Cuadernos del CES 12

Jorge Padua

Paquete estadístico para las ciencias sociales (SPSS): oferta y condiciones para su utilización e interpretación de resultados

Centro de Estudios Sociológicos
EL COLEGIO DE MEXICO

Jorge Padua

PAQUETE ESTADISTICO PARA LAS CIENCIAS SOCIALES (SPSS): OFERTA Y CONDICIONES PARA SU UTILIZACION E INTERPRETACION DE RESULTADOS

Centro de Estudios Sociológicos
El Colegio de México

Primera edición, 1975

Guanajuato 125, México 7, D. F.

Impreso y hecho en México
Printed and made in Mexico

INDICE

I. INTRODUCCION

El enorme desarrollo experimentado en los aparatos electrónicos en el almacenaje y procesamiento de datos, unido al acceso relativamente fácil de los investigadores a estos computadores, abre para el cientista social nuevas posibilidades tanto para el tratamiento de datos en gran escala, cuanto para la aplicación de técnicas estadísticas complejas, en lo matemático o en lo referente a la cantidad de tiempo requerido para su cómputo manual.

El uso de computadores para el análisis de datos requiere de una serie de pasos:

a) los datos deben ser preparados de manera tal que puedan ser "alimentados" a la máquina, es decir, los datos deben estar en la forma de tarjetas perforadas, cinta magnética, disco, o alguna otra forma especial (cintas de papel, p. ej.);

b) el investigador debe decidir qué es lo que quiere hacer con sus datos, es decir, qué tipo de cálculo va a solicitar; correlaciones, coeficientes de confiabilidad, análisis de varianza, etc.;

c) es necesaria la preparación de un "programa", que no es otra cosa que una guía de instrucciones que describe para la máquina la forma en cómo los cálculos deben ser realizados;

d) es necesario confeccionar además una serie de instrucciones para poner un conjunto particular de datos en un determinado

---

NOTA: La presente publicación forma parte de un estudio más amplio sobre técnicas de investigación.

centro de computación, para un determinado programa de cálculo. Esto es lo que se llama en la jerga técnica "job". Típicamente cada job consiste de varios mazos de tarjetas: a) Tarjetas de sistema (system cards); b) programa; c) control de tarjetas paramétricas; y d) datos. Las tarjetas de sistema varían de un centro de computación a otro, y por lo general éstos instruyen a sus usuarios sobre la preparación de tarjetas, o se encargan ellos mismos de confeccionarlas. Estas tarjetas incluyen nombre del usuario, de la investigación, nombre de variables, etc. En la mayoría de los casos incluye además una tarjeta con el nombre del programa. El programa es la lista de instrucciones que especifican el tipo y el orden de las operaciones que la computadora va a efectuar. Cuando el programa está ubicado en la memoria del computador, por medio de una de las tarjetas del sistema se lo ubica. Las tarjetas paramétricas (o tarjetas de control) son una lista de instrucciones específicas para un juego de datos en particular y para el problema específico. Por lo general estas tarjetas incluyen tarjetas de problemas, tarjetas opcionales y tarjetas de formatos. Las tarjetas de problemas contienen una descripción del job específico (número de observaciones, número de variables, forma del input y del out-put, etc.) Las tarjetas de formatos instruyen a la máquina sobre dónde y cómo encontrar los datos en cada tarjeta, cinta o disco. Las tarjetas opcionales programan a la máquina para hacer exactamente lo que debe hacer y con qué exactitud.

El objetivo específico de este trabajo es el de delinear algunos principios estadísticos y metodológicos referidos a los programas, particularmente a aquellos incluidos en el SPSS (Statistical Package for the Social Sciences). Paquete estadístico para las ciencias sociales, uno de los programas más completos y de mayor difusión en el medio académico y de investigación en el hemisferio occidental.

En la medida en que fueron los estadísticos y los psicometristas los que -de los cientistas sociales- han utilizado más los computadores, no es de extrañar que los programas más sofisticados se concentren en cálculos estadísticos; sin embargo, hay una generalización cada vez más creciente del uso de computadores en una variedad de disciplinas (medicina, música, administración, simulación de relaciones internacionales, estrategia, etc.) que son una indicación de la flexibilidad de estas máquinas para realizar tareas complejas, al mismo tiempo que nos alerta sobre la necesidad de tener acceso a programadores y analistas de sistemas con cierto grado de familiaridad con los problemas sustantivos de disciplinas en particular y que sean capaces de preparar a los computadores para cálculos y operaciones no complementados en programas de tipo "paquete".

Finalmente quiero alertar al lector sobre las descripciones que aparecen en las secciones siguientes: me he preocupado de hacer accesible cada una de las posibilidades que aparecen en el SPSS; las más de las veces he seguido de cerca el estilo y el ejemplo de los

autores del Manual, otras veces incluyo algo más de información. Todo el razonamiento es de carácter verbal, más que matemático, y la idea es presentar al investigador no familiarizado con la estadística matemática las opciones que hay en el programa, cuándo utilizarlas, cómo interpretar los out-puts, etc. Para mayores detalles, en cada una de las diferentes técnicas, existe abundante bibliografía que puede ayudar al lector a un uso más preciso del rico material de cálculo disponible en el SPSS.

