DIRECCION TECNICA DE DEMOGRAFIA E
INDICADORES SOCIALES

GUIA PARA LA APLICACIÓN DEL ANALISIS MULTIVARIADO A LAS ENCUESTAS DE HOGARES

Lima, Enero 2002

DIRECCIÓN Y SUPERVISION

Rosario Aquije Valdez
Directora Técnica de Demografía e Indicadores Sociales

RESPONSABLES DEL ESTUDIO

Econ. Rofilia Ramírez Ramírez
Ing. Estad. Herman Edgar Castillo Ramón

SOPORTE INFORMATICO

Sr. Walter Ayala Godiño

Preparado	:	Dirección Técnica de Demografía e Indicadores Sociales del Instituto Nacional de Estadística e Informática (INEI)
Impreso	:	Talleres de la Oficina Técnica de Administración del INEI
Diagramación	:	Centro de Edición de la Oficina Técnica de Difusión del INEI
Tiraje	:	500 Ejemplares
Domicilio	:	Av. General Garzón 658, Jesús María. Lima - Perú
Orden de Impresión	:	Nº 170-OTA-INEI
Depósito Legal Nº	:	150113-2002-0155

PRESENTACION

El Instituto Nacional de Estadística e Informática (INEI), en el marco de su política de difusión de las técnicas multivariadas de uso más frecuente, pone a disposición de las autoridades, investigadores y usuarios en general el documento "**GUIA PARA LA APLICACIÓN DEL ANALISIS MULTIVARIADO A LAS ENCUESTAS DE HOGARES**".

Los temas que se analizan en el presente documento, aplicando las técnicas multivariadas son el desempleo, la fecundidad y el desarrollo humano. La fuente de información que se ha utilizado para este análisis proviene de la Encuesta Nacional de Hogares (ENAHO 2000) y la Encuesta Demográfica y de Salud Familiar (ENDES 2000).

Las técnicas multivariadas permiten el estudio interrelacionado de las variables sociales y demográficas, a partir de modelos matemáticos con los que se representan las múltiples dimensiones de la realidad, facilitando al investigador un mayor acercamiento a los fenómenos sociales. Para una mejor comprensión y utilidad de estas técnicas se utilizan dos enfoques. El primero tiene que ver con "las variables y sus interrelaciones" y el segundo enfoque está relacionado con la evaluación de "la calidad de los datos".

En el documento se han aplicado las principales técnicas del análisis multivariado: el análisis factorial, el análisis discriminante y el análisis de conglomerados . Con el análisis factorial se estudian los factores explicativos del desempleo. Con el análisis discriminante se estudia la fecundidad, mediante la conformación de grupos de mujeres de acuerdo a un conjunto de variables seleccionadas y con el análisis de conglomerados se estudia del índice de desarrollo humano, a partir de la clasificación de los departamentos en función de un conjunto de variables determinadas.

Esperamos que este documento aporte los instrumentos para el análisis de los fenómenos sociales y sea de utilidad para las autoridades y usuarios en general, a quienes agradeceremos sus opiniones o sugerencias sobre las técnicas presentadas.

Lima, Enero 2002

Gilberto Moncada Vigo
Jefe del INEI

INDICE

I. OBJETIVOS

1.1 Proponer un enfoque multivariado para el análisis de las encuestas de hogares

Las encuestas de hogares captan información periódica de un conjunto de variables sociales, constituyendo la fuente de datos más importante para el análisis de los problemas sociales, el diseño y seguimiento de las políticas sociales en el país. El análisis multivariado a su vez, es la herramienta más apropiada para el estudio sistemático y simultáneo de dos o más variables. Por ello, en este documento se propone el modo más adecuado de aplicar las técnicas multivariadas utilizando la información de las encuestas de hogares. De este modo los investigadores dispondrán de los elementos necesarios para contrastar sus hipótesis planteadas.

OBJETIVOS ESPECIFICOS

1.1 Presentar nuevas herramientas analíticas para el análisis de los problemas sociales

En este documento se presenta un conjunto de nuevas herramientas analíticas para la investigación de los problemas sociales y el análisis de los resultados de las encuestas de hogares. Con esto se pretende ampliar el conocimiento de los métodos descriptivos tradicionalmente utilizados en los estudios demográficos y sociales mostrando el aumento de su efectividad cuando estos métodos son complementados por las técnicas multivariadas para el tratamiento simultáneo de las variables.

1.2 Proporcionar nuevos criterios para evaluar la consistencia de los datos

El análisis multivariado no se ha enfocado suficientemente para evaluar la consistencia de los datos habiéndose desarrollado la mayor parte de sus aplicaciones para la formulación de modelos causales y no causales y el contraste de hipótesis. Por ello, en el presente documento se explican nuevos criterios e instrumentos para verificar la consistencia simultánea de dos o más variables y así evaluar la base de datos de las encuestas de hogares.

II. ANÁLISIS MULTIVARIADO

Las investigaciones sociales y demográficas proporcionan cuantiosa información por la diversidad de temas enfocados en ellas así como por el gran número de observaciones que integran las muestras. A ello se agrega el hecho que las variables investigadas se expresan en diferentes escalas (nominal, ordinal, de razón e interválica) ¿Cómo analizar toda esa información? ¿De qué manera se puede reducir el número de variables y/o datos sin afectar el objeto social en estudio?. **La técnica matemática** que permite el análisis simultáneo de dos o más variables, la reducción de los datos, la descomposición en factores del fenómeno social, la clasificación y el ordenamiento de las unidades investigadas, es el Análisis Multivariado. Para un dominio de este tipo de análisis se requiere del conocimiento de las matemáticas avanzadas, sin embargo, en esta guía se presentan de un modo didáctico las técnicas multivariadas aplicadas a tres casos que se generan a partir de las encuestas de hogares.

Los tres casos prácticos presentados, analizan el problema social, desde su formulación, su representación en modelos y la interpretación de los resultados. Se utiliza como fuente de información la Encuesta Nacional de Hogares (ENAHO) del III trimestre del 2000 y la Encuesta Demográfica y de Salud Familiar (ENDES), 2000.

El Análisis Multivariado es una técnica matemática que permite el estudio simultáneo de las relaciones entre más de dos variables y de las unidades de análisis en un período de tiempo determinado.

Las estadísticas univariadas (una sola variable), son empleadas con frecuencia en los estudios sociales. En estos casos es suficiente conocer la media y la varianza del objeto en estudio para determinar la distribución de la variable y de este modo inferir los resultados de la investigación mediante las pruebas de hipótesis y la conformación de intervalos de confianza. Estos aspectos fueron explicados con mayor detalle en el documento " Guía Metodológica para la Evaluación de Indicadores Sociales de las Encuestas de Hogares ", elaborado por la DTDIS-INEI.

El avance de la Tecnología Informática (hardware y software) hizo posible que en las investigaciones se apliquen modelos analíticos complejos para estudiar los fenómenos sociales, dando lugar al uso frecuente de las **técnicas multivariadas.** Esto facilitó el estudio simultáneo de las variables y una mejor aproximación al conocimiento de la realidad social. Mediante el análisis multivariado se pueden analizar más de dos variables en forma simultánea, generándose por cada variable una media y varianza, dando lugar a la formación de la matriz de covarianzas. La

formulación de un modelo analítico, mediante el cual se hacen explícitas las relaciones entre más de dos variables, y el establecimiento de algunos supuestos previos, facilitarán la interpretación posterior de los resultados.

Para seguir un procedimiento sistemático y de control de calidad de los procesos, se recomienda iniciar el análisis multivariado elaborando los mapas conceptuales y los diagramas de procesos. "Los primeros permitirán la presentación de los conceptos sociales de un modo estructurado y de fácil comprensión, mientras que los diagramas de procesos hacen posible identificar las actividades secuenciales que se siguen en la investigación[1/]", haciendo posible incorporar más adelante medidas estadísticas de calidad en cada etapa del proceso.

La técnica multivariada a utilizar se determina a partir del conocimiento de la forma en que se relacionan las variables inherentes al fenómeno social en estudio.

2.1 Principios básicos

Para la aplicación eficiente del análisis multivariado tenga presente lo siguiente:

2.1.1 En cuanto a las variables

El análisis presenta restricciones según el tipo de variables existiendo una técnica específica de acuerdo al tipo de variables investigadas. Por ejemplo el análisis factorial y de conglomerados se aplica generalmente cuando las variables son cuantitativas (interválicas y de razón), mientras que el análisis discriminante exige que la variable dependiente sea cualitativa y las independientes cuantitativas o dicotómicas.

2.1.2 En cuanto a las unidades de análisis

El análisis es de corte transversal (en un determinado período de tiempo) y permite el estudio simultaneo de todas las unidades de análisis, y además se puede identificar unidades particulares, como PEA desocupada, ocupada, mujeres en edad fértil, adultos mayores, entre otros.

2.1.3 Principales elementos que intervienen en el análisis multivariado

Los principales elementos del análisis multivariado son:

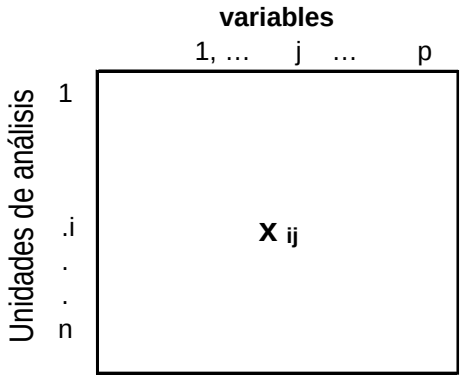
La unidad de análisis: Es la unidad de investigación o estudio, sobre la cual se realiza el análisis. Por ejemplo: la vivienda, el hogar, la persona o un ámbito geográfico (departamento, provincia, distrito, localidad, etc.).

Las variables: Son las características observables o conceptuales de la unidad de análisis. Estas características pueden registrarse en diferentes tipos de escalas. Las variables cuantitativas son medidas en una escala ordinal, interválica y de razón, por ejemplo: años de estudio, el ingreso,

[1/] Guía Metodológica "Evaluación de Indicadores Sociales de las Encuestas de Hogares, 2001.

la edad, los gastos del hogar. Las variables cualitativas son medidas en una escala nominal por ejemplo el sexo, la condición de alfabeto, la tenencia de la vivienda, etc.

La Matriz o Tabla de Datos: Una matriz o tabla de datos esta conformada por la disposición en filas de las unidades de análisis , mientras que en las columnas se ubican las variables. La intersección de una fila y una columna da lugar al dato. Por ejemplo la matriz de datos que se muestra consta de n unidades de análisis y p variables, xij representa un dato en particular. Se tendrá tantas matrices de datos como submuestras de población lo permita la encuesta.



2.2 El análisis multivariado en la Investigación Social y Demográfica

Los métodos multivariados aportan los elementos analíticos y operativos para que las Investigaciones Sociales y Demográficas, se realicen en un marco científico de tal manera que con el gran volumen de datos proveniente de las Encuestas de Hogares pueda estudiarse mejor la realidad social. Con la formulación de un modelo multivariado pueden evaluarse las hipótesis establecidas las mismas que serán contrastados empíricamente[2/].

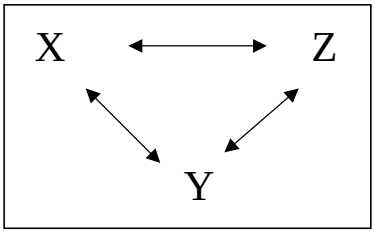
El tipo de relación entre las variables define la forma funcional del modelo. Por ejemplo si entre las variables no hay dependencia o todas son interdependientes, las técnicas de análisis más apropiadas son el Análisis Factorial, el Análisis de Conglomerados (Cluster), el Análisis de Correlación Canónica, el Análisis de Componentes Principales, entre otros. En cambio, si alguna variable (s) es dependiente(s) de otra u otras entonces se aplica: el Análisis de Regresión Multivariado, el Análisis Discriminante y el Análisis de Contingencia Múltiple.

[2/] Debe entenderse por contraste empírico al procedimiento que permite comprobar a partir de los datos recogidos las hipótesis formuladas.

Para facilitar la interpretación de los resultados se recomienda que el investigador formule a priori un modelo en el cual se hagan explícitas las relaciones entre las variables.

2.2.1 Análisis de la Interdependencia

Cuando no se puede determinar que una variable(s) determina a la otra(s), se dice que la relación entre las variables es de interdependencia. La condición de interdependencia queda establecida por el conocimiento a priori y/o los supuestos que el investigador tenga del fenómeno social en estudio. Por ejemplo en el siguiente diagrama causal, las variables X, Y , Z son interdependientes, así X es causa de Y, a su vez Y es causa de X , así Y lo es de Z y Z lo es de X. La interdependencia entre las variables puede resumirse en el siguiente diagrama causal[3/]:

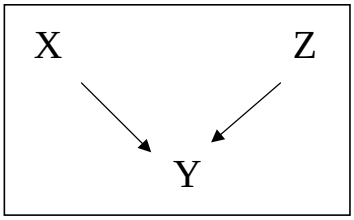


Ejemplo práctico: El índice de desarrollo educativo de la niñez y la adolescencia queda determinado a partir de un conjunto de factores denominados como de contexto, de proceso y de resultado. Esta denominación se realizó atendiendo a las características de las variables que conformaban cada factor. Así son variables de contexto: la proporción de niños con desnutrición crónica y el nivel educativo de la madre; de proceso: la proporción de alumnos en extra edad escolar, la proporción de menores de 15-17 años que estudia y trabaja, la tasa de desaprobación en educación primaria de menores y de resultado: el grado de estudio aprobado a los 17 años de edad, proporción de menores de 15-17 años que no saben leer ni escribir, proporción que no estudio ni trabaja. Entre estas variables no se estableció a priori ninguna relación de dependencia, ingresando todas ellas como explicativas del desarrollo educativo y por tanto interdependientes.

2.2.2 Análisis de la Dependencia

La relación entre las variables es de dependencia, cuando el conocimiento a priori del objeto social en estudio o los supuestos determinan que una variable (dependiente) es determinada por otras. Por ejemplo: en el siguiente diagrama causal la variable Y depende de X y Z. Así, Y depende de X, Z, y entre X y Z no hay relación.

[3/] El diagrama causal es una representación gráfica de los diferentes tipos de relación entre las variables. Para ello se utilizan diferentes figuras geométricas y flechas que indican el sentido de la relación entre las variables.



Ejemplo práctico: Un estudio de la calidad y la eficiencia de los hospitales determinó que las diferencias de atención en los diferentes hospitales son explicadas por las variables relacionadas con los recursos humanos que laboran en los hospitales (número, formación, compromiso con la institución), y otros factores como la disponibilidad de tecnología y el nivel de especialización, así como por las diferentes combinaciones de estos factores. La técnica multivariada utilizada para la comprobación empírica de esta relación causal, fue el análisis de regresión múltiple.

2.3 Aplicaciones al análisis de las encuestas de hogares

Las aplicaciones del Análisis Multivariado están dirigidas principalmente a la reducción de la estructura de datos y a la clasificación de las unidades de análisis o las variables en grupos. Además en este documento se presentan algunas aplicaciones dirigidas a evaluar la consistencia de las bases de datos de las Encuestas de Hogares.

2.3.1 Simplificación de la Estructura de Datos

Se busca encontrar la manera de reducir el universo de datos sin afectar al fenómeno social en estudio. Esto puede lograrse mediante la transformación (combinación lineal o no lineal) de un conjunto de variables interdependientes en otro conjunto de menor dimensión.

Ejemplos aplicativos:

1. La matriz de datos inicial compuesta por n-filas (unidades de análisis) y p-variables se reduce a una matriz de menor dimensión mediante el análisis de componentes principales (análisis factorial).

2. Cuando las unidades de análisis se disponen en grupos homogéneos mediante el análisis de conglomerados, se reduce el número de unidades a analizar, simplificándose su interpretación.

3. Cuando las n variables originales se reducen en n-1 factores (análisis factorial), siendo cada factor una combinación lineal de las variables, representando estos factores una dimensión diferente del fenómeno social observado, se logra un análisis simplificado y ordenado de los resultados.

2.3.2 Clasificación de las variables y las unidades de análisis

Se persigue encontrar el modo mas eficiente de agrupar las variables o las unidades de análisis.

ejemplos aplicativos:

1. Los departamentos del Perú (unidades de análisis) pueden ordenarse mediante el análisis factorial, a partir del puntaje obtenido como combinación lineal ponderada de un conjunto de variables interrelacionadas. Además pueden disponerse en grupos homogéneos y heterogéneos entre si, mediante el análisis de conglomerados.

2. Las variables relacionadas con el desarrollo educativo pueden ser clasificadas en dimensiones diferentes (contexto, proceso y resultado) aplicando el análisis factorial.

2.3.3 Evaluación de la consistencia transversal de los datos

Además de las aplicaciones mencionadas, los modelos analíticos multivariados pueden emplearse para evaluar la consistencia transversal de los datos.

ejemplos aplicativos:

1. A partir del análisis factorial se formula un modelo explicativo, cuyos coeficientes calculados (magnitud y signo) permitirán confirmar la consistencia de los datos. Si la magnitud de estos coeficientes excede los rangos esperados entonces debe revisarse la base de datos, de igual manera si se conoce el sentido de la relación entre las variables (signo) cualquier resultado diferente al esperado conducirá a la revisión de los datos. Así, al formular un modelo factorial que incorpore las variables gastos e ingresos F (G, Y), se espera que el coeficiente que mide la relación entre estas variables tenga signo positivo (relación directa), por cuanto "los gastos son directamente proporcionales al ingreso", un signo negativo permitirá inferir la inconsistencia de la información.