La utilidad que puede tener este tipo de enfoque, está en relación directa con el desconocimiento que el lector tenga de las limitaciones de cada uno de los estadísticos. Con demasiada frecuencia se solicita a los centros de computación cálculos que utilizan valioso tiempo de programación y de computadora para resultados finales sin ninguna significación. Por ejemplo, ocurre que el usuario solicita cálculos de correlaciones Pearsons para variables y atributos tales como sexo, opiniones, pertenencia a clase, etc. Ahora bien, una de las limitaciones de la correlación Pearsons es que las variables tienen que estar medidas a nivel intervalar al menos, y algunas de las variables que el usuario empleaba estaban medidas a nivel nominal u ordinal. ¿Quiere decir que no es posible utilizar correlaciones? De ninguna manera; simplemente lo que quiere decir es que había que computar el coeficiente de correlación o de asociación apropiado. Más adelante se analizarán algunas alternativas.

Otro problema bastante común y asociado a los niveles de medición es el de seleccionar las estadísticas apropiadas para un conjunto determinado de datos, a partir del out-put de la computadora. Muchos programas (por ejemplo el subprograma CROSSTAB) producen una serie de coeficientes de asociación, de los cuales el investigador debe seleccionar los que corresponden a sus datos; muchas veces ocurre que se publican todos o se utilizan algunos indebidamente.

En este trabajo, vamos a ocuparnos de los pasos b) y en parte del paso c) en el uso de computadoras. Es decir vamos a proponer algunos estadísticos -la mayoría de los cuales están contenidos en un programa denominado SPSS- de manera tal que el investigador decida más apropiadamente qué quiere hacer con sus datos, qué tipo de cálculo va a solicitar, etc.

II. NIVELES DE MEDICION

Uno de los requisitos teóricos más importantes para la utilización eficiente de modelos matemáticos o estadísticos es que éstos sean isomórficos con el concepto o el conjunto de conceptos que los modelos representan. En otras palabras, el modelo matemático debe tener la misma forma que el concepto. De no ser así, cualquier tipo de operación es ilegítima.

Las reglas para la asignación de números a objetos, conceptos o hechos están determinadas por una serie de teorías distintas, donde cada una de ellas se denomina nivel de medición. La teoría de la medición especifica las condiciones en que una serie determinada de datos se adaptan legítimamente a un nivel u otro, de manera que exista isomorfismo entre las propiedades de las series numéricas y las propiedades del objeto. De esta manera es posible utilizar el sistema matemático formal como un modelo para la representación del mundo empírico o conceptual.

Toda medición tiene tres postulados básicos, que son necesarios para igualar, ordenar y añadir objetos. Estos principios o postulados son:

a) $a=b$ o $a\neq b$, pero no ambos al mismo tiempo

b) si $a=b$ y $b=c$, entonces $a=c$

c) si $a>b$ y $b>c$ entonces $a>c$

El primer postulado es necesario para la clasificación. Nos va a permitir determinar si un objeto es idéntico o no a otro en virtud del atributo que consideramos. Manteniendo constante la dimen-

sión tiempo, establece relaciones excluyentes.

El segundo postulado nos capacita para establecer la igualdad de un conjunto de elementos con respecto a una característica determinada. Es el principio de la transistividad de igualdades.

El tercer principio, o principio de la transistividad de desigualdades o inecuaciones, nos permite establecer proposiciones ordinales o de rango.

Sobre la base de estos postulados y de acuerdo al tipo de operaciones empíricas que se puedan realizar con los atributos del universo que se desea escalar, se tienen cuatro distintos tipos de niveles de medición:

a) nominal

b) ordinal

c) intervalar

d) por cocientes o racionales.

Cada uno de estos niveles se caracteriza por el grado en que permanecen invariantes. La naturaleza de esta invariancia, fija los límites a los modos de manipulación estadística que pueden aplicarse legítimamente a los datos incluidos en el nivel de medición.

A) Nivel nominal

Es la forma más elemental de medición, en la que simplemente se sustituyen a los objetos reales por símbolos, números, nombres. Esta clasificación de los elementos de un universo de acuerdo a determinados atributos, da a la medición a este nivel un significado más cualitativo que cuantitativo.

Para "medir" en este nivel, se asignan símbolos o signos al atributo del objeto o conjunto de objetos que se desea medir, con la condición básica de no asignar el mismo signo a categorías que son diferentes; o diferentes signos a la misma categoría.

Por medio de esta escala simplemente diferenciamos a los objetos de acuerdo a la categoría a la que pertenecen. Ejemplos de medición de variable a nivel nominal:

Sexo (masculino-femenino)

Religión (católico, protestante, judío, mahometano, otra)

B) Nivel ordinal

En este nivel de medición, los objetos no solamente aparecen como diferentes, sino además existe una cierta relación entre grupos de objetos. Es decir la relación "mayor que" es válida para todos los pares de objeto de diferente clase.

Se obtiene una escala ordinal "natural" cuando los datos originales admiten una relación "más grande que" para todos los pares de unidades.

Los numerales asignados a los objetos rangueados son llamados valores de rango. Ejemplo: autoridad militar (capitán, teniente, sargento, etc.), distribución de poder o de prestigio Status, socio-económico (alto, medio, bajo).

C) Nivel intervalar

En las dos escalas examinadas más arriba, los elementos del sistema eran clases de objetos y las relaciones se reducían a igualdad a más grande que. Ninguno de los dos niveles especificaba distancia entre clases, es decir que cuando hablábamos de que A era mayor que B, y que B era mayor que C, no podíamos hacer ninguna afirmación sobre si la distancia que separaba A de B, era mayor, igual, más o menos importante y cuán intensa, que la que separaba B de C. En lo que se refiere a distancia que separa a objetos o clases de objeto tanto el nivel nominal como el ordinal son nominales

En una escala intervalar, podemos afirmar no solamente que tres objetos o clases a, b, c, están en una relación $a>b>c$, sino también que en los intervalos que separan a los objetos se da la relación $\overline{ab}>\overline{ij}$ o $\overline{ij}>\overline{ab}$.