2. Utilizando el análisis de conglomerados (cluster) los grupos de unidades de análisis conformados de acuerdo a un criterio de clasificación se espera sean semejantes a los grupos formados con otro método, de tal manera que al correlacionar el ordenamiento obtenido por el análisis de conglomerados con el ordenamiento generado por otro método, se obtenga un coeficiente de correlación significativo, (al menos 0,5). Cualquier otra situación determinará la necesidad de revisar la base de datos. Por ejemplo al aplicar el análisis cluster para ordenar los departamentos del Perú de acuerdo al nivel de pobreza (criterio) se obtienen una clasificación de los departamentos en estratos pobres y no pobres. Esta clasificación se compara con el ordenamiento simple de los departamentos según el Indice de Necesidades Insatisfechas (NBI), esperando encontrar un ordenamiento semejante (correlación significativa). De no encontrarse esta relación puede inferirse que la información es inconsistente.

3. De una muestra total se seleccionan aleatoriamente dos submuestras, al aplicar el análisis discriminante a cada muestra se analiza la tabla de ordenamiento de las variables esperando encontrar cierta semejanza en ambas, lo que permitirá inferir la consistencia de la información. Por ejemplo del modulo de empleo de la ENAHO se seleccionan aleatoriamente dos submuestras, se aplica a cada una de ellas el análisis discriminante para encontrar las variables que explican la condición de subempleo. Las variables mas explicativas se espera sean las mismas en ambas submuestras. Cualquier otro resultado conducirá a la revisión de la base de datos, en especial de aquellas variables cuya importancia relativa difiera en las submuestras.

III. ANÁLISIS FACTORIAL CONFIRMATORIO

El análisis factorial es una técnica del Análisis Multivariado que permite obtener a partir de un conjunto de variables un grupo menor de nuevas variables denominadas factores, los mismos que estarían explicando la variación conjunta o dependencia mutua entre dichas variables. Estos factores denominados también variables "latentes" se caracterizan por no estar **correlacionados entre sí.** Con esta reducción se hace más sencillo el análisis de los resultados.

Los tipos más frecuentes del análisis factorial son: **el análisis factorial exploratorio y el análisis factorial confirmatorio**. El primero se utiliza cuando el investigador requiere clasificar las variables en dimensiones excluyentes (factores). Por ejemplo, mediante el análisis factorial exploratorio podemos disponer las variables relacionadas con el Indice de Desarrollo Educativo en sus tres dimensiones: **contexto** (condicionantes y medio social), **proceso** (variables explicativas), y **resultados** (rendimiento educativo).

Con el análisis factorial las variables se disponen en factores no correlacionados entre sí, donde cada factor representa una dimensión diferente del fenómeno social de este modo se logra una apreciación integral del objeto en estudio.

3.1 El análisis factorial confirmatorio

El análisis factorial confirmatorio se aplica cuando el investigador tiene un conocimiento a priori del fenómeno en estudio, lo cual le permite formular las hipótesis necesarias acerca de la relación de causalidad entre las variables así como restringir el valor de algunos parámetros del modelo antes de calcularlo (coeficientes del modelo). Así, el modelo a priori formulado es contrastado con los resultados muestrales obtenidos.

El análisis factorial confirmatorio permite contrastar sistemáticamente las relaciones formuladas a priori entre las variables de tal manera que se compruebe empíricamente los supuestos y los resultados tengan la consistencia estadística esperada.

3.2 El análisis factorial confirmatorio y su aplicación a la ENAHO

La encuesta nacional de hogares (ENAHO), es una encuesta que periódicamente ejecuta el INEI y está orientada a obtener información tanto del hogar como de las personas que lo habitan en temas relacionados con las características de la vivienda y del hogar, empleo e ingresos, educación, salud, programas sociales y condiciones de vida de la población. Las variables que integran cada uno de los temas, pueden ser analizadas simultáneamente y de una manera eficiente mediante las técnicas multivariadas.

El análisis factorial, permite desarrollar investigaciones sociales con variables tanto cuantitativas como cualitativas. Así, se puede estudiar el desempleo abierto aplicando un modelo factorial confirmatorio, con las diferentes variables seleccionadas de la base de datos que integran el modulo empleo de la ENAHO.

3.3 Estudio de los factores explicativos del desempleo abierto

En el siguiente caso práctico se estudia el desempleo abierto utilizando la técnica factorial confirmatoria. Como toda investigación esta se inicia con la revisión conceptual. Así, se considera que están en situación de desempleo abierto, las personas de 14 años y más de edad que cumplen las tres condiciones siguientes: desean trabajar, están disponibles para hacerlo y se encuentran buscando activamente un empleo. En el país esta situación afecta aproximadamente el 10% de la población económicamente activa (PEA), siendo necesario efectuar un estudio sistemático de este problema social. A partir de un modelo factorial confirmatorio se identificarán las variables más relacionadas con el desempleo y que contribuyen a explicar las condicionantes de esta situación.

3.3.1 Etapas para realizar el análisis factorial confirmatorio

Para realizar el análisis factorial confirmatorio siga las siguientes etapas:

1. **Formule las hipótesis:** Para plantear las hipótesis se formulan las siguientes preguntas: ¿Cuáles son los factores explicativos del desempleo? ¿Qué variables contribuyen más a explicar estos factores? A partir de estas preguntas se pueden formular las hipótesis siguientes:

   Primera hipótesis: El desempleo abierto está determinado por factores relacionados con la demanda del mercado laboral (las expectativa del empleador), la oferta de mano de obra (el perfil profesional del desempleado) y otro factor no considerado en los anteriores.

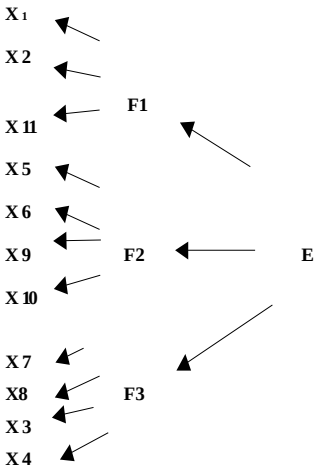
   Segunda Hipótesis: El nivel educativo es la variable más determinante en la condición de desempleo.

2. **Genere un modelo multivariado:** El conocimiento a priori del fenómeno social determina la selección de las variables, su número y el sentido de la relación entre ellas. Así, mediante un diagrama causal podemos representar las relaciones entre las variables. En este ejemplo se han seleccionado once variables explicativas del problema social.

Modelo factorial confirmatorio del desempleo abierto:

$$E_j = \mu_j + \lambda_{j1} \cdot F1 + \lambda_{j2} \cdot F2 + \lambda_{j3} \cdot F3 + \varphi_j$$
$$j = 1, \ldots, n$$

Diagrama causal:



donde:

E : Desempleo (variable explicada)

F1 : Primer Factor. Demanda del mercado laboral (expectativa del empleador)

F2 : Segundo Factor. Oferta de mano de obra (el perfil profesional del desempleado)

F3 : Tercer Factor. Otro factor no considerado entre los anteriores

Xj : j - ésima variable explicativa

3. Variables explicativas: Generalmente el análisis factorial se realiza con variables medidas en una escala intervalar, porque la matriz de correlación punto inicial del análisis se basa en el cálculo del coeficiente de Pearson. Para el presente ejemplo, se ha considerado en el modelo factorial confirmatorio algunas variables nominales las cuales han sido convertidas a dicotómicas, codificándose sus categorías con 0 y 1. El valor 1 indicará la presencia de la cualidad correspondiente a una de las dos categorías y el 0 la ausencia de dicha cualidad. Este artificio permitirá obtener el valor analítico (signo) de los coeficientes estimados, al margen de la escala en que se han medido estas variables. Así, se ha considerado las siguientes variables explicativas:

No	Variable	Tipo	Código
1	¿Ha trabajado antes?	Nominal-dicotomica	EXPLABOR
2	¿Ha aprendido un oficio a través de la experiencia?	Nominal-dicotomica	P592
3	¿Cuántas semanas ha buscado trabajo?	Interválica	p564
4	¿Tiene profesión?	Nominal-dicotomica	p584b
5	Edad	Interválica	EDA500
6	Sexo	Nominal-dicotomica	SEXO
7	¿Total de miembros del hogar?	Interválica	MIEPERHO
8	Area	Nominal-dicotomica	AREA
9	Condición de Unión	Nominal-dicotomica	CONCIVIL
10	Años de estudios	Interválica	ANOEST
11	¿Actualmente lleva o llevó cursos de capacitación?	Nominal-dicotomica	P587

4. **Fuente de datos:** La fuente de datos para probar las hipótesis planteadas es la ENAHO 2000 III trimestre, se utiliza la base de datos sin expandir, para que los coeficientes estimados, no se vean afectados por los factores de expansión. Además la base de datos debe estar conformada por la población objetivo (unidades de análisis) , que para el presente caso práctico es la población de 14 años y más de edad que se encuentran en situación de desempleo abierto.

5. **Inicie el Análisis Factorial Confirmatorio** utilizando el programa SPSS. Abra la base de datos con las variables seleccionadas. Luego en el menú de barras del SPSS ubique la opción Analyze / Data Reduction / Factor, de acuerdo al cuadro de diálogo siguiente:

La opción **Data Reduction** contiene las técnicas de reducción de datos también llamadas de reducción de las dimensiones del fenómeno en estudio. Estas son el "Factor Analysis" (Análisis Factorial), el "Correspondence Analysis" (Análisis de Correspondencias) y el "Optimal Scaling" (Escalamiento Optimo).

El "Factor Analysis" (Análisis Factorial), se aplica principalmente a las variables cuantitativas, mientras que las otras técnicas son apropiadas para variables cualitativas, permitiendo encontrar las relaciones entre las categorías de las variables.

El "Correspondence Analysis" (Análisis de Correspondencias), se utiliza para estudiar la semejanza entre variables con igual número de categorías. Por ejemplo se puede estudiar la relación entre el sexo (hombre / mujer) de la persona y su condición de ocupación (ocupado / desocupado).

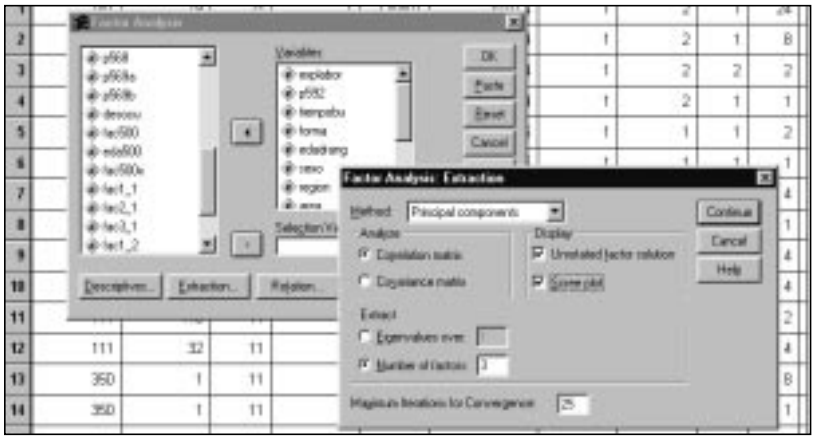
El "Optimal Scaling" (Escalamiento Optimo), permite el estudio de variables cualitativas con diferente número de categorías. Así se puede estudiar la relación entre el nivel educativo de una persona (primaria, secundaria, superior) y el ámbito (urbano / rural).

Regresando al análisis factorial confirmatorio, la opción factor habilita una ventana que facilita la selección de las variables explicativas, de modo que las once variables explicativas seleccionadas se trasladen al recuadro de "Variables". Las opciones del Data Reduction / Factor, son: "Descriptives", "Extraction", "Rotation", "Scores" y "Options", que a continuación se explican con más detalle.

La Ventana DESCRIPTIVES: Habilita el cuadro de diálogo siguiente "Factor Analysis: Descriptives"

La opción estadística **"Statistics"** - Univariate descriptives, activada facilita la generación de las estadísticas descriptivas (media, desviación estándar, moda, entre otras) para cada una de las variables seleccionadas. Al activar "Initial solution" el programa mostrara la solución inicial (comunalidad inicial) en la salida ("output"). Las opciones activadas en **"Correlation Matrix"** -Matriz de Correlación-, permiten que se calculen los coeficientes de correlación de Pearson, ("coefficients"), el determinante de la matriz (determinant) y los niveles de significación estadística de los coeficientes estimados (significance levels). Las pruebas de esfericidad de Barlett y el test de Kaiser-Meyer-Olkin (KMO) permitirán evaluar la calidad de las estimaciones.

La Ventana EXTRACTION: En esta ventana seleccione las siguientes opciones:



En **Method**, elija el método de componentes principales "Principal Components" por ser el punto de partida para para estimar los factores.

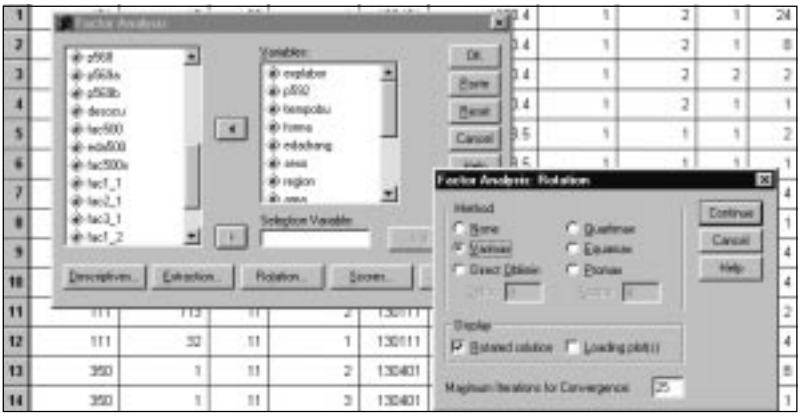
En **Analyze** active la opción de la matriz de correlaciones, de modo que se muestren las asociaciones entre las variables. En caso de que no exista asociación entre las variables, la matriz de correlación será igual a la matriz identidad. La siguiente opción Covariance matriz (matriz de covarianzas) es opcional.

En **Extract** active cualquiera de las opciones que se muestran, pues ambas permiten fijar el número valores eigen o de factors a extraer del resto de variables. En este caso se ha elegido extraer tres factores.

En **Display** debe activar las dos opciones "unrotated factor solution" (solución factorial no rotada) para comparar las soluciones entre factores sin rotar y rotadas, el "scree plot" para obtener el gráfico respectivo.

En Maximum Iterations for convergence, especificar el número máximo de pasos que el algoritmo puede tomara para estimar la solución. Por defecto este número es de 25.

La Ventana ROTATION: En esta ventana debe seleccionar las siguientes opciones:



En **Method** active el método de rotación Varimax, para la rotación ortogonal de las componentes o factores, de modo que las variables fuertemente correlacionadas entre sí presenten concentraciones altas sobre un mismo factor (por ejemplo las variables correlacionadas con las características demográficas estén agrupadas en un mismo factor) y su concentración en otros factores sean bajas. Así, se optimiza la solución.

En **Display** active la opción "Rotated solution" para mostrar la solución rotada. El número de iteraciones para la convergencia de la solución que por defecto aparece es 25.

La Ventana SCORES: La ventana scores muestra la siguientes opciones:

Los puntajes factoriales (scores) son las proyecciones de los valores de las variables sobre cada uno de los factores o componentes hallados. Debe activarse la opción **Save as variables**, para grabar en la base de datos estos puntajes factoriales (F_1, F-2....) como nuevas variables. Estos puntajes permitirán ordenar las unidades de análisis, recomendándose obtenerlos por el método de regresión por ser este un procedimiento más estandarizado. Además debe activarse la opción Display factor **score coefficient matrix** que muestra la matriz de transformación de las componentes para calcular los puntajes factoriales.

La Ventana OPTIONS: En la ventana Options, active las siguientes funciones según se indican:

En **Missing Values**, active la opción "Exclude cases listwise" por la que se excluye del análisis las unidades de análisis con valores incompletos. En **Coefficient Display Format** active la opción "Sorted by size" que permite presentar en forma ordenada las variables de acuerdo a su mayor correlación con los factores estimados.

Una vez realizada todas las selecciones pulsar OK en el cuadro de dialogo principal para ejecutar todo el procedimiento de análisis factorial.

6. ¿Cómo se interpretan los resultados del análisis factorial realizado?

El análisis factorial realizado ha permitido la obtención de los tres factores en los cuales se agrupan las once variables inicialmente consideradas y la generación de los indicadores de calidad de las estimaciones. Las interpretaciones de las salidas que otorga el programa son:

Matriz de correlación

La matriz de correlación contiene los coeficientes de correlación de Pearson (medida de asociación lineal entre las variables). La magnitud y el signo del coeficiente determina el grado y sentido de la relación entre las variables explicativas. Cuanto mayor sea el valor del coeficiente mayor será la relación entre las variables. Si es positivo entonces la relación entre las variables es directa, pero si el signo es negativo la relación es inversa. Mientras más cerca a uno se encuentren los coeficientes de correlación implica que la relación entre las variables es mas alta,

mientras que los valores próximos a cero implican ausencia de la correlación (relación). El determinante de la matriz se indica al pié de la tabla. Debajo de los coeficientes de correlación se muestra la significación estadística de los coeficientes calculados, la misma que viene dada por la proximidad a cero de los valores mostrados. Así, cuanto más próximos a cero se encuentren estos valores los coeficientes serán estadísticamente significativos.

Correlación Matrix[a]

	Ha trabajado antes?	Ha aprendido algún oficio a través de la expe-riencia?	Cuantas semanas ha estado buscando trabajo sin interrupción	Tiene profesión	Edad (Años)	Sexo	Total de miembros del hogar	Area	Condición de unión	Años de Estudio	Actual-mente lleva o ha llevado cursos de capaci-tación?
Ha trabajado antes?	1.000	-0.190	0.000	0.027	-0.216	0.095	-0.010	0.049	0.194	-0.009	-0.087
Ha aprendido algún oficio a través de la experiencia?	-0.190	1.000	0.006	-0.187	0.260	-0.308	0.016	0.008	-0.199	-0.200	-0.120
Cuantas semanas ha estado buscando trabajo sin interrupción?	0.000	0.006	1.000	0.023	0.138	-0.028	0.066	-0.1	-0.046	0.058	0.028
Ninguna profesión	0.027	-0.187	0.023	1.000	-0.077	0.090	-0.130	-0.1	0.103	0.725	0.277
Edad (Años)	-0.216	0.260	0.138	-0.077	1.000	-0.162	-0.145	-0.1	-0.524	-0.214	0.014
Sexo	0.095	-0.308	-0.028	0.090	-0.162	1.000	-0.007	0.001	0.006	0.076	0.103
Total de miembros del hogar	0.010	0.016	-0.066	-0.130	-0.145	-0.007	1.000	0.070	0.040	-0.102	-0.071
Area	0.049	0.008	-0.054	-0.134	-0.093	0.001	0.070	1.0	0.017	-0.190	-0.153
estado civil	0.194	-0.199	-0.046	0.103	-0.524	0.006	0.040	0.017	1.000	0.154	0.004
Años de estudio	-0.009	-0.200	0.058	0.725	-0.214	0.076	-0.102	-0.2	0.154	1.000	0.373
Actualmente lleva o ha llevado cursos de capacitación?	-0.087	-0.120	0.028	0.277	0.014	0.103	-0.071	-0.2	0.004	0.373	1.000
Ha trabajado antes?		0.000	0.499	0.158	0.000	0.000	0.355	0.037	0.000	0.365	0.001
Ha aprendido algún oficio a través de la experiencia?	0.000		0.411	0.000	0.000	0.000	0.275	0.384	0.000	0.000	0.000
Cuantas semanas ha estado buscando trabajo sin interrupción?	0.499	0.411		0.197	0.000	0.156	0.007	0.024	0.047	0.017	0.156
Ninguna profesión	0.158	0.000	0.197		0.002	0.000	0.000	0.000	0.000	0.000	0.000
Edad (Años)	0.000	0.000	0.000	0.002		0.000	0.000	0.000	0.000	0.000	0.309
Sexo	0.000	0.000	0.156	0.000	0.000		0.393	0.487	0.413	0.003	0.000
Total de miembros del hogar	0.355	0.275	0.007	0.000	0.000	0.393		0.005	0.073	0.000	0.005
Area	0.037	0.384	0.024	0.000	0.000	0.487	0.005		0.266	0.000	0.000
estado civil	0.000	0.000	0.047	0.000	0.000	0.413	0.073	0.266		0.000	0.447
Años de estudio	0.365	0.000	0.017	0.000	0.000	0.003	0.000	0.000	0.000		0.000
Actualmente lleva o ha llevado cursos de capacitación?	0.001	0.000	0.156	0.000	0.309	0.000	0.005	0.000	0.447	0.000	

a. Determinant = 0.169

La matriz de correlación es una matriz simétrica cuyos coeficientes de correlación se muestran por encima y por debajo de la diagonal principal (formada por uno). Como se esperaba algunos coeficientes están más cerca de la unidad, mientras que otros valores están próximos a cero. Así, la correlación entre los años de estudio y la tenencia de profesión, es 0.725 (coeficiente más alto), asimismo este coeficiente es estadísticamente significativo (0.000). La significación estadística viene dada por los valores que se muestran debajo de la matriz de correlación. El coeficiente de correlación es significativo si el valor que se muestra en la parte inferior a la matriz de correlación es inferior a 0.05. Cualquier otro valor, indica que el coeficiente de correlación no es significativo y por tanto no son válidas las inferencias, siendo sus valores solo referenciales.

Es importante que todas las variables tengan al menos un coeficiente de correlación significativo en la matriz. El coeficiente de correlación de Pearson es apropiado cuando las variables son cuantitativas. Sin embargo, las variables cualitativas incorporadas al modelo permitirán la disposición adecuada de las variables en los factores al margen de la magnitud y signo de los coeficientes de correlación a partir de ellas generados.

Pruebas de Kaiser-Meyer-Olkin (KMO) y de Bartlett

La prueba de Bartlett, está referida a la matriz de correlaciones. Se contrasta la siguiente hipótesis nula (Ho): La matriz de correlaciones es una matriz de identidad; versus la hipótesis alternante: la matriz de correlaciones no es una matriz de identidad. En caso de rechazar la Ho se concluye que las variables están correlacionadas entre sí, lo que da sentido al análisis factorial a realizar.

La prueba de **Kaiser-Meyer-Olkin** (KMO) estima un valor que de acuerdo a su ubicación en una escala permitirá concluir si el análisis realizado es conveniente. Este KMO se basa en la relación entre los coeficientes de correlación de Pearson y los coeficientes de correlación parcial entre las variables. En la medida que los primeros sean más altos, el valor estimado estará mas cerca de uno, y por tanto el modelo factorial empleado será más efectivo.

Prueba KMO y Bartlett

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.602
Bartlett's Test of Sphericity	Approx. Chi-Square	2375.412
	df	55
	Sig.	.000

Realizada la prueba de Bartlett, por ser el Sig (0.00) inferior al valor de 0.05 a priori fijado, se rechaza la Ho. Se concluye que es poco probable que la matriz de correlación sea una matriz de identidad, y por tanto la correlación entre las variables es estadísticamente significativa.

Para interpretar el KMO obtenido se requiere ubicar este valor (0.602) en la siguiente tabla:

1	>= KMO >	0.9	excelente
0.9	>= KMO >	0.8	bueno
0.8	>= KMO >	0.7	aceptable
0.7	>= KMO >	0.6	regular
0.6	>= KMO >	0.5	deficiente
KMO <=		0.5	inaceptable

El KMO calculado en el ejemplo es igual a 0.6 alcanzando un nivel deficiente de acuerdo a la escala presentada. Esto probablemente sea consecuencia de que más de la mitad de variables empleadas en el caso práctico analizado son dicotómicas.

Comunalidad

La comunalidad, es una medida del aporte de los factores a la explicación de las variables, mientras más próximos a uno estén las comunalidades, significa que los factores explican en su totalidad la variabilidad.

Para el caso práctico las variables: Años de estudios, edad y tiene profesión son las variables mejor explicadas por los tres factores, al registrar las comunalidades más altas. El valor de 0.803 se interpreta de la siguiente manera: el 80% de la variabilidad de los años de estudios es explicada por los tres factores, mientras que el valor de 0.31 significa que la variable "ha trabajado antes" es explicada apenas en un 30% por los tres factores.

Porcentaje de varianza

El cuadro muestra el porcentaje de varianza del modelo que es explicado por los tres factores o componentes. En la columna "Rotation Sums of Squared Loadings" se indica que el primer componente explica el 19.1% de la variación total, el segundo componente explica el 14.6%, mientras que el tercero el 13.9%. Así, entre los tres factores explican el 48% del comportamiento de la variable explicada (E).

Porcentaje de Varianza Explicada

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.331	21.187	21.187	2.331	21.187	21.187	2.105	19.137	19.137
2	1.778	16.164	37.351	1.778	16.164	37.351	1.609	14.630	33.767
3	1.133	10.302	47.653	1.133	10.302	47.653	1.527	13.886	47.653
4	1.051	9.552	57.205						
5	.936	8.514	65.718						
6	.904	8.216	73.934						
7	.803	7.296	81.230						
8	.731	6.648	87.878						
9	.650	5.909	93.788						
10	.446	4.055	97.843						
11	.237	2.157	100.000						

Extraction Method: Principal Component Analysis.

Los valores Eigen, vienen a ser la expresión numérica de las componentes, explicando su valor un porcentaje de la varianza total. Por ejemplo en la columna "Rotation Sums of Squared Loadings" el valor Eigen que corresponde a 2.105 explica el 19.1%, mientras que los valores eigen de 1.609 y 1.527 explican el 14.6% y el 13.8%, respectivamente. Es decir el porcentaje de variación explicado crece en relación directa a la magnitud del valor Eigen. De este modo los tres primeros valores Eigen (equivalentes a tres factores o componentes) explican el 48% de la variabilidad total del modelo, lo que puede interpretarse como un porcentaje

aceptable, teniendo mas de la mitad de variación del desempleo explicada por los tres factores. En el caso de los modelos de prognosis se recomienda que el porcentaje de variación total del caso estudiado sea explicado por los factores al menos en el 60%.

Representación Gráfica

En el gráfico se representan en el eje de abcisas el número total de factores o componentes y en el eje de ordenadas su valor numérico. Se aprecia la relación inversa entre la magnitud del coeficiente y el número de factores. Desde que la magnitud del coeficiente mide el poder explicativo, se puede inferir que conforme se calculen más factores el poder explicativo decrece.

El punto de inflexión de la curva, señala el número ideal de factores a determinar. En el caso práctico desarrollado el gráfico confirma que el número ideal de factores es tres.

Matriz de Cargas o Pesos Factoriales Rotada

La tabla muestra la disposición de las variables en los factores según su grado de importancia. Así en el primer factor o componente, las variables: "años de estudio", "tiene profesión", "actualmente lleva o ha llevado cursos de capacitación" y "área" tienen mayor peso. A este factor se le denominó "Demanda del mercado laboral". En el segundo factor denominado "Oferta de mano de obra", tienen mayor importancia las variables: "edad", "condiciones de unión", "total de miembros del hogar" y "semanas buscando trabajo sin interrupción". En el tercer factor denominado "otro factor" se congregan las variables "sexo", "aprendió un oficio por la experiencia" y "ha trabajado antes". Las variables más representativas de cada factor es decir aquellas con los pesos más altos son el nivel educativo, la edad y el sexo, las mismas que podemos inferir son las más condicionantes del desempleo.

Matriz de Componentes Rotadas[a]

	Component		
	1	2	3
AÑOS DE ESTUDIO	.881	.151	7.184E-02
Tiene profesi¢n	.825	6.501E-02	8.613E-02
Actualmente lleva o ha llevado cursos de capacitaci¢n?	.596	-.160	7.279E-02
AREA	-.391	.186	.105
Edad (A¤os)	-9.495E-02	-.773	-.325
Condición de Unión	.146	.762	.155
Total de miembros del Hogar	-.221	.367	-.118
Cuantas semanas ha estado buscando trabajo sin interrupcion?	9.889E-02	-.314	4.586E-02
Sexo	2.759E-02	-.184	.772
Ha aprendido alg£n oficio a traves de la experiencia?	-.188	-.104	-.727
Ha trabajado antes?	-.125	.259	.477

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Principales conclusiones del análisis factorial realizado

El modelo factorial ha permitido determinar que las variables más condicionantes de la situación de desempleo abierto son: el nivel educativo, la edad y el sexo (son las que tienen el mayor peso o carga factorial en cada factor).

La disposición de las variables según su coeficiente de correlación determina que el primer factor (F1) se reúnan las variables: años de estudio, tenencia de profesión y cursos de capacitación llevados por el desocupado denominación este factor como **"calificación laboral"**; en el siguiente factor (F2) , se congregan las variables como: la edad, la condición de unión, el tamaño de la familia y las semanas que ha buscado trabajo por lo que se considera este factor como **"características sociodemográficos"**. En el tercer factor (F3) son más importantes las variables: sexo, ha aprendido algún oficio a través de la experiencia y si ha trabajado antes por lo cual se le ha denominado **"experiencia laboral"**.

La identificación de estas tres dimensiones hace posible realizar el análisis de los resultados a partir de ellas prescindiéndose de las variables originales.

En cuanto a la consistencia del modelo y los datos.

- **Consistencia del modelo:** las medidas de calidad para la validez del modelo (Bartlett y KMO) analizadas,

permiten inferir que la técnica factorial utilizada ha sido adecuada. Además el gráfico de componentes otorga validez al número de componentes estimadas. Otra medida de calidad son los valores que se presentan en la tabla de comunalidad. En esta tabla se muestra el porcentaje del comportamiento de la variable que es explicado por los factores estimados. En el caso práctico desarrollado la variable "años de estudio" es explicada en los tres factores en más del 80%, existiendo algunas variables como cuántas semanas ha estado buscando trabajo sin interrupción, total de miembros del hogar y área que no alcanzan el 20% en los tres factores.
La tabla de resultados de la Matriz de componentes rotados muestra la disposición de las variables en los factores o componentes según su grado de importancia. Esto permitirá confirmar las hipótesis a priori establecidas de tal manera que si las variables tienen un agrupamiento diferente al esperado, el modelo puede no ser el mas adecuado. En el ejemplo desarrollado las variables más explicativas del desempleo son aquellas con los coeficientes mas altos en los factores, confirmándose que el nivel educativo, la edad y el sexo contribuyen más a explicar esta situación.

- **Consistencia de los datos:** la matriz de correlación estimada permite verificar la consistencia de los datos, analizándose el sentido de la relación entre las variables (signo de los coeficientes de correlación) y la magnitud de estos. A partir de estos valores pueden inferirse si los datos confirman el conocimiento teórico sobre el problema, en caso contrario debe evaluarse la base de datos en especial de aquellas variables con coeficientes de correlación no significativos y con signos contrarios al esperado.

Para el ejemplo, los resultados mostrados en la tabla de comunalidad recomiendan revisar las variables cuántas semanas ha estado buscando trabajo sin interrupción, total de miembros del hogar y área, por cuanto el poder explicativo de los tres factores en ellas no alcanzan ni la mitad. En estos casos se recomienda realizar un análisis descriptivo de estas variables que permita detectar algunas inconsistencias (% de omisiones, alta varianza, valores extremos, entre otras).

En La Matriz de componentes rotados aquellas variables cuyos coeficientes rotados presentan valores semejantes que no permite definir su ubicación en algunos de los factores, deben ser revisadas. Para el ejemplo la variable área, cuántas semanas ha estado buscando trabajo, ha trabajado, pueden ubicarse en cualquiera de los factores dada la semejanza de sus valores, por ello se recomienda realizar el análisis de la distribución de sus valores en la base de datos original.

Además, se recomienda como procedimiento general efectuar una partición de los datos en dos muestras, seleccionadas aleateoriamente, a fin de replicar los procedimientos en varias submuestras Los resultados de la primera submuestra deberán contrastarse con los resultados de la segunda muestra, de tal manera que se alcancen la consistencia estadística.

IV. ANALISIS DE CONGLOMERADOS

4.1 El análisis de conglomerados

El análisis de conglomerados es una técnica del análisis multivariado que permite agrupar un conjunto de individuos o de variables en grupos (cluster) de acuerdo a ciertos criterios de distancia y similaridad fijados, de tal manera que cada grupo esté integrado por unidades homogéneas y los grupos entre sí sean muy heterogéneos. A diferencia de otras formas de análisis multivaríado (discriminante) los grupos no están definidos a priori y la conformación de los mismos tiene un carácter exploratorio. La cercanía o alejamiento entre las unidades de análisis o de variables se determina por ejemplo con la distancia euclídea[5/], esto condiciona el análisis a variables cuantitativas. De este modo dos unidades se consideran muy semejantes cuando menor es la distancia entre ellas.

El análisis de conglomerados (cluster) es una técnica multivariada que permite la conformación de grupos homogéneos de unidades de análisis o variables, según una medida de distancia o proximidad determinada.

4.2 El análisis de conglomerados y su relación con la ENAHO

A través de las encuestas de hogares se recogen los datos que permiten obtener los indicadores para cuantificar los problemas sociales como el analfabetismo, la pobreza, el ingreso, entre otros. Estos indicadores o variables se expresan en diferentes escalas : ordinales, interválicas y de razón lo que determinaría distintas formas de clasificación de las unidades de análisis (vivienda, hogar, individuo, etc.) en grupos excluyentes, según el tipo de escala considerada generándose tantas clasificaciones como variables se tengan. La herramienta analítica que integra los diferentes modos de clasificación de las unidades de análisis y variables y la conformación de grupos excluyentes, es el análisis de conglomerados (cluster).