Es decir que es una escala o nivel que está caracterizado por un orden simple de los estímulos sobre la escala, y por un orden en los tamaños que miden las distancias en los estímulos adyacentes sobre la escala. Los datos contienen especificaciones relativas al tamaño exacto de los intervalos que separan a todos los objetos en la escala, a más de las propiedádes que se obtienen en la escala nominal y ordinal. Aquí estamos realmente a nivel de lo que entendemos por "cuantificación" propiamente tal y se requiere el establecimiento de algún tipo de unidad física de medición que sirva como norma, y que por lo tanto pueda aplicarse indefinidamente con los mismos resultados.

La escala de intervalos supone la adjudicación de un cero arbitrario, y las operaciones aritméticas se aplican sobre las diferencias entre los valores de la escala. Ejemplos: temperatura, tests de IQ, etc.

D) Nivel por cociente o racional

Supone un 0 absoluto y es posible cuando existen operaciones para determinar cuatro tipos de relaciones:

-similitud

-ordenación de rangos

-igualdad de intervalos

-igualdad de proporciones (razones o cocientes)

Una vez determinada la igualdad por cociente, los valores numéricos pueden transformarse con sólo multiplicar cada valor por una constante. Con este tipo de escala es posible realizar todo tipo de operaciones aritméticas. Ejemplo: distancia, peso, volumen, etc.

III. PROGRAMA ESTADISTICO DEL SPSS

El SPSS contiene programas estadísticos para:

- -estadística descriptiva y distribuciones de frecuencia para una variable
- -tablas de contingencia y tabulaciones cruzadas
- -correlaciones bivariatas
- -correlaciones parciales
- -regresiones múltiples
- -análisis de la varianza
- -análisis discriminatorio
- -análisis factorial
- -análisis de correlaciones canónicas
- -análisis de escalograma, para escalas Guttman

Y una serie de subrutinas, para modelos lineares en análisis de regresiones, como regresiones con variables mudas (Dummy variables) y Path análisis. Nosotros vamos a tratar de especificar para qué sirven cada uno de estos sub-programas dando algunos detalles sobre las condiciones para su utilización.

IV. ESTADISTICA DESCRIPTIVA

A) Medidas de tendencia central

Incluye solamente la media aritmética, la mediana y el modo. Estos promedios indican los valores centrales de observaciones. Sirven para: describir en forma sintética al conjunto de datos; los promedios provenientes de muestras pueden ser utilizados como una buena estimación de los valores parámetros, existiendo para ello una serie de técnicas de estimación a partir de valores muestrales que serán examinadas en la parte correspondiente a estadística inferencial.

La media aritmética, es lo que conocemos familiarmente como promedio, esto es, el resultado de dividir la suma total de todas las mediciones por la cantidad total de casos.

La mediana es el punto en la distribución que la divide en dos partes iguales, esto es, por encima de la mediana se encuentra el 50% de los casos y por debajo el otro 50%.

El modo es el punto en la distribución que registra la frecuencia máxima; la media aritmética es la más exacta y confiable de las tres medidas.

Empleo de media, mediana y modo

El nivel de medición apropiado para cada uno de los niveles es:

nominal ..... modo
ordinal ..... modo, mediana
intervalar o racional ..... modo, mediana, media

Existen algunos casos en los cuales, además del nivel de medi

ción apropiado, es necesario tener en cuenta la forma de la distribución de los datos.

En síntesis se computa:

Media Aritmética Cuando	a) los datos están medidos a nivel intervalar al menos b) cuando la distribución es simétrica, aproximadamente normal y unimodal c) cuando se van a efectuar cálculos posteriores
Mediana	a) cuando los datos están medidos a nivel ordinal al menos b) cuando se cuenta con distribuciones incompletas c) cuando la distribución es necesariamente asimétrica
Modo	a) cuando la escala es nominal b) cuando se desea conocer el caso más típico

B) Medidas de variabilidad o de dispersión

Las medidas de tendencia central por sí solas constituyen una información valiosa, pero insuficiente para un análisis de la distribución, necesitando el complemento de lo que se conoce como medidas de variabilidad o de dispersión. Estas medidas indican cómo se distribuyen los valores alrededor de las medidas de tendencia central.

Las medidas de variabilidad más importantes son:

Amplitud total (range) que denota simplemente la diferencia entre los valores máximo y mínimo de la distribución;

La amplitud semi-intercuartil (Q) que es la mitad de la amplitud de 50% central de casos;

La desviación media (AD) es la media aritmética de todos los desvíos con relación al promedio, cuando no se toman en consideración los signos algebraicos.

La desviación estándar (sigma), que es un desvío cuadrático medio, o en términos operacionales la raíz cuadrada de la media aritmética del cuadrado de las desviaciones de cada una de las medidas en relación al promedio aritmético.

Cada una de estas medidas de variabilidad complementan la información de las medidas de tendencia central. Por ejemplo:

Medida de tendencia central	Medida de variabilidad
Modo	Amplitud total
Mediana	Amplitud semiintercuartil
Media	Desviación media-desviación estándar

Los valores provenientes de la amplitud total, son útiles para tener una idea general del rango de variación en los datos; sin embargo, es poco confiable en la medida que para su cálculo solamente se utilizan dos valores extremos, por lo cual es imprecisa.

La amplitud semiintercuartil, complemento de la mediana es útil como índice de la simetría de la distribución total. En distribuciones perfectamente simétricas el cuartil 1 (Q_1) y el cuartil 3 (Q_2), están a igual distancia del centro de la distribución o mediana (Q_3). Si las distancias son desiguales hay asimetría. En resumen:

Asimetría positiva cuando: $(Q_3-Q_2) > (Q_2-Q_1)$

Asimetría negativa cuando: $(Q_3-Q_2) < (Q_2-Q_1)$

Asimetría cero cuando: $(Q_3-Q_2) = (Q_2-Q_1)$

La desviación media nos informa sobre la magnitud de las desviaciones respecto a la media. Cuando la distribución es normal, aproximadamente el 58% de las observaciones caen en el espacio comprendido entre la media aritmética más una desviación media y la media aritmética y menos una desviación media.

La desviación estándar es la medida de variabilidad más exacta y confiable y la más empleada en cálculos posteriores (correlación, varianza, etc.) La interpretación más común de la desviación estándar es idéntica a la que realizamos con la desviación media, esto es, en términos de distribución normal; sumando y restando a la media aritmética una desviación estándar debe esperarse el 68.26% de todos los casos, caiga en esa área de la curva. De esta manera podemos estimar para una distribución cualquiera cuanto se acerca o se aleja de una distribución normal. Cuando la distribución es muy asimétrica el cálculo de la desviación estándar no es conveniente, recomendándose más el cálculo de desviación media o desviación intercuartil.

El SPSS contiene además una serie de medidas para la determinación de la forma de la distribución como la curtosis (kurtosis) y la oblicuidad (skewness).

La oblicuidad es un estadístico que indica el grado en que la distribución se aproxima a la distribución normal. Cuando la distribución es completamente simétrica la oblicuidad es igual a 0. Los valores positivos indican que los casos se concentran más a la izquierda de la media, mientras que los valores extremos a su derecha. Los valores negativos se interpretan exactamente al revés. Se aplica únicamente cuando los datos están a nivel intervalar al menos.

La curtosis es una medida relativa a la forma de la distribución (mesocúrtica o platicúrtica). La curtosis en una distribución normal es cero. La curtosis es positiva cuando la distribución es estrecha y en forma de pico, mientras que los valores negativos indican una curva aplanada.

El cálculo de estadística descriptiva en el SPSS, está contenido en dos subprogramas: condescriptive y frecuencies.

El subprograma condescriptive en variables continuas a nivel intervalar.

El subprograma frecuencies, trabaja con variables discretas, por lo consiguiente se corresponde a los niveles nominales y ordinales.

El investigador debe tener cuidado en la selección de cuáles de los valores que aparecen en las descripciones sumarias en las hojas del out-put de la computadora, a fin de seleccionar solamente aquellos estadísticos que se corresponden con la naturaleza de sus datos.

V. CONFIABILIDAD DE LOS ESTADISTICOS

Las características de las poblaciones (o universos) son denominados valores parámetros. ¿Qué es una población, desde el punto de vista estadístico?, es una materia de definición arbitraria, aunque en un sentido general puede definirse como el conjunto total de unidades (individuos, objetos, reacciones, etc.) que pueden ser delimitados claramente por la posesión de un atributo o cualidad propio y único. Muestra es cualquier subconjunto de ese conjunto mayor que constituye la población. Téngase bien claro entonces que dos muestras pueden considerarse como proveniendo de la misma población o de dos poblaciones diferentes, en relación a un solo aspecto (por ejemplo, coeficiente intelectual, actitud frente al cambio, etc.)

Gran parte de las investigaciones en ciencas sociales se basan en estudio de muestras, y a partir de ellas se desea estimar o inducir las características de la población o universo.

El error estándar nos da una estimación de la discrepancia de las estadísticas muestrales con relación a los valores de la población. Si pretendemos utilizar el estadístico muestral como una estimación de los valores parámetros, cualquier desviación de la media muestral sobre la media parámetro puede considerarse como un error.

El error estándar de la media aritmética nos dice cuál es la magnitud del error de estimación para la media aritmética. El error estándar de la media es el desvío estándar de la distribución de las medias muestrales provenientes de una misma población. El error estándar de la media es directamente proporcional al tamaño de su desviación estándar e indirectamente proporcional al tamaño de la muestra. La interpretación del error estándar de una media se realiza en términos de distribuciones normales; esto es, la finalidad es de-

terminar en qué medida las medias aritméticas de muestras similares a la que estamos considerando, se alejan o aproximan a la media de la población. Cuanto menor sea el valor del error estándar, tanto mayor será la fiabilidad de nuestra media como estimación de la media de la población. Por lo general la probabilidad de error se fija con valores de .05 y .01, es decir que la probabilidad de error aceptada es de 5 sobre 100 o de 1 sobre 100, de allí que la estimación de los parámetros se haga en intervalos de confianza, donde a la media muestral le agregó y le disminuyó tantos valores de error estándar como para cubrir 95% del total de casos, o el 99%.

La fórmula genérica para la determinación de intervalos de confianza es:

Para .05 : $M \pm 1{,}96\ \sigma_M$

Para .01 : $M \pm 2{,}58\ \sigma_M$

Para la distribución de probabilidades distintas de .05 y .01, simplemente se buscan en la tabla de curva de distribución normal los valores correspondientes.

Las interpretaciones señaladas tanto para la media aritmética como para el resto de los estadísticos, son posibles únicamente cuando las muestras son aleatorias.

El cálculo del error estándar de la media está contenido en el subprograma condescriptive.

Confiabilidad de la mediana

Se interpreta de la misma forma que el error estándar de la media, y en poblaciones normalmente distribuidas, el error estándar de la mediana da una variabilidad aproximadamente de un 25% mayor que el de las medias.

Error estándar de desviación estándar, Q, proporciones, porcentajes, frecuencias.

Las fórmulas para su cálculo difieren, pero la interpretación es idéntica a la realizada más arriba, esto es, en términos de estimación de valores parámetros por medio de intervalos de confianza.