4.3 Clasificación de los departamentos del Perú en función de las variables más relacionadas con el índice de desarrollo humano

En el siguiente caso práctico se clasifica los departamentos del Perú en función de un conjunto de variables relacionadas con el índice de desarrollo humano (IDH). El ordenamiento resultante no persigue fines analíticos comparativos constituyendo solamente un ejercicio didáctico, para demostrar la aplicación de la técnica.

[5/] Es la opción que por defecto proporciona el programa. Se define como la raíz cuadrada de la suma de diferencias al cuadrado entre dos elementos en la variable o variables consideradas $D(X,Y)=\sqrt{S(X_i - Y_i)^2}$.También es usualmente considerada el cuadrado de la distancia euclidea $D(X,Y)=S(X_i - Y_i)^2$

4.4 Etapas para realizar el análisis de conglomerados

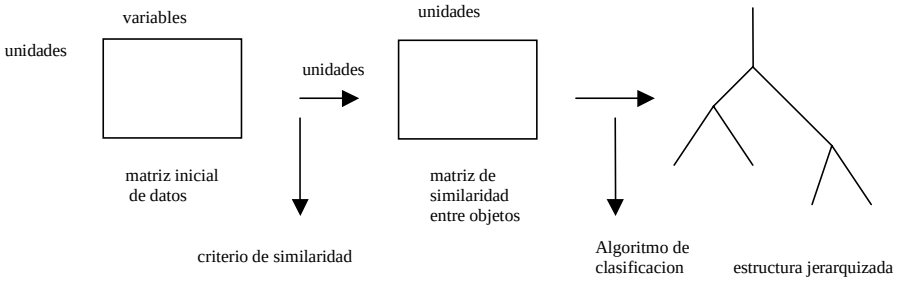
Para realizar el análisis de conglomerados siga las siguientes etapas:

1. **Formule las hipótesis:** Para plantear las hipótesis se formula la siguientes preguntas: ¿Cuántos grupos de departamentos se puede conformar de acuerdo a las variables relacionadas con el IDH? ¿Cuál es el departamento con mayor desarrollo humano relativo? ¿Cuál es el departamento con menor desarrollo humano relativo?

   Primera hipótesis: De acuerdo a las variables que explican el IDH los departamentos del Perú se agrupan en tres grupos bien diferenciados: los muy desarrollados, los medianamente desarrollados y los menos desarrollados.

   Segunda hipótesis: Lima es el departamento que tiene el mayor desarrollo humano relativo y Huancavelica es el de menor desarrollo.

2. **Genere un esquema de análisis:** Para realizar el análisis de conglomerados deben identificarse primero las variables/características que diferencian más a los grupos. Si el objetivo es formar conglomerados de individuos deben ubicarse las variables que presentan los valores más altos y más bajos. El análisis factorial explicado anteriormente ayuda en la selección de estas variables. Así, se escogerán las variables que integran cada factor y las más representativas de las dimensiones del problema social en estudio. El siguiente diagrama muestra los procesos seguidos para realizar un análisis de conglomerados.



   Tener presente que los conglomerados se generan a partir de las relaciones interdependientes entre las variables.

3. **Identifique la fuente de datos y la unidad de análisis:** La unidad de análisis es el departamento y los indicadores están referidas al período 1999. Cuando las fuentes de información proceden de muestras se recomienda utilizar los datos sin expandir, así se evitarán resultados que distorsionen la interpretación de los coeficientes estimados.

4. **Variables seleccionadas:** El análisis generalmente está restringido a las variables interválicas o de razón. Para el ejemplo las variables seleccionadas para conformar los conglomerados (cluster) son:

No	Variable	Tipo	Código
1	Esperanza de Vida	Interválica	ESPERANZ
2	Tasa de alfabetismo	Razón	ALFABETISM
3	Años Promedio de Estudio	Interválica	AÑOS_ PRO
4	Ingreso Promedio Mensual	Interválica	INGRESO_PR

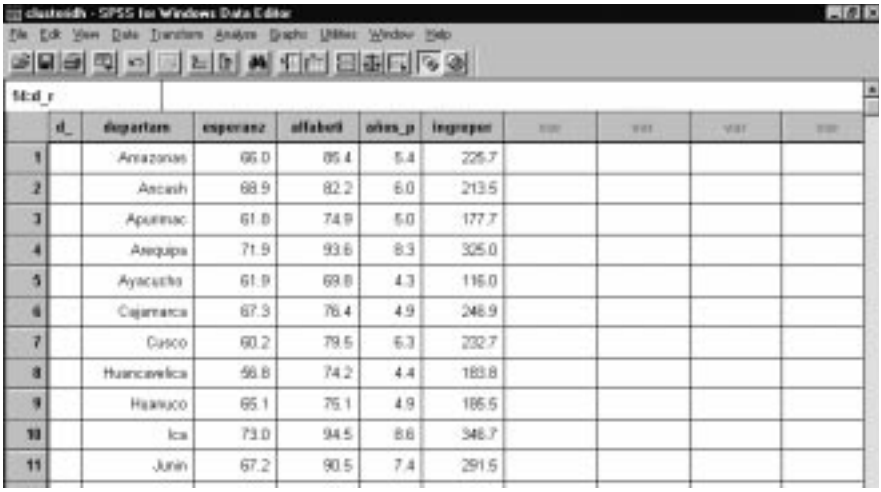
Las variables utilizadas en este caso práctico son las que intervienen en el calculo IDH. La esperanza de vida corresponde al periodo 1995-2000, mientras que el resto de variables se han obtenido a partir de la ENAHO 1999.

6. **Iniciando el análisis de conglomerados:** No se realizó el análisis factorial para la selección de variables debido a que el Programa de las Naciones Unidas para el Desarrollo (PNUD) determina un conjunto de indicadores como explicativos del desarrollo humano.

   El análisis se inicia ubicando en la base de datos los cuatro indicadores que explican el IDH a nivel departamental.

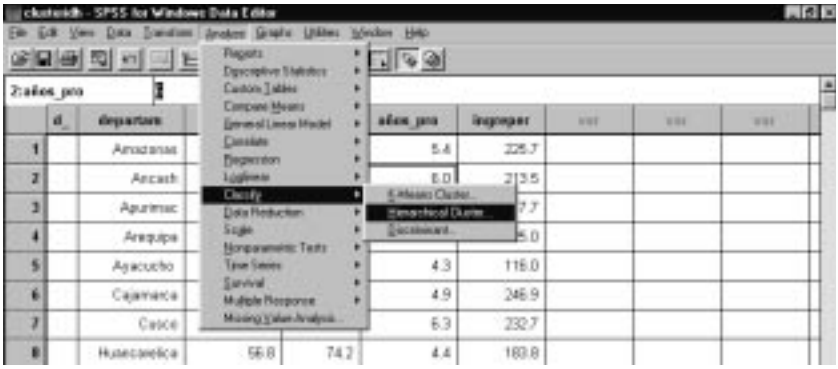
Con el análisis de conglomerados se conformaran grupos de departamentos a los cuales se denominarán cluster. Para la formación de los cluster se utilizan dos métodos: el Análisis Cluster Jerárquico o el K-Means Cluster. El primer método se utiliza cuando se dispone de una muestra relativamente pequeña de individuos. El segundo método se recomienda cuando se dispone de un tamaño de muestra grande y a priori se conoce el número de cluster. En el ejemplo, los 24 departamentos constituyen una muestra pequeña lo que determina aplicar en este caso el análisis cluster jerárquico.

A continuación se muestra la base de datos activa y las variables luego de la selección:

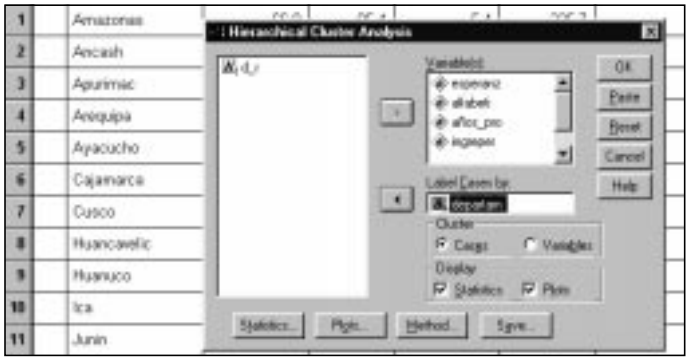


	cl_	departam	esperanz	alfabeti	años_p	ingreper
1		Amazonas	66.0	85.4	5.4	225.7
2		Ancash	68.9	82.2	6.0	213.5
3		Apurimac	61.8	74.9	5.0	177.7
4		Arequipa	71.9	93.6	8.3	325.0
5		Ayacucho	61.9	69.8	4.3	116.0
6		Cajamarca	67.3	76.4	4.9	248.9
7		Cusco	60.2	79.5	6.3	232.7
8		Huancavelica	56.8	74.2	4.4	183.8
9		Huanuco	65.1	75.1	4.9	186.5
10		Ica	73.0	94.5	8.6	346.7
11		Junin	67.2	90.5	7.4	291.5

En el menú de barras del SPSS elija la opción **Analyze/Classify/Hierarchical Cluster**.



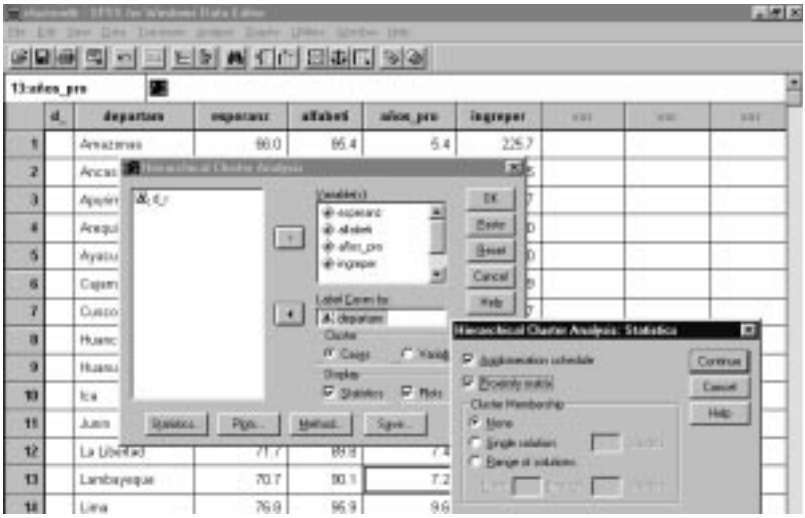
Pase al recuadro "variable(s)" los indicadores: Esperanza de vida, Alfabetismo, Años promedio de estudios, e Ingreso Per cápita. En el recuadro "Label cases by" pase la variable que identifica las unidades a clasificar "departam". Esta variable tiene que ser nominal. Luego en Cluster active la opción "Cases " y en Display las opciones Statistics y Plots " :



En la opción Cluster se selecciona "Cases" en lugar de "Variables", de tal manera que el análisis se efectúe a nivel de departamentos (casos). En caso contrario el análisis será a nivel de variables. En "Display" se habilita las opciones "Statistics" y "Plots". La primera permitirá calcular las estadísticas de tendencia central y de dispersión para cada conglomerado (cluster), mientras que la segunda generará los respectivos gráficos.

Ventana STATISTICS

En el subcuadro que se muestra, agregue la opción **"Proximity matriz"** a las seleccionadas por defecto.



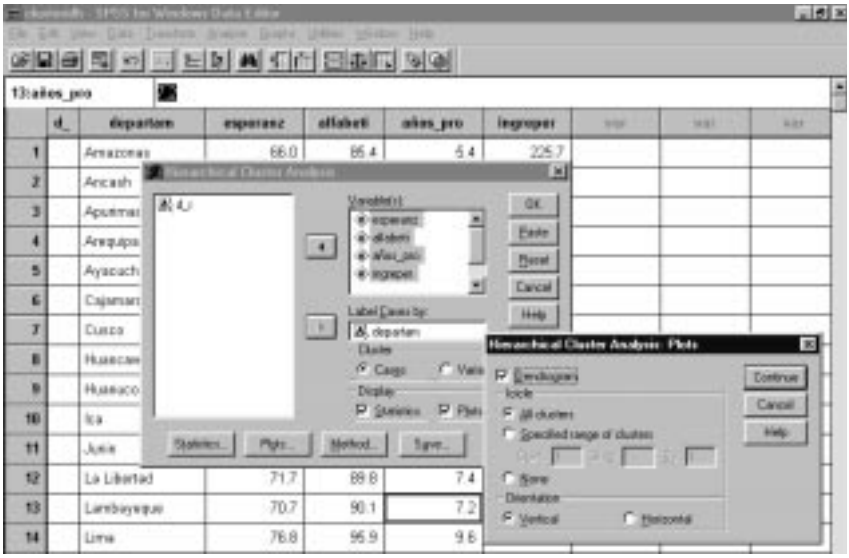
El **"Agglomeration schedule"** es un método acumulativo de formación de cluster. Consiste en formar primero tantos cluster como departamentos se tengan. Es decir se inicia con 24 cluster. Cuando se forma el primer cluster se van añadiendo otros, de tal manera que al final todos los cluster conformarán un solo grupo. Si un cluster se ha formado, ya no se desintegra hasta el final.

Con la opción **"Proximity matrix"** se muestra la matriz de distancias, que para el presente ejemplo se compone de la matriz de distancias euclideas al cuadrado. La opción **"Cluster membership"** permite mostrar la formación progresiva de los cluster, al inicio y en cada paso iterativo. Si selecciona "None" no se muestra nada. El "Single Solution", da la posibilidad de mostrar un número determinado de cluster. En "Range of solution" da la posibilidad de fijar un rango determinado de clusters en que desea dividir la muestra, que va desde dos hasta un número equivalente al total de unidades de análisis menos uno.

Pulse el botón "continue" para continuar la selección.

Ventana PLOTS

En **"Plots"** seleccione las siguientes opciones. En el subcuadro que se muestra, dejar todas las opciones seleccionadas por defecto y añada la opción "Dendograma". Seguidamente pulse continue.

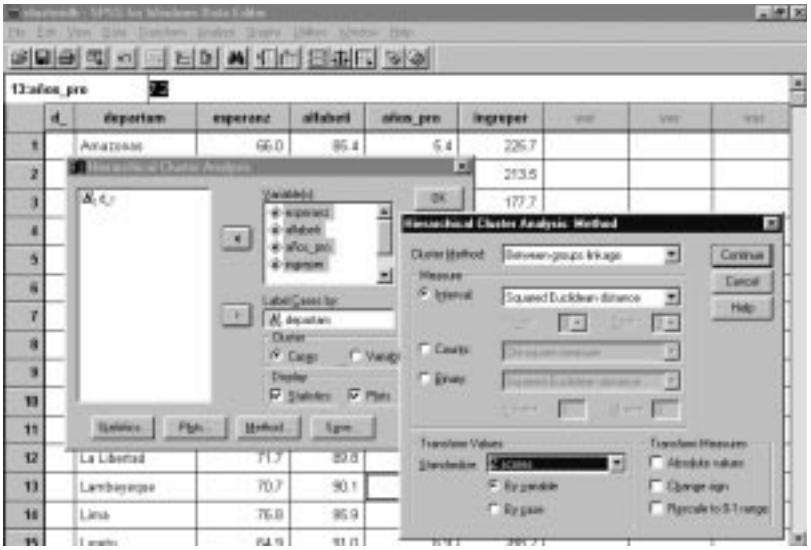


El "Dendogram" es una representación de los resultados del análisis cluster. Se lee de izquierda a derecha. Las líneas verticales dan a conocer la unión de dos cluster. En la parte superior se muestra la escala de distancias entre los diversos cluster (coeficientes), la cual ha sido reconvertida a otra escala de 0-25. La posición de la línea vertical en esta escala indica la distancia de unión de los cluster. El vertical icicle plot (gráfico vertical) que se configura al seleccionar las opciones "icicle all cluster" y "orientation vertical", permite representar gráficamente la conformación de los grupos. Al trazar una línea horizontal, en el gráfico "vertical icicle plot", se configura el número de grupos y los departamentos que lo integran.

Pulse continue para continuar la selección.

Ventana *METHOD*

Pulse el botón de la opción **Cluster Methods** del cuadro de diálogo principal de la figura. En la ventana que se muestra, dejar todas las opciones seleccionadas por defecto excepto la relacionada al procedimiento "Transform Values", en la cual debe seleccionar de la lista desplegable, la opción Z score. Pulse continue

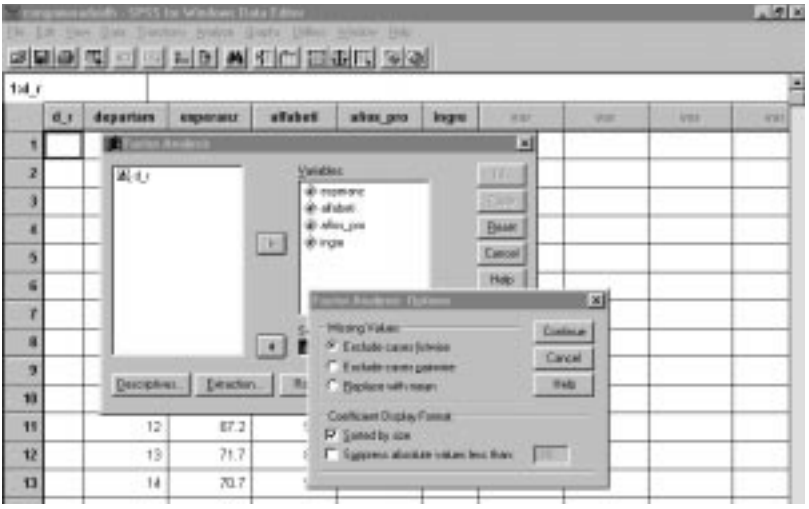


La opción "Measure" permite seleccionar la expresión para medir las distancias, la misma que estará en función al tipo de escala en que se han medido las variables: intervalo y de razón.