Confiabilidad de un coeficiente de correlación.

Lo mismo que los demás estadísticos, un coeficiente de correlación

está sujeto a los errores de muestreo. La variabilidad entonces estará en función del tamaño del error estándar del coeficiente. Sin embargo, la distribución de los coeficientes obtenidos no será uniforme ya que dependerá de la magnitud de r, así como el número de observaciones que componen la muestra. En la medida en que los coeficientes varían entre + 1.00 y − 1.00 cuando la r parámetro se aproxima a esos valores extremos, la distribución será más asimétrica; negativamente asimétrica para los valores positivos de r y positivamente asimétrica para los valores negativos de r. Solamente en el caso en que el r parámetro sea 0, entonces la distribución de las r muestrales será normal.

Para muestras grandes el problema de la asimetría carece de importancia significativa; cuanto más grande la muestra, menor será la dispersión de las r, de allí que aun cuando r es cero, en muestras menores de 25 casos sea necesario tener alguna precaución en las estimaciones.

Cuando r es grande y la muestra es grande, el error estándar del coeficiente será mínimo.

VI. CONFIABILIDAD DE DIFERENCIA ENTRE ESTADISTICOS (SUBPROGRAMA BREAKDOWN)

El subprograma breakdown calcula e imprime las sumas, medias, desviaciones estándar y varianza de la variable pendiente, en los distintos subgrupos que la componen, según clasificaciones complejas que incluyan de 1 a 5 variables independientes, cualesquiera sea el nivel de medición (en las variables independientes, la variable dependiente debe ser medida a nivel intervalar).

Antes de desarrollar las alternativas del subprograma de breakdown conviene introducir conceptualmente la idea de confiabilidad de la diferencia entre estadísticos.

Para la investigación es importante no solamente estimar los valores poblacionales, sino utilizar el error estándar para interpretar varios resultados, en lo relativo a las diferencias que pueden existir entre ellos.

El tipo de preguntas que nos planteamos aquí es ¿cuál es la fiabilidad de la diferencia entre medias proporciones, etc., que hemos registrado en nuestras observaciones?, ¿son los hombres o las mujeres más capaces en comprensión verbal?, ¿el rendimiento intelectual en las clases medias es superior o inferior o igual al rendimiento intelectual en las clases bajas?, etc.

a) Error estándar de la diferencia de medias (Subprograma T-Test)

La magnitud de la oscilación en la diferencia entre medias obtenidas de muestras distintas, dependerá naturalmente de la magnitud de la oscilación que es propia de las medias. La estabilidad de las medias estará representada por sus respectivos errores estándares.

Cuando las N son lo suficientemente grandes, las medias oscilan alrededor de un valor central (parámetro que por lo general no conocemos.) Nuestra finalidad es entonces determinar primero si existe diferencia, para luego definir su magnitud.

El problema reside entonces en determinar si la diferencia que se examina entre las dos medias muestrales, implica además una diferencia en la distribución de la población; en otras palabras si la diferencia es la expresión de diferencias reales a niveles poblacionales, o se deben simplemente a los efectos del azar (y por lo consiguiente del error) en las muestras.

El test T de student nos ayuda a establecer cuando la diferencia entre dos medias es significativa. Para ello se formula una hipótesis nula. Una hipótesis nula supone que las dos muestras provienen de la misma población; consecuentemente las desviaciones son interpretadas como debidas al efecto del azar sobre las muestras. Según la hipótesis nula se supone que la distribución de las diferencias es normal, de donde $M_1 - M_2 = 0$

El nivel de significación para la aceptación o rechazo de la hipótesis nula es seleccionado por el investigador. Los más comunes son de .05 y .01 aunque esto depende más bien del área que se está investigando (para aceptar o rechazar una vacuna nueva que cure el cáncer puedo elegir un nivel menor; cuando se trata de la introducción de una medicina para suplantar alguna en uso con cierto grado de efectividad, eligiré un nivel de significación mayor).

El valor de t nos informará entonces sobre la probabilidad o improbabilidad para la aceptación o rechazo de la hipótesis nula, o de alguna hipótesis alternativa. Es decir no se afirma que no existe una diferencia en los resultados, sino únicamente que la diferencia no es, o es significativa.

El subprograma T-TEST, computa los valores t y sus niveles de probabilidad para dos tipos de casos:

a) Muestras independientes o error estándar de la diferencia para medias no correlacionadas, es decir, para situaciones en las que las dos series de observaciones son independientes. Por ejemplo, comparación del rendimiento de hombres y mujeres en una situación de test.

b) Muestras apareadas, o error estándar de la diferencia para medias correlacionadas. El ejemplo típico es el de las mediciones antes-después en diseños experimentales.

Existen casos en los cuales el investigador no plantea la hipótesis nula (la hipótesis de las no-diferencias), sino plantea una

hipótesis alternativa, en la que trata de demostrar que la media en un grupo es más grande que la media del otro. En estos casos la interpretación del valor t obtenido se hace a partir de lo que se llama test de una sola cola, es decir, se toma en cuenta solamente una mitad de la distribución.

VII. TABLAS DE CONTINGENCIA Y MEDIDAS DE ASOCIACION (SUBPROGRAMA CROSSTABS)

Este subprograma contiene tanto tabulaciones cruzadas para tablas de n x k, así como una serie de medidas de correlación, asociación y de confiabilidad de la diferencia entre estadísticos.

Comenzaremos por los análisis de tipo más sencillo para continuar luego a los cálculos de medidas más complejas.

Tabulaciones cruzadas

Una tabulación cruzada es simplemente la combinación de dos o más variables discretas o clasificatorias en la forma de tabla de distribuciones de frecuencia. El cuadro resultante puede ser sometido a análisis estadístico, en términos de distribuciones porcentuales, aplicación de test de significación, coeficientes de asociación y de correlación, etc.

Las tablas cruzadas son muy utilizadas en análisis de encuestas, en tablas de 2 x 2 o con la introducción de variables de prueba o de control o intervinientes, constituyendo así tablas de n x k. Aquí el investigador debe tener especial cuidado cuando solicita tablas de n x k que el tamaño de su muestra sea lo suficiente grande para permitir que cada uno de los casilleros contenga las frecuencias esperadas (ver J. Padua: Muestras para hipótesis sustantivas y para hipótesis de generalización).

De todos modos existen una serie de restricciones al uso de los distintos estadísticos que señalaremos más adelante que imponen algunas limitaciones en cuanto a la cantidad total de casilleros, ya sea por cantidad de variables o por cortes en cada una de ellas. Por

ejemplo: el investigador debe recordar que si combina digamos 4 variables, todas dicotomizadas, la cantidad total de casilleros será de 16; si las variables estarían tricotomizadas la cantidad de casilleros ascendería a 81. La fórmula genérica para el cálculo del tamaño final de la matriz es:

$$M = r_1 . r_2 . r_3 \ldots . r_n$$

donde

r: cantidad de cortes o divisiones en cada una de las variables

M: tamaño de la matriz de datos

Hay que recordar, que en este tipo de tablas hay que esperar un promedio de 10 a 20 casos en cada uno de los casilleros, lo que hace que las muestras deben tener tamaños considerables cuando se desea cuadros muy complejos.

Normalmente las tablas imprimen tanto las frecuencias dentro de cada casillero o celda, como los porcentajes con respecto al marginal horizontal y al marginal vertical y al total general (en ese orden), además de todos los coeficientes que incluye la subrutina y que a continuación pasamos a detallar.

Ji cuadrado (x^2)

Es un modelo matemático o test para el cálculo de la confiabilidad o significado de diferencias entre frecuencias esperadas (f_e) y frecuencias observadas (f_o). La utilidad de este test no-paramétrico para variables nominales, reside en su aplicación para prueba de hipotesis para tres tipos de situaciones:

a) prueba de hipótesis referidas al grado de discrepancia entre frecuencias observadas y frecuencias esperadas, cuando se trabaja sobre la base de principios apriorísticos;

b) pruebas de hipótesis referidas a la ausencia de relación entre dos variables. Se trata de pruebas de independencia estadística y son trabajadas en base a tablas de contingencia; y

c) pruebas referidas a la bondad de ajuste. En este caso se trata de comprobar si es razonable aceptar que la distribución empírica dada (datos observados), se ajusta a una distribución teórica, por ejemplo, binomial, normal, Poisson, etc. (datos esperados).

Supuestos y requisitos generales

-las observaciones deben ser independientes entre sí.

-los sucesos deben ser mutuamente excluyentes.

-las probabilidades que figuran en las tablas de X^2 están basadas en una distribución continua, mientras que el X^2 calculado en la práctica lo está en base a variables discretas. Se supone que esta última puede aproximarse a la primera.

-el nivel de medición mínimo es nominal.

-las frecuencias esperadas mínimas por casillero deben ser 5, cuando esto no se cumple es necesario aplicar un factor de corrección (correccion de Yates).

-la prueba de X^2 es útil solamente para decidir cuando las variables son independientes o relacionadas. No nos informa acerca de la intensidad de la relación, debido a que el tamaño de la muestra y el tamaño de la tabla ejercen una influencia muy fuerte sobre los valores del test. Existen numerosos estadísticos basados en la distribución de X^2 que son útiles para la determinación de la intensidad de la relación (ver coeficiente Fi, Cramer, C, etc.)

Coeficiente Fi (φ)

Es una medida de asociación (fuerza de la relación) para tablas de 2 x 2. Toma el valor cero cuando no existe relación, y el valor + 1.00 cuando las variables están perfectamente relacionadas.

Coeficiente V. de Cramer

Es una versión ajustada del coeficiente φ para tablas de r x k. El nivel de medición es nominal y el coeficiente varía entre 0 y 1.00

Coeficiente de contingencia (C)

Basado como los dos anteriores en X^2 se pueden utilizar matrices de cualquier tamaño. Tiene un valor mínimo de 0, y sus valores máximos varían según el tamaño de la matriz (por ejemplo para matrices de 2 x 2 el valor máximo de C es .707; en tablas de 3 x 3 es .816, etc., la fórmula genérica $\sqrt{\frac{k-1}{k}}$) consecuentemente para una interpretación del coeficiente obtenido en cualquier tabla de 2 x 2 habría que dividir ese valor por .707.

Limitaciones:
-el límite superior del coeficiente está en función del número de categorías
-dos o + coeficientes C no son comparables, a no ser que provengan de matrices de igual tamaño.

El coeficiente Q de Yule

También como los anteriores para escalas nominales, se utiliza únicamente en tablas de 2 x 2. Los valores Q son 0 cuando hay independencia entre las variables, siendo sus límites $\pm$ 1.00 cuando cualquiera de las 4 celdas en la tabla contiene 0 frecuencias: por lo general cuál de los distintos coeficientes es preferible en este caso (φ o Q) depende del tipo de investigación y del tipo de distribución marginal.

Coeficiente lambda (λ)

Es un coeficiente de asociación para tablas de r x k, cuando las dos variables están medidas a nivel nominal.