La opción "Transform Values" permite homogenizar los diferentes tipos de escala en que se han medido las variables. Así las unidades de distancia entre dos individuos será expresada en el mismo tipo de unidades (Z-scores)

Ventana OPTIONS

En la ventana Options, seleccione las siguientes funciones:



"Exclude cases listwise", permite excluir los valores incompletos. "Sorted by size" permite ordenar los coeficientes estimados en forma creciente.

Sin realizar ningún cambio en la ventana "save", en el cuadro de diálogo principal pulse OK. De este modo se ejecutara todo el procedimiento.

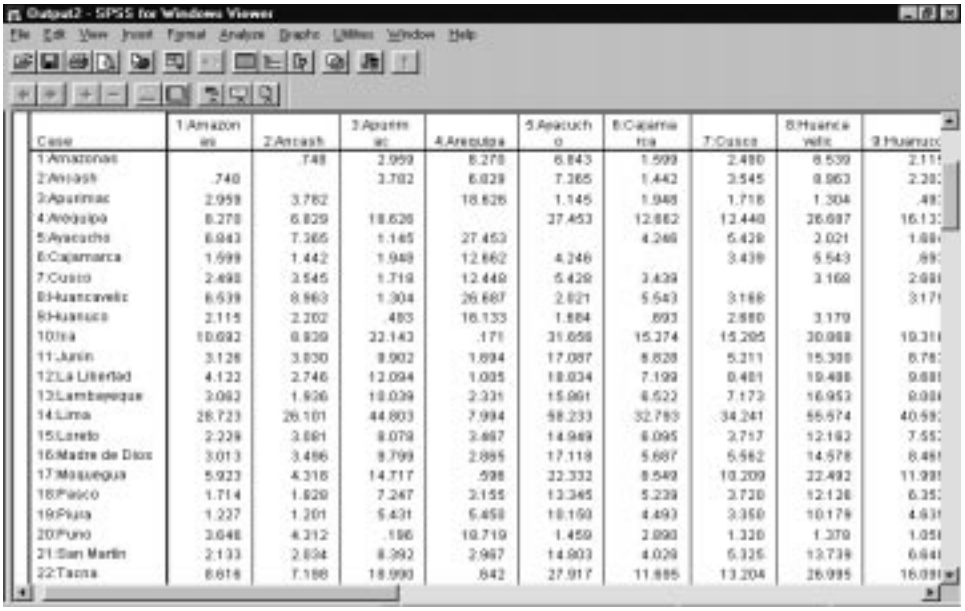
El agrupamiento del ejemplo desarrollado toma en cuenta las variables de tipo económico, social y demográfico lo cual determina una conformación específica de los departamentos, en los distintos grupos, probablemente muy diferente a la que se obtendría si el criterio de clasificación se hubiese realizado con variables de tipo cultural, político, ambiental o de salud, etc.

Así como se agrupan los departamentos según el criterio de distancia elegido, también existen muchos métodos para combinar los grupos. En el ejemplo se emplea el método aglomerativo, que consiste en un agrupamiento sucesivo en una serie de pasos. Al comienzo se tienen tantos grupos (cluster) como departamentos, en cada paso los cluster se van uniendo hasta que al final constituyen un solo grupo.

6. Interpretación de los resultados del análisis de conglomerados

Matriz de distancias

La matriz formada de orden 23 x 23 (total de departamentos menos uno), muestra las distancias euclideanas entre los departamentos. Un valor pequeño denota mayor semejanza en cambio un valor grande mayor diferencia.


Output2 - SPSS for Windows Viewer

File Edit View Insert Format Analyze Graphs Utilities Window Help

Case	1:Amazonas	2:Ancash	3:Apurimac	4:Arequipa	5:Ayacucho	6:Cajamarca	7:Cusco	8:Huancavelic	9:Huanuco
1:Amazonas		.748	2.959	8.278	6.843	1.599	2.480	6.539	2.11
2:Ancash	.748		3.782	8.628	7.365	1.442	3.545	6.963	2.20
3:Apurimac	2.959	3.782		18.626	1.145	1.948	1.718	1.304	.49
4:Arequipa	8.278	8.629	18.626		27.453	12.662	12.440	26.687	16.13
5:Ayacucho	6.843	7.365	1.145	27.453		4.246	5.429	3.021	1.68
6:Cajamarca	1.599	1.442	1.948	12.662	4.246		3.439	5.543	.89
7:Cusco	2.480	3.545	1.718	12.440	5.429	3.439		3.168	2.68
8:Huancavelic	6.539	6.963	1.304	26.687	3.021	5.543	3.168		3.17
9:Huanuco	2.115	2.202	.493	16.133	1.684	.893	2.680	3.179	
10:Ica	10.692	8.839	22.143	.171	31.656	15.374	15.295	30.988	19.31
11:Junin	3.128	3.830	8.902	1.694	17.087	8.828	5.211	15.388	8.78
12:La Libertad	4.122	2.746	12.094	1.035	18.034	7.199	9.481	19.488	9.88
13:Lambayeque	3.082	1.936	10.039	2.331	15.991	6.522	7.173	16.953	9.08
14:Lima	28.723	26.101	44.803	7.994	58.233	32.783	34.241	55.574	40.59
15:Loreto	2.229	3.881	8.079	3.467	14.969	8.095	3.757	12.182	7.55
16:Madre de Dios	3.013	3.486	9.799	2.895	17.118	5.687	5.562	14.578	8.48
17:Moquegua	5.923	4.316	14.717	.598	22.332	8.549	10.209	22.492	11.99
18:Pasco	1.714	1.829	7.247	3.155	13.345	5.239	3.720	12.126	6.35
19:Piura	1.227	1.201	5.431	5.458	10.150	4.493	3.350	10.179	4.93
20:Puno	3.648	4.312	.186	10.719	1.459	2.890	1.320	1.379	1.05
21:San Martin	2.133	2.634	8.392	2.987	14.903	4.029	5.325	13.739	6.64
22:Tacna	8.618	7.188	18.990	.842	27.917	11.885	13.204	26.985	16.08

Tabla de aglomeración

En esta tabla se muestra el número de grupos o cluster que se combinan en cada estado (Stage). La primera línea corresponde al estado stage = 1 . En este nivel se combinan el cluster 4 (Arequipa) con el cluster 10 (Ica) quedando 23 cluster. La distancia euclidea (coefficients) entre estos cluster es 0.171. La última columna (next stage) indica en que estado se incorpora un nuevo cluster (departamento) a esta primera unión. En el ejemplo, es el estado 9. Si se verifica en la columna stage = 9 , ubica el número 22 que corresponde al departamento de Tacna, quedando formado un nuevo cluster con los departamentos de Arequipa, Tacna e Ica. La columna "stage cluster first appears" indica en que nivel se dio la primera formación de cluster. De este modo la lectura del "aglomeration schedule" permite realizar un seguimiento a la conformación de los grupos. Cuanto menor sean los coeficientes, implica mayor homogeneidad entre los cluster.

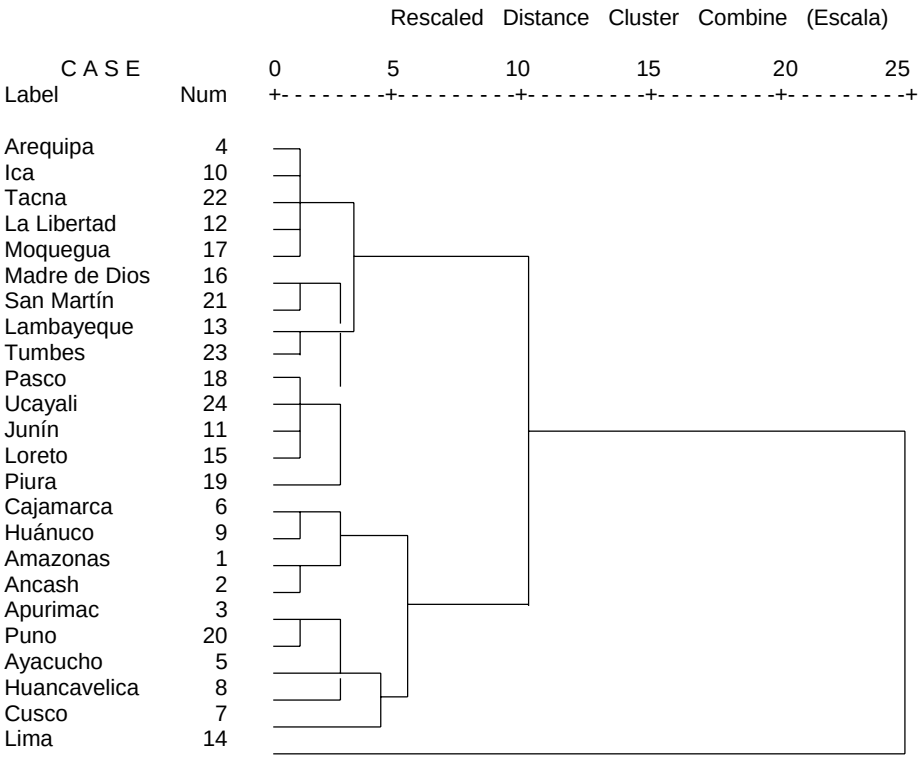
Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	4	10	.171	0	0	9
2	3	20	.196	0	0	14
3	13	23	.246	0	0	15
4	18	24	.273	0	0	7
5	16	21	.291	0	0	18
6	11	15	.370	0	0	7
7	11	18	.399	6	4	13
8	12	17	.438	0	0	12
9	4	22	.647	1	0	12
10	6	9	.693	0	0	17
11	1	2	.748	0	0	17
12	4	12	1.101	9	8	20
13	11	19	1.201	7	0	15
14	3	5	1.302	2	0	16
15	11	13	1.415	13	3	18
16	3	8	1.568	14	0	19
17	1	6	1.839	11	10	21
18	11	16	2.035	15	5	20
19	3	7	2.908	16	0	21
20	4	11	3.069	12	18	22
21	1	3	3.880	17	19	22
22	1	4	10.567	21	20	23
23	1	14	24.964	22	0	0

Dendograma

El dendograma es una representación gráfica de la conformación de los conglomerados o cluster. La línea horizontal superior indica la escala a la cual se combinan los cluster. Las líneas verticales indican la conformación de los departamentos en cluster. Cuanto los conglomerados se encuentren más próximo a cero en la escala (rescaled distance cluster combine) implica que las agrupaciones son más fuertes. Así tenemos Arequipa. Ica, Tacna, La libertad y Moquegua conforman el primer cluster cuya agrupación por estar próxima a cero en la escala indica cercanía en cuanto a las características de desarrollo estudiadas. En cambio Lima se encuentra muy alejado del resto de conglomerados, por cuanto sus indicadores denotan mayor desarrollo y por tanto mayor distancia.

* * * * * * HIERARCHICAL CLUSTER ANALYSIS * * * * * * *
Dendrogram using A verage Linkage (Between Groups)



7. Principales conclusiones del análisis factorial realizado

Los departamentos del Perú se han dispuesto en tres grupos bien diferenciados de acuerdo a su mayor relación con las variables consideradas, comprobándose la primera hipótesis formulada. Así, Lima con mejores indicadores de desarrollo humano y muy alejada del resto conforma un grupo, el segundo grupo y con un desarrollo humano relativo mediano está integrado por: Arequipa, Ica, Tacna, La Libertad, Moquegua, Madre de Dios, San Martín, Lambayeque, Tumbes, Pasco, Ucayali, Piura, Junín, y Loreto. El tercer grupo, con el menor desarrollo relativo está conformado por: Cajamarca, Huanuco, Amazonas, Ancash, Apurímac, Puno, Ayacucho, Cusco y Huancavelica.

También se comprueba que Huancavelica forma parte del grupo de departamentos con menor desarrollo relativo, comprobándose la segunda hipótesis.

8. Consistencia del modelo y los datos

En cuanto al modelo

En la **matriz de distancias** se muestran los coeficientes que permiten determinar la semejanza entre la unidades de análisis. Un valor alto significa que las unidades no se parecen, mientras que los valores pequeños estarán asociados a las unidades con características semejantes. En el ejemplo desarrollado, los departamentos con características sociales diferentes mostrarán los coeficientes de distancia más altos, mientras que los valores pequeños estarán asociados a departamentos con características muy semejantes. Así, entre Ica y Arequipa el coeficiente de distancia es 0.171 lo que denota su semejanza, mientras que entre Lima y Amazonas el coeficiente de distancia es 28.7. La consistencia del modelo queda determinada cuando los coeficientes de distancia hallados tienen relación con las diferencias encontradas entre las variables originales.

La tabla de aglomeración: En esta tabla pueden estudiarse paso a paso la formación de los grupos de unidades de análisis, esperando encontrar en el ejemplo desarrollado que las primeras uniones se den entre aquellos departamentos con desventajas sociales relativas muy parecidas. Si esta condición se verifica entonces podrá inferirse la efectividad del modelo.

El dendograma: En este gráfico se aprecia de manera global la disposición de los departamentos.

También la conformación de los diferentes grupos de acuerdo a su desarrollo relativo. En el dendograma, se observa que Lima se separa del resto de departamentos, por cuanto tiene los mejores indicadores de desarrollo humano, mientras que Cajamarca, Huanuco, Amazonas, Ancash, Apurímac, Puno, Ayacucho, Huancavelica y Cusco forman un conglomerado con una desventaja relativa mayor en relación al resto de departamentos.

En cuanto a la consistencia de los datos

La **matriz de distancias** sirve para realizar comparaciones entre los departamentos. Así, Apurímac y Ayacucho tienen un

coeficiente de distancia de 1.14, lo que confirma las diferencias mostradas en los valores de las variables originales y que se utilizan como criterios de clasificación (Esperanza de vida, Años promedio de estudios, Alfabetismo, e Ingreso Per cápita). Mientras que Ayacucho y Arequipa tienen un coeficiente de distancia de 27. 45 y entre Apurímac y Arequipa este coeficiente es 18.6. Así, Apurímac y Ayacucho son más parecidos entre sí, en cambio Arequipa dados los valores de los coeficientes de distancia no se parece a ninguno. Puede concluirse que los datos confirman los supuestos teóricos con respecto a las diferencias relativas en el desarrollo de los departamentos reflejados en los valores de las variables. Esta proximidad entre los resultados observados y los esperados es evidencia de la consistencia de los datos de la encuesta. En caso contrario la base de datos debe estudiarse nuevamente.

La tabla de aglomeración: Esta tabla puede ser útil para identificar aquellos departamentos con las variables que tienen datos muy alejados de la tendencia mostrada por los valores promedio del grupo al que pertenece. En estos casos se recomienda revisar en la base de datos dicha variables. En el ejemplo desarrollado las distancias entre los cluster que se muestran en la columna "coefficients" de la tabla "Agglomeration Schedule" sirven para realizar este tipo de comparaciones. Así entre el departamento de Arequipa, Ica y Tacna identificados en esta tabla con los dígitos 4, 10 y 22, se espera valores en las variables originales cercanos.

El dendograma: contribuye también a la evaluación de los datos, por cuanto la disposición y agrupamiento entre los departamentos obedece a los valores que toman entre las variables. De este modo a simple vista pueden ubicarse los departamentos con los coeficientes de distancia muy próximos entre sí, cuyos valores en las variables originales se espera también lo sean.

V. ANALISIS DISCRIMINANTE CLASIFICATORIO

5.1 El análisis discriminante

El análisis discriminante es otra técnica del análisis multivariado que permite **clasificar las unidades de análisis** en grupos definidos a priori y **analizar las causas** que han dado lugar a la formación de estos grupos. Los grupos se conforman a partir de un conjunto de variables seleccionadas (variables independientes), las cuales además de explicar la formación de los grupos pueden ordenarse según su mayor poder discriminatorio. De este modo las unidades de análisis son reclasificadas, generándose una clasificación distinta a la realizada a priori, donde el aporte de las variables independientes (explicativas) a la formación de los grupos determina la formación de los mismos. El algoritmo matemático **(función discriminante)** se determina a partir de la escala de medida de la variable dependiente: si la variable dependiente es dicotómica (discreta) entonces el modelo matemático más apropiado es el logístico; si es continua, entonces el modelo que mejor se adecua es el modelo de regresión lineal simple o múltiple.

El análisis discriminante es otra técnica del análisis multivariado que permite asignar las unidades de análisis a grupos a priori conformados a partir de un conjunto de variables explicativas que contribuyen más a la formación de los grupos .

El análisis discriminante clasificatorio

El análisis discriminante clasificatorio se aplica cuando se busca conformar grupos mutuamente excluyentes de unidades de análisis a partir de un conjunto de variables explicativas (independientes), estos nuevos grupos muy probablemente difieren de los conformados a priori.