El coeficiente lambda pertenece a la familia de un grupo de coeficientes (τ_b, λ y otros), que se utilizan para hacer interpretaciones probabilísticas en tablas de contingencia. El tamaño del coeficiente indica la reducción proporcional en errores de estimación en la variable dependiente cuando los valores en la variable independiente son conocidos.

El valor máximo de λ es 1.00 y ocurre cuando las predicciones pueden ser hechas sin ningún error. Un valor cero significa que no hay posibilidad de mejorar la predicción. Un coeficiente lambda .50 significa que podemos reducir el número de errores a la mitad, etc.

Coeficiente τ_b de Goodman y Kruskal

Sirve a los mismos propósitos que el coeficiente lambda y debe ser preferido cuando los marginales totales no son de la misma magnitud.

Coeficiente de incertidumbre

También para niveles nominales en tablas de contingencia de r x k. La computación del coeficiente toma en cuenta simetría y asimetría (el coeficiente lambda toma en cuenta, por ejemplo, solamente la asimetría.) El coeficiente asimétrico es la proporción de reducción de la incertidumbre conocido por efecto del conocimiento de la variable independiente. La ventaja de este coeficiente sobre lambda es que considera el total de la distribución y no solamente el modo.

El máximo valor del coeficiente de incertidumbre es 1.00 que denota la eliminación de la incertidumbre, y se alcanza cada vez que cada categoría de la variable independiente está asociada a solamente una de las categorías de la variable dependiente. Cuando no es posible lograr ningún avance en términos de disminución de la incer-

tidumbre el valor del coeficiente es 0. Una versión simétrica del coeficiente mide la reducción proporcional en incertidumbre que se gana conociendo la distribución conjunta de casos.

Coeficiente TAU b

Mide asociación entre dos variables ordinales en tablas de contingencia. Este coeficiente es apropiado para tablas cuadradas (es decir, donde el número de columnas es idéntico al número de filas. Sus valores varían de 0 a $\pm$ 1.00. El valor cero indica que no existe asociación entre pares concordantes y discordantes. El valor $\pm$ 1.00 se obtiene cuando todos los casos se ubican a lo largo de la diagonal mayor. En tablas de 2 x 2 el valor de Tau b es idéntico al de φ con la ventaja de que el coeficiente tau b proporciona información sobre la dirección de la relación a través del signo. Los valores negativos indican que los casos se distribuyen sobre la diagonal menor. Los valores intermedios entre 0 y 1 indican casos que se desvían de las diagonales. A mayor desviación mayor proximidad al valor cero (es decir cuando los pares discordantes son iguales a los pares concordantes).

Coeficiente TAU c

Sirve a los mismos propósitos que el coeficiente tau b, pero este coeficiente es más apropiado para tablas rectangulares (cuando el número de columnas difiere del número de líneas). La interpretación de ambos coeficientes es similar.

Coeficiente gamma (γ)

Mide asociación entre dos variables ordinales en tablas de contingencia de r x k. Mientras que el coeficiente tau c depende para su cómputo solamente del número de líneas y de columnas, y no las distribuciones marginales, tomando en cuenta los empates, el coeficiente gamma excluye los empates del denominador de la fórmula de cómputo, siendo además un coeficiente con posibilidades de aplicación en datos no agrupados. Además el coeficiente no requiere de cambios en la forma de la matriz. Los valores numéricos de gamma por lo general son más altos que los valores de tau b y de tau c.

El coeficiente gamma es simplemente el resultado del número de pares concordantes menos el número de pares discordantes, divididos por el número total de pares unidos. Los valores gamma varían entre 0 y $\pm$ 1.00, donde el signo indica la dirección de la relación y los valores la intensidad de la misma.

El SPSS provee valores gamma para tablas de tres a n entradas, en el que se calcula el gamma de orden cero y además gammas parciales. El gamma de orden cero mide la relación entre dos variables,

siendo exactamente el mismo al que se discute en los párrafos anteriores. Cuando la matriz tiene tres o más dimensiones, el SPSS (subprograma CROSSTABS) computa un coeficiente gamma de orden cero (reduciendo la tabla a variable dependiente e independiente) y además medidas de correlación parcial gamma de la relación entre las dos variables, controladas por una o más variables adicionales. El investigador puede analizar así cómo influyen en la relación de sus variables dependiente e independiente, la introducción de variables adicionales (en la sección correspondiente a correlaciones parciales indicaremos con mayor detalle el uso y significado de las correlaciones parciales).

Coeficiente D de Sommer

Para variables ordinales en tablas de contingencia, este coeficiente toma en cuenta los empates, pero el ajuste es realizado de manera distinta a la utilizada en los coeficientes tau b y tau c.

Coeficiente eta (η)

Se utiliza cuando la variable independiente es nominal y la variable dependiente intervalar. Este coeficiente indica cuán disimilar son las medias aritméticas en la variable dependiente dentro de las categorías establecidas por la variable independiente. Cuando las medias son idénticas el valor del coeficiente es 0. Si las medias son muy diferentes y sus varianzas son pequeñas, los valores de eta se aproximan a 1.00.

Correlación biserial (r_b)

Para utilizar cuando una de las variables está medida a nivel nominal y la otra a nivel intervalar, la variable a nivel nominal puede ser una dicotomía forzada. Sus valores oscilan entre 0 y $\pm$ 1.00.

Correlación punto-biserial (r_{pb})

Similar al coeficiente biserial, se aplica cuando la variable nominal es una dicotomía real. La interpretación de ambos coeficientes es idéntica y su utilización más común se encuentra en la construcción de pruebas, sobre todo para la determinación de validez.