5.2 El análisis discriminante y su relación con la ENDES

La encuesta demográfica y de salud familiar ENDES contiene un conjunto de información relacionada con las características demográficas de la mujer en edad fértil y de sus hijos menores de cinco años, además de datos relacionados con la vivienda y el hogar. El estudio puede estar referido a diferentes unidades de análisis: el hogar, la mujer adolescente, la mujer adulta mayor, entre otras. Estas unidades de análisis pueden disponerse en grupos según las características de las variables que las integran. La ENDES considera generalmente variables cualitativas.

Las variables de la ENDES medidas en escalas diferentes (nominales, ordinales e interválicas), determinarán diversos modos de clasificación no pudiéndose determinar cual de las variables influye más en la

conformación de los grupos de estudio. De allí la necesidad de aplicar otra técnica multivariada como el análisis discriminante que permita clasificar las unidades de análisis medidas en diferentes escalas e identificar aquellas variables más influyentes en la conformación de los grupos.

5.3 Estudio del número de hijos de las mujeres según variables seleccionadas

A partir del modelo discriminante y la relación de dependencia establecida se busca encontrar las causas que determinan la tenencia de hijos por las mujeres en edad fértil.

5.4 Etapas para realizar el análisis discriminante

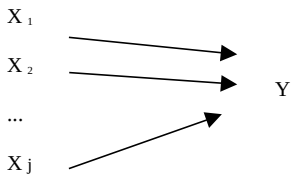
Para realizar el análisis discriminante siga las siguientes etapas:

1. **Formule las hipótesis:** Para plantear las hipótesis se formulan las siguientes preguntas: ¿El número de hijos de las mujeres esta determinado por el nivel educativo? ¿Qué variables contribuyen más a distinguir entre las mujeres con hijos y sin ellos? A partir de estas preguntas se pueden formular las hipótesis siguientes:

   Primera hipótesis: El nivel educativo determina el número de hijos en las mujeres.

   Segunda hipótesis: La edad, la condición de ocupación y el lugar de residencia son las variables que contribuyen más a diferenciar entre las mujeres con hijos y sin ellos.

2. **Genere un modelo multivariado:** El modelo multivaríado se conforma a partir de una ecuación lineal. A fin de alcanzar mayor precisión en las estimaciones se recomienda efectuar primero, un análisis factorial para clasificar las variables de acuerdo a las dimensiones del objeto social en estudio. Generalmente la variable dependiente es cualitativa mientras las independientes son cuantitativas.



Y : variable dependiente **X_i : variables independientes**

Modelo discriminante clasificatorio:

$$Y_{ji} = a X_{1i} + b X_{2i} + c X_{3i} + ... + m X_{ji}$$

Donde Y ij : puntaje discriminante del i-ésimo individuo correspondiente a la j-ésima variable

3. **Fuente de información y unidad de análisis:** La fuente de información para el presente ejemplo aplicativo es la base de datos de la encuesta demográfica y de salud familiar ENDES 2000. Se recomienda utilizar los datos sin expandir de modo que las estimaciones no se vean afectadas por las ponderaciones. La unidad de análisis es la mujer en edad fértil (15 a 49 años).

4. **Variables seleccionadas:** El análisis discriminante requiere que la variable para definir los grupos sea cualitativa mientras que las variables discriminantes deberán ser cuantitativas (intervalares),

en caso contrario se recomienda convertirlas en variables dicotómicas (0-1). Tener presente que el valor de uno debe esta asociado a la presencia de la cualidad. Así, por ejemplo las categorías de la variable estado conyugal pueden hacerse dicotómicas. El valor 0 indica no unida mientras que el valor 1 indica unida.

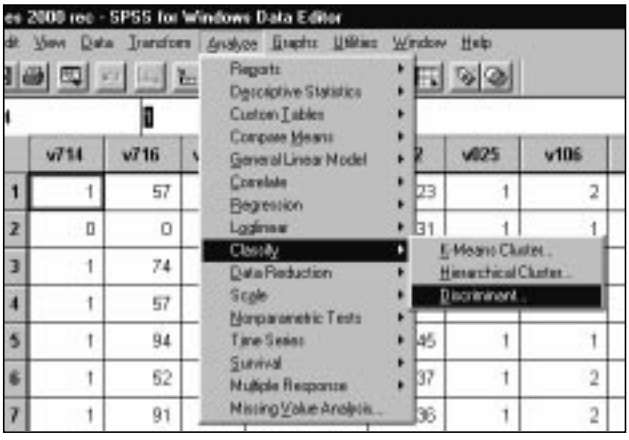
No	Variable	Tipo	Código
1	Edad	Interválica	V012
2	Lugar de Residencia	Nominal-Dicotómica	V0251
3	Educación Básica Alcanzada	Nomiinal- Dicotómica	V1061
4	Estado Conyugal	Nominal-Dicotómica	V5011
5	Condición de Actividad	Nominal-Dicotómica	RECV7171

Variable Dependiente: mujer en edad fértil sin hijos, con uno a dos hijos y con tres o mas hijos (REV201).

5. Procedimientos para realizar el análisis discriminante

Para proceder a realizar el análisis discriminante con el SPSS siga los siguientes pasos:

1. Cargue la base de datos con las variables seleccionadas de la fuente de información mencionada



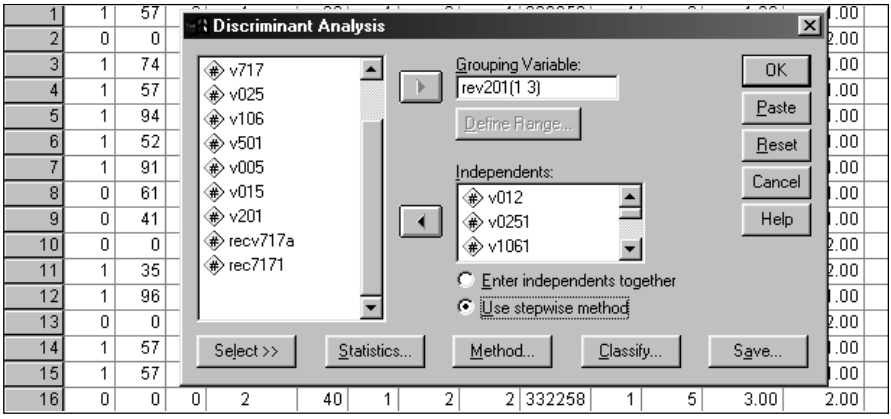
Iniciando el análisis discriminante: Ingrese a la siguiente ventana del SPSS Analyze / Classify/ Discriminant. La opción Classify permite clasificar la información de acuerdo a las opciones K-means Cluster y Hierarchical Cluster, explicadas en la sección anterior. A estas dos se añade la opción discriminante, seleccione de acuerdo a la ventana mostrada. Una vez que ha ingresado al análisis discriminante realice lo siguiente.

En el campo Grouping Variable, ingrese la variable dependiente que en el ejemplo es el número de hijos (REV201). Esta variable tiene tres categorías (1 = sin hijos, 2 = 1 y 2 hijos, 3 = 3 o mas hijos). En "Define Range" especifique el valor mínimo y máximo de la variable dependiente. En el campo Independents, ingrese las variables independientes. La opción seleccionada por defecto es "enter independents together", mediante la cual el software evalúa todas las variables independientes al mismo tiempo. La opción alternativa es el "stepwise method" (método stepwise) , mediante el cual las variables explicativas son evaluadas una por una y salen de acuerdo a ciertos criterios. Las dos formas son válidas quedando la elección a criterio del investigador. Si elige el "stepwise", se presentan algunas opciones adicionales.

Las siguientes opciones son el "Select", "Statistics", "Classify", "Save" y "Method"(en caso de optar por el stepwise), las cuales se explican detalladamente:

Ventana SELECT

Esta opción se utiliza cuando se trabajan con grupos de muestras (submuestras). Por ejemplo podemos analizar si el número de hijos en las mujeres jóvenes se explica por las variables seleccionadas. En la caja de diálogo "selection variable", incluya la variable dependiente. En el desarrollo del presente ejemplo no se ha considerado esta opción.



Ventana STATISTICS

En el cuadro de diálogo, seleccione todas las opciones, tal como se indica, y pulse continue. Estas opciones permiten realizar lo siguiente:

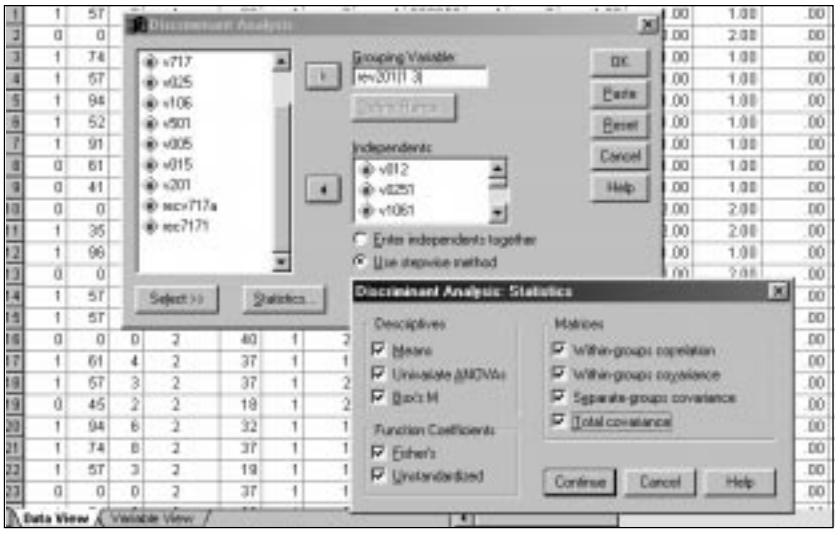
La opción "Descriptive", calcula el promedio y la desviación estándar para las variables independientes en cada grupo.

La opción "Univariate ANOVA's", calcula la significación estadística de las diferencias entre los promedios de los grupos conformados.

La opción "Box M", evalúa las diferencias entre las matrices de covarianza.

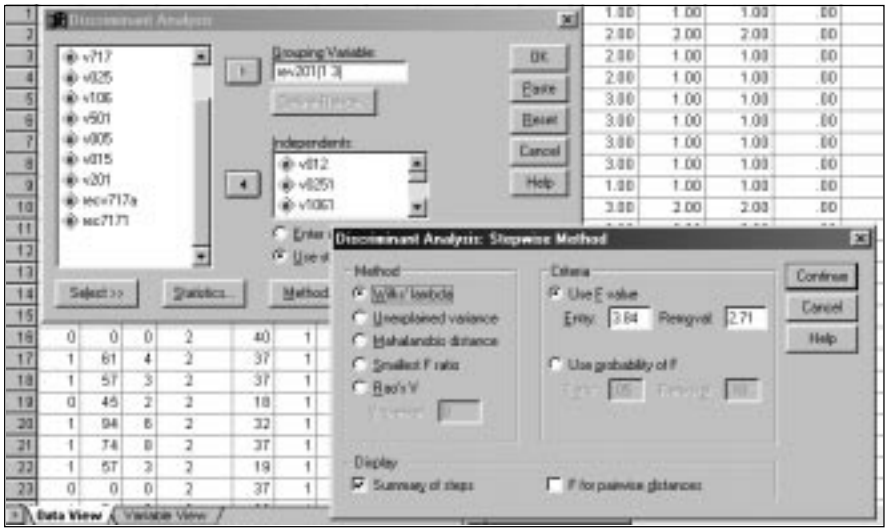
La opción "Function Coefficients: Fisher y Unstandardized", permite el cálculo de los puntajes discriminantes de Fisher y no estandarizados.

La opción "Matrices", permite calcular todas las matrices de correlación y covarianza intragrupos, por cada grupo y total. En el cuadro de diálogo, seleccione todas las opciones, tal como se indica.



Ventana METHOD

Este cuadro de diálogo estará activo si se selecciona la opción Stepwise. En caso de hacerlo deje las opciones seleccionadas por defecto: en Method deje el "Lambda de Wilks", en Display active "summary of steps", en Criteria "Use F value" . Seguidamente pulse continue.



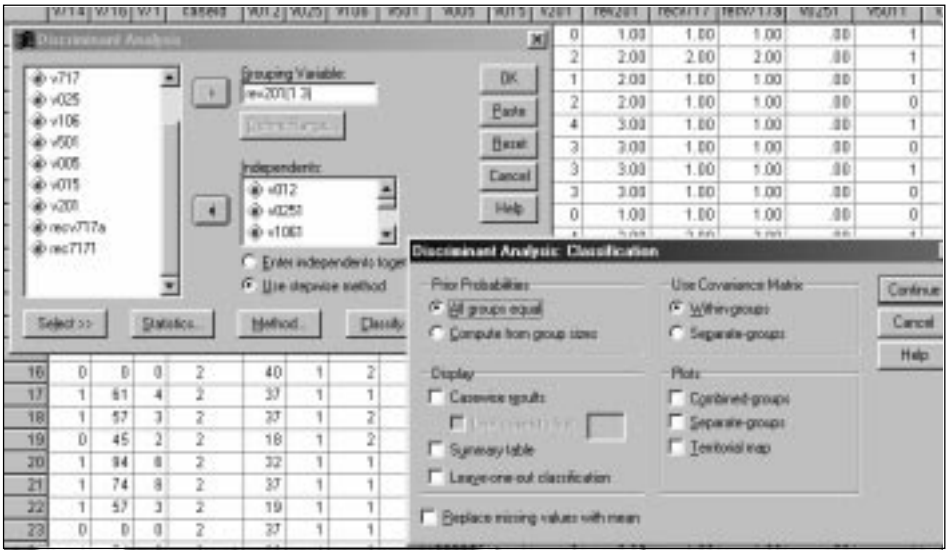
El "Lambda de Wilks" calculado es un valor numérico que sirve para decidir el ingreso o salida de las variables en el modelo. Estos valores van acompañados de los "F value" y se interpretan de manera inversa a los lambda de Wilks. Así, un mayor F implica un mayor poder discriminatorio de la variable. Este proceso iterativo de selección se muestra cuando se activa "summary of steps".

Ventana CLASSIFICATION

En esta ventana deben definirse las características de la clasificación. En Prior probabilities, se tienen dos opciones: "All groups equal" que quiere decir "todas las unidades de análisis ingresan con igual probabilidad a priori" y "Compute from group sizes" que significa "probabilidades de ingreso diferente para cada unidad de análisis de acuerdo al tamaño del grupo a priori conformado". En Display se muestran los resultados de la clasificación. Con "casewise results" se muestran todos, y con "limit cases to first", se fija el número de casos a mostrar. "Summary table"y "Leave-one-out classification" muestra la tabla y los resultados de la clasificación.

En "Use Covariance matrix" se muestran las matrices de covarianza entre grupos y de cada grupo por separado. El "Plots"permite representar gráficamente las unidades de análisis en los grupos, separadamente y en un mapa territorial, en el cual se muestra su nueva ubicación de acuerdo al análisis realizado.

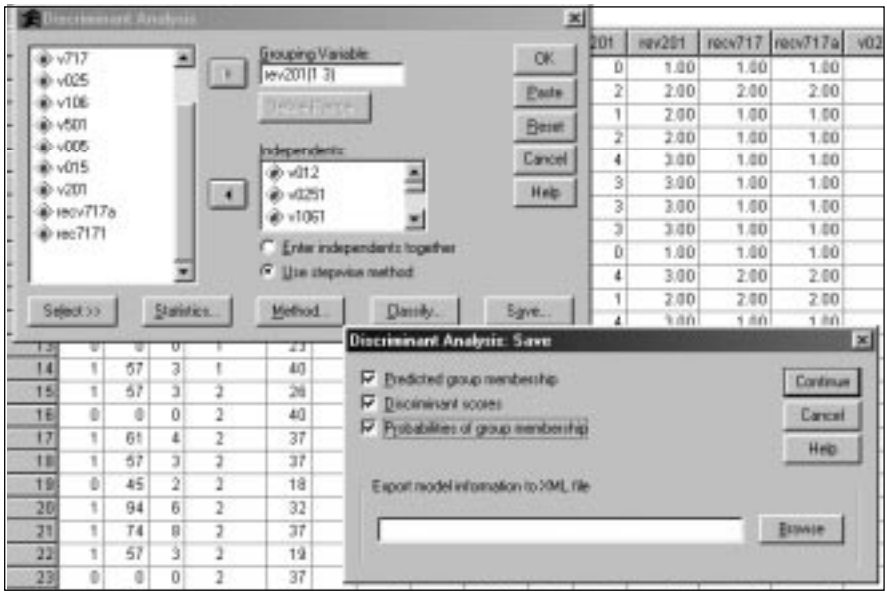
Ejecutamos el proceso con continue.



Ventana SAVE

Seleccionamos todas las opciones, pulsamos continue.

La opción Discriminant Score, muestra los puntajes del análisis discriminante. Estos puntajes se estiman al reemplazar el valor de las variables en la ecuación discriminante correspondiente.



Cuando en el cuadro de diálogo principal, se selecciona OK debe ejecutarse el análisis discriminante con las opciones seleccionadas.

6. Interpretación de los resultados del análisis discriminante

Las unidades de análisis originalmente dispuestas en tres grupos según el número de hijos han sido reclasificadas de acuerdo a sus valores en las variables explicativas consideradas. Los resultados (output) del modelo discriminante aplicado son:

Medidas descriptivas

La calidad del análisis realizado se determina mediante las medidas descriptivas como: el promedio y la desviación estándar de cada grupo. Se espera que los promedios (mean) de cada grupo sean diferentes, mientras que los coeficientes de variación que resultan de dividir la desviación estándar (std. deviation) y el promedio sean semejantes. Si las medias son diferentes, se comprueba que los grupos conformados difieren entre sí, justificándose el análisis realizado. Estas comparaciones serán posibles en la medida que se compruebe también la homogeneidad de las varianzas.

Group Statistics

REV201 Nº DE HIJOS		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
1.00 NO TIENE	V012 Current age - respondent	20.9106	6.5638	8912	8912.000
	V0251	.2819	.4499	8912	8912.000
	V1061	.8209	.3834	8912	8912.000
	V5011	8.640E-02	.2810	8912	8912.000
	RECV717 CONDTRABAJO	1.5252	.4994	8912	8912.000
2.00 1 A 2	V012 Current age - respondent	28.6531	7.6882	8450	8450.000
	V0251	.3515	.4775	8450	8450.000
	V1061	.6923	.4616	8450	8450.000
	V5011	.7776	.4159	8450	8450.000
	RECV717 CONDTRABAJO	1.4460	.4971	8450	8450.000
3.00 3 A MAS HIJOS	V012 Current age - respondent	37.2453	6.9774	10481	10481.000
	V0251	.5047	.5000	10481	10481.000
	V1061	.3529	.4779	10481	10481.000
	V5011	.9248	.2637	10481	10481.000
	RECV717 CONDTRABAJO	1.3462	.4758	10481	10481.000
Total	V012 Current age - respondent	29.4093	9.8214	27843	27843.000
	V0251	.3869	.4870	27843	27843.000
	V1061	.6057	.4887	27843	27843.000
	V5011	.6118	.4874	27843	27843.000
	RECV717 CONDTRABAJO	1.4338	.4956	27843	27843.000

La eficacia del análisis discriminante realizado se verifica cuando los puntajes promedio (mean) de cada grupo (1; 2; 3), son lo más diferentes entre sí, mientras que las desviaciones estándar (std. deviation) son mínimas. El número de casos analizados sin ponderar y ponderados se muestra en las dos últimas columnas (Valid N -listwise).

Matriz de covarianza

Existen tantas matrices de covarianza como grupos de la variable dependiente se dispongan. Cuando se analizan el signo y la magnitud de los coeficientes de la matriz de covarianza ésta se convierte en una medida de calidad de las estimaciones. La semejanza entre los valores calculados para cada grupo, será también un indicador de la consistencia de las estimaciones.

Covariance Matrices[a]

REV201 Nº DE HIJOS		V012 Current age - respondent	V0251	V1061	V5011	RECV717 CONDTRABAJO
1.00 NO TIENE	V012 Current age - respondent	43.083	-.311	-7.713E-03	.395	-.343
	V0251	-.311	.202	-7.139E-02	9.198E-03	-2.911E-02
	V1061	-7.713E-03	-7.139E-02	.147	-1.101E-02	3.897E-02
	V5011	.395	9.198E-03	-1.101E-02	7.894E-02	-4.538E-03
	RECV717 CONDTRABAJO	-.343	-2.911E-02	3.897E-02	-4.538E-03	.249
2.00 1 A 2	V012 Current age - respondent	59.109	-.956	.489	3.565E-02	-3.535E-02
	V0251	-.956	.228	-9.624E-02	2.655E-03	-2.186E-02
	V1061	.489	-9.624E-02	.213	-1.202E-03	3.405E-02
	V5011	3.565E-02	2.655E-03	-1.202E-03	.173	2.712E-02
	RECV717 CONDTRABAJO	-3.535E-02	-2.186E-02	3.405E-02	2.712E-02	.247
3.00 3 A MAS HIJOS	V012 Current age - respondent	48.684	-.450	-.101	-6.352E-02	-.132
	V0251	-.450	.250	-.102	1.133E-02	-2.215E-02
	V1061	-.101	-.102	.228	-5.334E-03	3.183E-02
	V5011	-6.352E-02	1.133E-02	-5.334E-03	6.954E-02	1.238E-02
	RECV717 CONDTRABAJO	-.132	-2.215E-02	3.183E-02	1.238E-02	.226
Total	V012 Current age - respondent	96.460	8.237E-02	-1.243	2.428	-.680
	V0251	8.237E-02	.237	-.110	3.778E-02	-3.138E-02
	V1061	-1.243	-.110	.239	-6.752E-02	4.975E-02
	V5011	2.428	3.778E-02	-6.752E-02	.238	-1.365E-02
	RECV717 CONDTRABAJO	-.680	-3.138E-02	4.975E-02	-1.365E-02	.246

a. The total covariance matrix has 27842 degrees of freedom.

Las matriz de covarianza **(covariance matrices)** en general, es una matriz simétrica cuya diagonal principal contiene las varianzas, mientras que por encima y debajo de esta diagonal se muestran las covarianzas. En el ejemplo se aprecia cierta semejanza entre las valores correspondientes a las diagonales en cada grupo conformado.

Lambda de Wilks

La suma de cuadrados de la variación total se descompone en una suma de cuadrados intra-grupo y una suma de cuadrados entre-grupos. El Lambda de Wilks, es una medida de calidad, que se forma a partir de la relación entre la suma de cuadrados dentro de grupos y la suma de cuadrados total. Así, si el Lambda de Wilks es uno quiere decir que toda la variación se explica por la variación dentro de grupos y no hay diferencia entre los grupos. En cambio, cuanto más cerca de cero este el Lambda, implica que la diferencia entre los grupos es mayor, lo que significa que las variables son adecuadas para construir las funciones discriminantes. Las variables con menor Lambda de Wilks son las más discriminantes. En el cuadro "variables Entered-Removed" se muestra el ingreso/ salida de las variables.

Variables Entered/Removed[a,b,c,d]

			Wilks' Lambda							
							Exact F			
Step	Entered	Removed	Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	V5011		.438	1	2	27840.000	17893.902	2	27840.000	.000
2	V012 Current age - respondent		.305	2	2	27840.000	11287.309	4	55678.000	.000
3	V1061		.279	3	2	27840.000	8287.134	6	55676.000	.000
4	V0251		.277	4	2	27840.000	6253.582	8	55674.000	.000
5	RECV717 CONDTRABAJO		.276	5	2	27840.000	5026.355	10	55672.000	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 10.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

Los Lambda de Wilks calculados permiten inferir que todas las variables deben ingresar al modelo, lo cual se deba probablemente al gran tamaño de muestra considerado.

Al realizar el análisis de varianza (descomposición de la variación total) para cada una de las variables por separado encontramos que todas las variables muestran diferencias significativas (prueba F), lo cual es atribuible al tamaño de muestra grande. Como se esperaba el nivel educativo (v1061), el lugar de residencia (v0251), la condición de actividad (RECV717), la edad (V012) y el estado conyugal (V501) tienen los menores Lambda de Wilks y por tanto explican mejor la variabilidad entre los grupos de mujeres (las que no tienen hijos, las mujeres con uno y dos hijos y las que tienen tres hijos).

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.276	35819.352	10	.000
2	.892	3194.053	4	.000

Wilks' Lambda

Step	Number of Variables	Lambda	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	1	.438	1	2	27840	17893.902	2	27840.000	.000
2	2	.305	2	2	27840	11287.309	4	55678.000	1.898E-14
3	3	.279	3	2	27840	8287.134	6	55676.000	.000
4	4	.277	4	2	27840	6253.582	8	55674.000	.000
5	5	.276	5	2	27840	5026.355	10	55672.000	.000

El Lambda de Wilks para los grupos muestra la efectividad del análisis realizado. Las funciones discriminantes 1 y 2 son estadísticamente significativas, como lo muestra la prueba estadística Jí-Cuadrado (Chi-square).

Prueba M de Box

¿Existe diferencias significativas entre los grupos conformados? ¿Son las matrices de varianza y covarianza de cada grupo estadísticamente significativas? La prueba M de BOX, es un indicador que permite responder estas interrogantes. El M de Box de 3,997.2 determina un valor F alto. Bajo la hipótesis nula que no hay diferencias significativas, se analizan los valores del F calculado (F) y el sig (nivel de significación). En la tabla "test results", si el sig es inferior a 0.01 entonces se rechaza la hipótesis nula y se concluye que los grupos conformados difieren significativamente.

Test Results

Box's M		3997.235
F	Approx.	133.207
	df1	30
	df2	2315436188.341
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

Se encuentran diferencias significativas entre las matrices de varianza y covarianza de cada grupo. La F = 133 y el grado de significación (sig. p = 0.00) así lo señalan. Los grupos de mujeres con hijos conformadas son los adecuados.

Funciones discriminantes

¿Es concordante la clasificación de las unidades de análisis a partir de los puntajes discriminantes con la clasificación a priori?

El análisis discriminante permite calcular las funciones discriminantes, para determinar el puntaje discriminante con el cual se clasifican las unidades de análisis.

Canonical Discriminant Function Coefficients

	Function	
	1	2
V012 Current age - respondent	.087	-.081
V0251	.183	-.254
V1061	-.506	1.151
V5011	2.190	2.152
RECV717 CONDTRABAJO	-.155	-.160
(Constant)	-3.450	.706

Unstandardized coefficients

Con estos coeficientes se determinan las siguientes ecuaciones discriminantes:

D1= - 3.45 + 0.08 V012 + 0.18 V0251 - 0.5 V1061 + 2.19 V5011 - 0.15 RECV7171

D2= 0.7 - 0.08 V012 - 0.25 V0251 + 1.15 V1061 + 2.15 V5011 - 0.16 RECV7171

Los valores correspondientes de las unidades de análisis en las variables deben sustituirse en las ecuaciones de modo que se obtengan los puntajes discriminantes. Cuando se ejecuta todo el procedimiento y habiendo seleccionado la opción save **("discriminant score")** estos puntajes se muestran en la última columna de la base de datos . En el ejemplo desarrollado por tener la variable dependiente tres categorías se generan dos funciones discriminantes y dos puntajes por cada unidad de análisis, los cuales se muestra en dos columnas con los encabezados siguientes "discriminant score from function 1" y "discriminant score from function 2". Junto a ellas se muestra también la columna "predicted group" (que indica el grupo pronosticado al que pertenece la unidad de análisis) asimismo otras columnas son "probabilities of membership in group 1" "probabilities of membership in group 2" y "probabilities of membership in group 3" que indican la probabilidad de pertenencia de la unidad de análisis a cada categoría de la variable dependiente.

Importancia relativa de las variables

En los grupos de mujeres conformados (mujeres sin hijos, con uno o dos hijos y con tres o mas hijos) la importancia de las variables no es la misma. Así tenemos:

Classification Function Coefficients

	REV201 Nº DE HIJOS		
	1.00 NO TIENE	2.00 1 A 2	3.00 3 A MAS HIJOS
V012 Current age - respondent	.488	.630	.807
V0251	5.162	5.400	5.838
V1061	5.088	4.745	3.177
V5011	-.527	5.983	7.091
RECV717 CONDTRABAJO	6.510	6.044	5.972
(Constant)	-13.956	-19.405	-25.454

Fisher's linear discriminant functions

En el grupo de mujeres con tres o más hijos la variable edad -v012- es más importante en la identificación de este grupo.

El lugar de residencia -v0251-(urbano / rural) explica también las diferencias en el número de hijos de las mujeres. El coeficiente estimado es más alto para el grupo de mujeres con tres y más hijos.

El nivel educativo -v1061- es para el grupo de mujeres sin hijos más determinante. Comprobándose empíricamente la relación "a mayor nivel educativo menor número de hijos".

El estado conyugal -v5011-(nunca unidad / unida) es más importantes en las mujeres con tres o mas hijos. Se comprueba que las mujeres unidas tienden a tener más hijos que las no unidas.

La condición de actividad es más determinante en la disminución del número de hijos. Así, las mujeres que tienen menos hijos son aquellas que tienen ocupación laboral.

La constante (constant) comprende todo aquello que no es explicado por las variables consideradas en el modelo. Para el ejemplo, esta constante crece en relación directa al número de hijos, por ello se recomienda considerar más variables a fin de reducir esta constante.

Clasificación de las unidades de acuerdo a las funciones discriminantes

Los puntajes discriminantes llevan asociadas una probabilidad, la cual se convierte en una regla de clasificación de las unidades de análisis. Esta regla se basa en el teorema de Bayes. La probabilidad que una unidad de análisis con un puntaje discriminante, pertenezca a uno de los tres grupos (mujeres sin hijos, con uno o dos hijos y con tres o mas hijos) se estima mediante la siguiente expresión:

$$P(Gi/D) = \frac{P(D/Gi)\,P(Gi)}{\sum_{i=1} S(D/Gi)P(Gi)}$$

Esta expresión significa lo siguiente: ¿Cuál es la probabilidad que una unidad de análisis con un puntaje discriminante pertenezca a alguno de los tres grupos? Esta probabilidad viene dada por el cociente de dos expresiones. La primera expresión es el numerador, donde se multiplica la probabilidad condicional que una unidad de análisis pertenezca a alguno de los tres grupos, por la probabilidad a priori, que en este caso viene a ser igual para todas ("all groups equal"). El denominador, es la sumatoria de las combinaciones de las probabilidades condicionales para cada uno de los grupos.

Determinadas las probabilidades posteriores, la unidad de análisis pertenece al grupo cuya probabilidad calculada ha sido la mayor. Estos resultados se muestran en una tabla desagregada, siempre que se active la opción "Display-Casewise results-limit cases to first ". Los resultados globales se presentan en una tabla resumen, en una matriz denominada "matriz de confusión". En la diagonal principal, de esta tabla, se presentan el número de casos correctamente clasificados, es decir aquellos que coinciden con la clasificación a priori. Por encima y por debajo de la diagonal, se muestran los casos que a priori

se clasificaban en una categoría y luego por el análisis discriminante han cambiado de categoría. En el ejemplo, si en el análisis realizado se encuentra el 30% perteneciendo a una de las tres categorías, implica que las variables seleccionadas no han tenido efecto en la conformación de los grupos.

En el ejemplo desarrollado se muestra la tabla siguiente que resume los resultados del análisis realizado:

Classification[b,c]

			Predicted Group Membership			
		REV201 Nº DE HIJOS	1.00 NO TIENE	2.00 1 A 2	3.00 3 A MAS HIJOS	Total
Original	Count	1.00 NO TIENE	7997	648	267	8912
		2.00 1 A 2	1697	5031	1722	8450
		3.00 3 A MAS HIJOS	433	2195	7853	10481
	%	1.00 NO TIENE	89.7	7.3	3.0	100.0
		2.00 1 A 2	20.1	59.5	20.4	100.0
		3.00 3 A MAS HIJOS	4.1	20.9	74.9	100.0
Cross-validated [a]	Count	1.00 NO TIENE	7997	648	267	8912
		2.00 1 A 2	1697	5031	1722	8450
		3.00 3 A MAS HIJOS	433	2195	7853	10481
	%	1.00 NO TIENE	89.7	7.3	3.0	100.0
		2.00 1 A 2	20.1	59.5	20.4	100.0
		3.00 3 A MAS HIJOS	4.1	20.9	74.9	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 75.0% of original grouped cases correctly classified.

c. 75.0% of cross-validated grouped cases correctly classified.

En la sección "original" los valores de la diagonal de la tabla indican los casos clasificados correctamente en los grupos. En el primer grupo (mujeres sin hijos) existen 89.7% de casos correctamente clasificados, en el grupo de mujeres con uno a dos hijos existen 59.5% de casos correctamente clasificados, mientras que en el grupo de mujeres con tres o mas hijos existe un 74.9% de casos en que la clasificación original coincide con la clasificación hallada por el método indirecto.

En promedio el 75% de los casos de la muestra, para los tres grupos de mujeres, la clasificación original ha coincidido con la clasificación hallada por métodos indirectos. Esto indica que el análisis realizado ha sido efectivo.

Otros criterios estadísticos para determinar la calidad del análisis

Existen otros criterios estadísticos para evaluar la calidad de la Función Discriminante, entre ellos tenemos:

1. Los histogramas de los puntajes discriminantes para cada uno de los grupos, en lo cuales debe verificarse la distribución normal de dichos puntajes y detectar la existencia de valores extremos.
2. Los eigen-value o valores propios, que explican un porcentaje de la varianza total. En la medida que el mayor porcentaje de varianza, sea explicado por ellos, la efectividad del método será mejor.
3. La alta correlación entre los puntajes discriminantes y cada variable independiente.

7. Principales conclusiones del análisis discriminante realizado

Las hipótesis planteadas se han contrastado con los datos de la encuesta. Así, se ha podido determinar que la variable más explicativa del número de hijos de las mujeres en edad fértil, es el nivel educativo, comprobándose la validez de la primera hipótesis.

Otras variables que contribuyen a establecer diferencias entre las mujeres sin hijos y las mujeres con hijos son la edad, el lugar de residencia (urbano, rural), la condición de ocupado (trabaja, no trabaja) y el estado conyugal (Nunca unida, unida). Los datos confirman la segunda hipótesis e incorporan una variable adicional (estado conyugal) como explicativa de estas diferencias.

En cuanto al Modelo

El análisis realizado ha permitido comprobar empíricamente que existen diferencias entre las mujeres en edad fértil sin hijos, con uno o dos y con tres o más hijos, las cuales se deben principalmente a las variables edad, nivel educativo, lugar de residencia y estado conyugal. Esto se comprueba al interpretar los indicadores como el **Test M de Box** y el **Lambda de Wilks**. Estas pruebas permiten comprobar la pertinencia del modelo analizado e identificar las variables más discriminatorias.

Además la tabla **"Classification Results"** resume los resultados de la clasificación realizada. Así, a mayor porcentaje de coincidencias entre la clasificación a priori determinada y la obtenida por el modelo discriminante será más efectiva la contribución del modelo a la predicción del comportamiento de la variable dependiente.

En cuanto a la Consistencia de los Datos: En la tabla "Statistical Groups" se muestran el promedio y la desviación estándar a partir de las cuales se puede calcular el coeficiente de variabilidad para cada grupo. Este valor debe ser muy semejante en los diferentes grupos. Además, las varianzas de los grupos conformados deben ser muy parecidas mientras que entre los promedios se espera encontrar ciertas diferencias. Estos valores calculados permitirán evaluar la consistencia de los datos, comparándose los resultados esperados con los obtenidos. Así, podrán encontrarse las variables que presenta la mayor distorsión respecto al conjunto general de información y revisar la base de datos correspondiente. Por ejemplo, en el caso analizado la variable edad es más homogénea en los grupos de mujeres con tres o más hijos que entre aquellas que no los tienen. Este comportamiento observado está de acuerdo al comportamiento esperado en la población, por lo cual podemos concluir que los datos para esta variable son consistentes.

La **matriz de covarianzas** también contribuye al análisis de la información por cuanto el signo de los coeficientes calculados se espera tenga correspondencia con el comportamiento teórico de la variable. En el ejemplo, se conoce que el número de hijos de las mujeres tiene una relación directa con el nivel educativo. Esta relación se verifica empíricamente observando los resultados de la matriz de covarianzas.

La tabla en la que se muestra la importancia relativa de la variable en cada grupo conformado permite evaluar la consistencia de los datos. Así, en el ejemplo la variable nivel educativo explica mejor la ausencia de hijos en las mujeres y la condición de ocupado explica más la tenencia de tres o más hijos. Estos resultados al ajustarse a los esperados permiten inferir la consistencia de la información.

VI. CONCLUSIONES

1. Una aplicación eficiente de los modelos multivariados da lugar a la elaboración de diagramas causales en los cuales se representen los diferentes tipos de relación entre las variables. Estos diagramas deben elaborarse antes de iniciar la aplicación del análisis mulvitaviado de modo que se facilite el análisis posterior de los resultados.
2. El análisis factorial confirmatorio (AFC) permite reducir la cantidad de variables a investigar y además agrupar en factores excluyentes las mismas
3. La aplicación del AFC es más efectiva cuando todas las variables que intervienen son cuantitativas.
4. Cuando intervienen variables cualitativas y cuantitativas en el AFC, los coeficientes de la matriz de correlación en muchos casos no son interpretables, debido a que el algoritmo del programa SPSS, que sirve para calcular este coeficiente se aplica solo a variables cuantitativas.
5. En el modelo factorial las variables que pertenecen a un factor pueden ser reemplazadas entre sí, toda vez que cada factor representa una dimensión del fenómeno social en estudio, de este modo se puede reducir la cantidad de variables sin afectar el objeto en estudio
6. Los conglomerados de unidades de análisis formados mediante el análisis de conglomerados (AC) tienen la característica de ser homogéneos y diferir significativamente entre sí.
7. Para alcanzar mayor efectividad con el AC se requiere que las variables se representen al menos en una escala ordinal.
8. Si el objetivo del investigador es tener las variables más representativas de cada dimensión del fenómeno en estudio, se recomienda aplicar el AFC, antes del AC, de este modo se garantiza que las variables seleccionadas representen una dimensión diferente del fenómeno en estudio.
9. El análisis discriminante clasificatorio (ADC) permite la disposición de las unidades de análisis en grupos, de acuerdo a ciertos criterios a priori fijados y en función de un conjunto de variables. A diferencia del AC donde se desconoce la cantidad de grupos a conformar, en el ADC este número es conocido a priori y lo que se procura es encontrar las variables que contribuyen más a la conformación de estos grupos
10. Se recomienda utilizar el análisis discriminante clasificatorio (ADC) cuando las variables estén expresadas en al menos una escala ordinal.
11. Para evaluar la consistencia de los datos formule un modelo causal hipotético y aplique el AFC. A fin de comprobar la naturaleza de la relación entre las variables y los supuestos a priori de tal manera que los datos sean confirmatorios de tales supuestos, en caso contrario deben ser evaluados nuevamente.

12. La evaluación de la consistencia de los datos aplicando el AC, se efectúa a partir del análisis las medias y las varianzas calculadas para cada cluster o conglomerado. Las varianzas altas en los cluster indican la probable presencia de valores extremos en la base de datos, por lo cual se recomienda verificar estos datos, potencialmente influyentes.
12. Una de las aplicaciones del análisis discriminante para evaluar la consistencia de los datos consiste en calcular los puntajes discriminantes con los cuales se puede clasificar las unidades de análisis en grupos excluyentes. Al observar el porcentaje de coincidencias entre la clasificación a priori y la resultante del análisis discriminante, se espera u porcentaje de coincidencias superiores al 60%, caso contrario se recomienda revisar la base de datos.
13. Las aplicaciones del análisis multivariado son diversas, las mostradas en esta guía constituyen sólo una pequeña fracción de sus usos posibles. Así, el modelo factorial es útil además para evaluar las políticas aplicadas en determinado sector de la economía a través del análisis de la magnitud de los coeficientes factoriales estimados en el modelo. El análisis cluster además puede aplicarse para la selección de variables de las encuestas de hogares, a partir de los conglomerados conformados. Otras aplicaciones del análisis discriminante permitirán predecir el comportamiento de cierto grupo de individuos a partir del conocimiento de las variables explicativas del modelo discriminante y determinar el perfil de un individuo a partir del puntaje discriminante calculado, lo que facilitará la aplicación de políticas focalizadas.

VII. RECOMENDACIONES

1. Las técnicas de análisis multivariado explicadas deben ser vistas como un complemento al análisis descriptivo de los datos, sin las cuales no se podría alcanzar un conocimiento completo del problema ni aplicar las pruebas estadísticas más apropiadas para confirmar las hipótesis planteadas.

2. Para la formulación de los modelos es importante tener un conocimiento teórico de las relaciones entre las variables de tal manera que los procesos iterativos de estimación conduzcan a resultados consistentes.

3. Los modelos multivariados de dependencia (discriminante y regresión por ejemplo) no están exentos de algunas deficiencias. Así la relación de causalidad entre las variables fijada como supuesto inicial en estos modelos puede ocultar el verdadero sentido de la relación quedando sus efectos confundidos con las variables explícitamente consideradas. Por ejemplo, se espera que el nivel educativo tenga una fuerte relación con los ingresos y el lugar de residencia. Aunque pueden existir poblaciones donde esta relación se muestre en otro sentido. Por ello se recomienda realizar pruebas repetidas para contrastar los resultados obtenidos.

4. Cuando se analiza de las encuestas de corte transversal generalmente el investigador es un observador pasivo con poco control sobre los resultados. Por ello las pruebas multivariadas que posteriormente se efectúen deben contar con un marco conceptual de tal manera que las relaciones entre las variables (magnitud y el signo de los coeficientes del modelo) tengan un sólido fundamento teórico. Así, cualquier resultado alejado del esperado será atribuible a los datos.

5. El INEI también cuenta con información de encuestas panel a las cuales se recomienda aplicar las técnicas multivariadas desarrolladas, a fin de obtener los coeficientes que permitan analizar los ciclos y tendencias del fenómeno social en estudio.

Referencias Bibliográficas

1. **Bienvenido Visauta Vinacua,** "Modelos Causales" - Editorial Hispano Europea, España 1986.

2. **Bienvenido Visauta Vinacua,** "Análisis Estadístico con SPSS para Windows" - Mc Graw Hill, Volumen II Estadística Multivariante 1998.

3. **Andrew L. Comrey,** "Manual del Análisis Factorial " - Cátedra, España 1985.

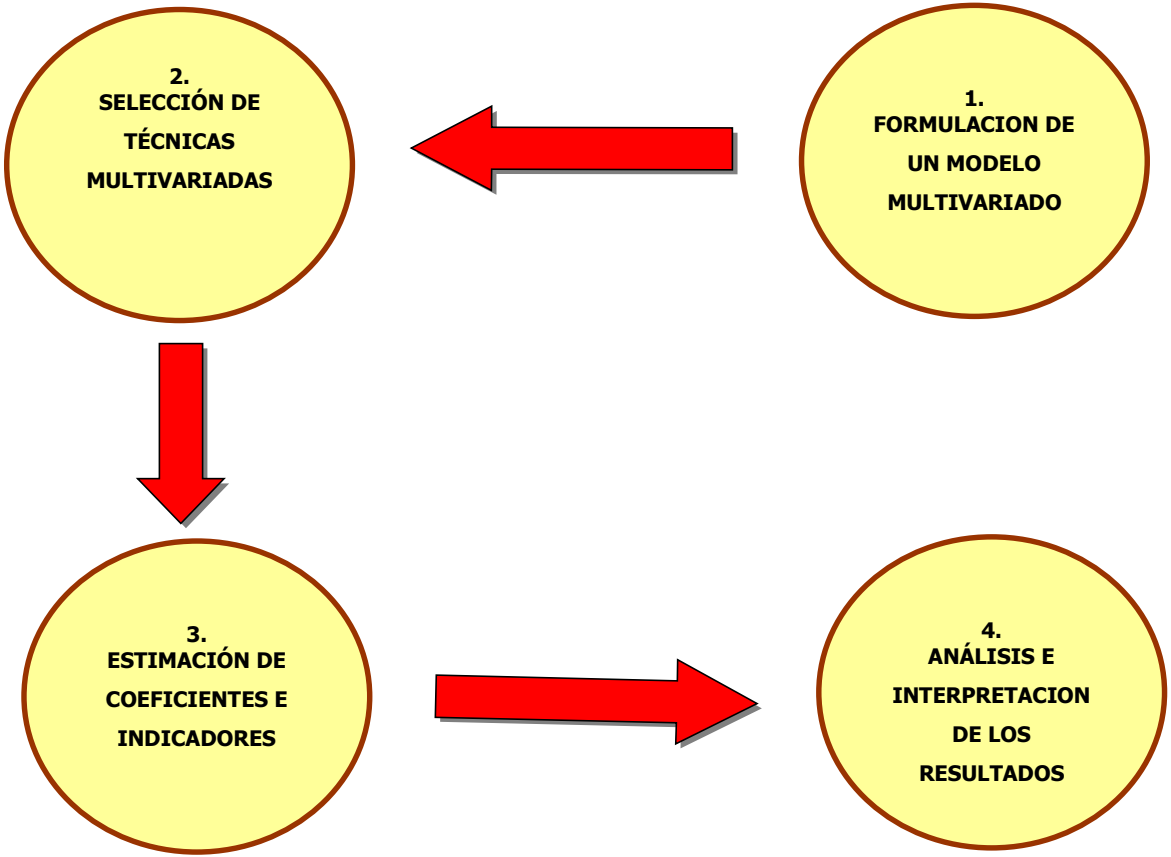
4. **Manuel Mora y Araujo, Paul Lazarsfeld, Warren Torgenson, y otros,** "Medición y Construcción de Indices" - Editorial Nueva Visión, Argentina 1971.

5. **Programa MECOVI PERU,** "Compendio de Cuestionarios Trimestrales en la Encuesta Nacional de Hogares" - Centro de Investigación e Informática -OTDETI -INEI, Lima, Mayo 2000.

6. **Centro de Investigación y Desarrollo - INEI, Setiembre 2001,** "Variables Investigadas con la ENDES".

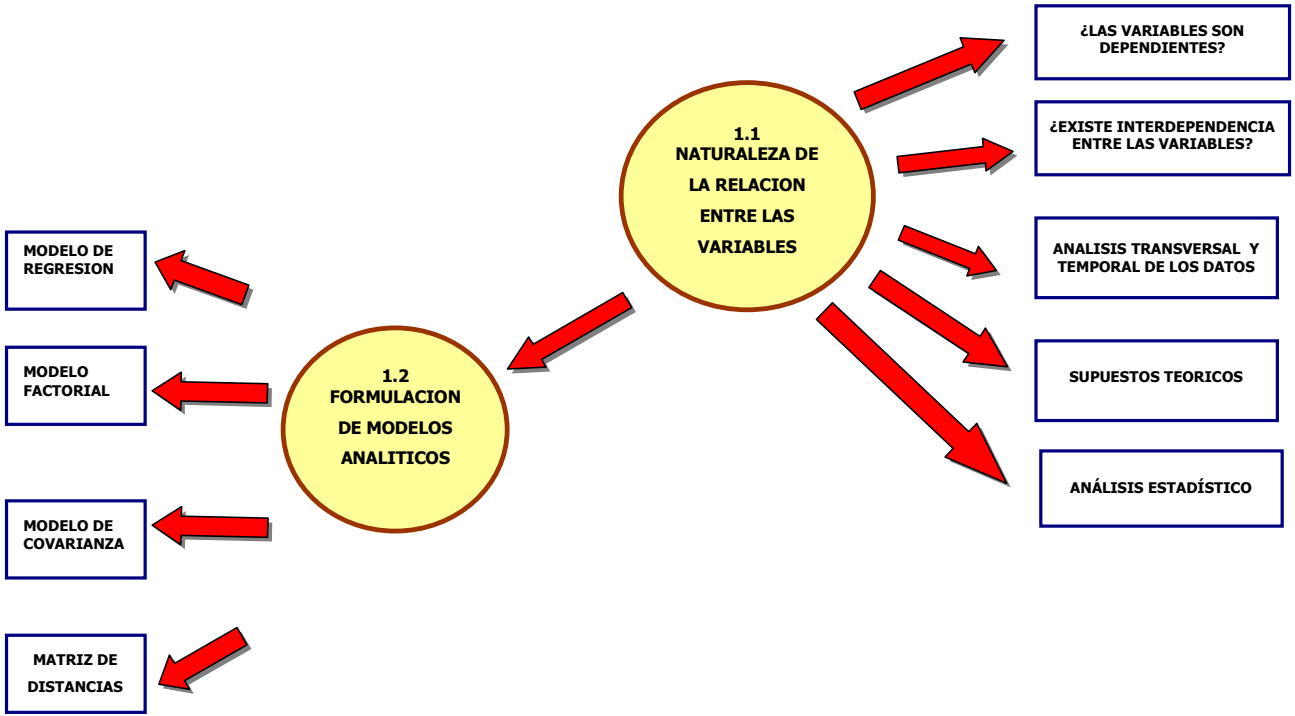
ANEXO Nº 1

ETAPAS PARA REALIZAR EL ANÁLISIS MULTIVARIADO



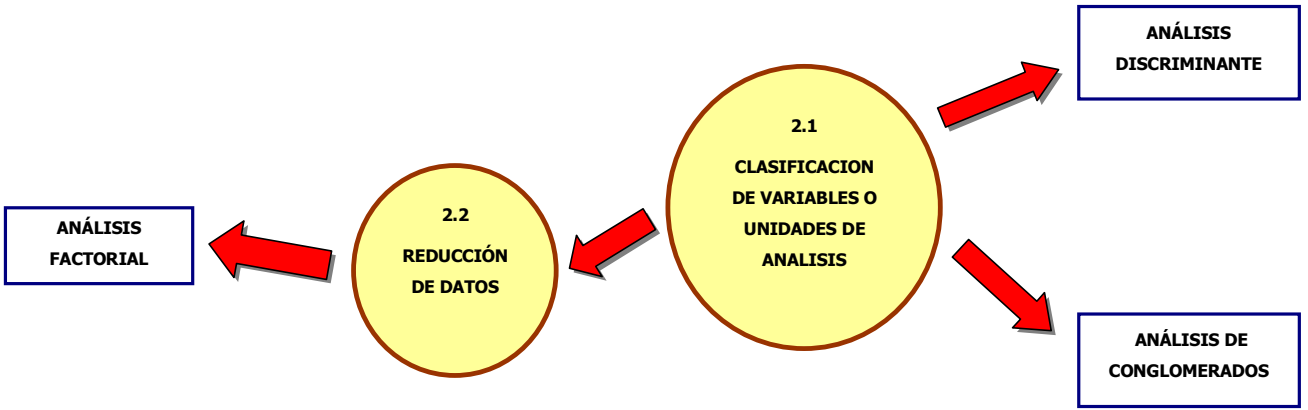
ANEXO Nº 2

1. FORMULACION DE UN MODELO MULTIVARIADO



ANEXO Nº 3

2. PRINCIPALES TÉCNICAS MULTIVARIADAS



ANEXO Nº 4

3. CONSISTENCIA DE LOS RESULTADOS