Coeficiente de correlación Spearman (ρ)

Es un coeficiente de correlación por rangos, cuando las dos variables están medidas a nivel ordinal, e indica el grado en que la variación o cambio en los rangos de una de las variables están relacionados a las variaciones o cambios en los rangos en la otra variable. Tanto el coeficiente ρ de (rho) Spearman como el coeficiente τ (tau) de Kendall, son coeficientes no paramétricos, es decir que no se hacen supuestos acerca de la distribución de los casos sobre las variables. Ambos coeficientes suponen la no existencia de muchos empates, por lo cual el sistema de organización de los datos y de cómputo, son distintos a los de las tabulaciones cruzadas, y por ello se encuentran en subprogramas diferentes (en este caso el subprograma correspondiente en el SPSS es denominado Nonpar Corr).

Para el cómputo del coeficiente correlación ρ de Spearman (así como para el τ de Kendall), no se toman en consideración los valores absolutos en las variables, sino su orden de rango. El coeficiente rho de Spearman se aproxima más que el coeficiente tau de Kendall al coeficiente de correlación producto-momento de Pearson, cuando los datos son aproximadamente continuos. Los valores del coeficiente varían entre -1.00 y +1.00

Coeficiente de correlación Tau de Kendall (τ)

Similar al coeficiente Ro, se utiliza cuando las dos variantes son ordinales. Por lo general debe preferirse cuando existe abundante número de empates entre rangos, caso que se da especialmente cuando el número total de casos es grande y se los clasifica en un número relativamente pequeño de categorías. El subrprograma nonpar corr contiene factores de corrección para empates tanto para el coeficiente tau como para el coeficiente rho. Los valores de este coeficiente oscilan entre -1.00 y +1.00

Coeficiente de correlación producto-momento de Pearson (r)

Para dos variables medidas a nivel intervalar por lo menos, éste es un coeficiente de correlación paramétrico que nos indica con la mayor precisión cuando dos cosas están correlacionadas, es decir, hasta qué punto una variación en una se corresponde a una variación en otra. Sus valores varían de +1.00 que quiere decir correlación positiva perfecta; a través de 0 que quiere decir independencia completa o ausencia de correlación, hasta -1.00 que significa correlación perfecta negativa. El signo indica por lo tanto la dirección de la convariación y la cifra la intensidad de la misma. Una correlación perfecta de +1.00 indica que cuando una variable se "mueve" en una dirección, la otra se mueve en la misma dirección y con la misma intensidad. La

interpretación de la magnitud de r depende en buena medida del uso que se quiera dar del coeficiente, el grado de avance teórico en el área, etc. Guilford* sugiere como orientación general, la siguiente interpretación descriptiva de los coeficientes de correlación producto-momento;

r menor que .20 - correlación leve, casi insignificante

r de .20 a .40 - baja correlación, definida, pero baja

r de .40 a .70 - correlación moderada, sustancial

r de .70 a .90 - correlación marcada, alta

r de .90 a 1.00 - correlación altísima, muy significativa

De todos modos la interpretación del coeficiente está además condicionada a su grado de significación (ver significación de los estadísticos).

Premisas o suposiciones fundamentales para el cómputo de r

- ambas variables deben ser medidas a nivel intervalar al menos
- la dirección de la relación debe ser rectilínea
- la distribución tiene que ser homoscedástica (las dispersiones en las columnas y en las líneas del diagrama de dispersión deben ser similares). Esta condición prevalece cuando las dos distribuciones son simétricas entre ellas.

El programa imprime el valor del coeficiente de correlación, la cantidad de casos, y la significación estadística.

Existen varios coeficientes que se derivan del coeficiente de correlación producto-momento, entre otras, por ejemplo:

r^2: mide la proporción de la varianza en una variable que es "explicada" por la otra.

Diagrama de dispersión (Scattergram)

El SPSS puede imprimir además, a través de su subprograma Scattergram,

---

* Guilford, J. P.: Psychometric Methods; McGraw-Hill, N. Y., 1954

el diagrama de dispersión para dos variables, computando además la regresión linear simple. El diagrama de dispersión es un gráfico de puntos donde, basado en los valores en las dos variables, una de las variables define el eje horizontal y la otra el eje vertical. Estos diagramas son de mucha utilidad ya que nos dan una imagen de la relación, que puede ser utilizada para la determinación de la homoscedasticidad, por ejemplo y para decidir si vale o no la pena continuar más adelante.

Para la confección de los diagramas, el usuario tiene que tomar algunas decisiones sobre cómo se van a manejar la falta de datos (missing data), qué clase de escala tiene que ser utilizada y cómo se van a colocar las líneas segmentadas.

Comúnmente dos líneas verticales y dos líneas horizontales segmentadas dividen cada eje con tres secciones, de manera tal que el gráfico consiste en 9 rectángulos iguales. Si el investigador prefiere, las líneas segmentadas pueden ser diagonales que atraviesen el gráfico.

Los datos (es decir, cada punto sobre el diagrama) están representados por asteriscos (*) cuando un caso cae en alguna intersección, de dos a ocho casos el número es impreso. Nueve o más casos están representados por el número 9. Cuando la escala contiene muy pocas categorías, existe la posibilidad de que los puntos sobre el diagrama se den muy amontonados, lo que limita la utilización del diagrama de dispersión recomendándose para esas situaciones una tabulación cruzada.

Los estadísticos que acompañan al diagrama de dispersión, son aquellos asociados a las regresiones lineares simples: correlación producto-momento, error estándar de la estimación, r^2, significación de la correlación, intersección con el eje vertical, e inclinación.

Es necesario discutir con algún detalle el concepto de regresión, ya que sirve de base para la utilización de predicciones, así como de ayuda para la comprensión del concepto de correlaciones parciales y múltiples.

El concepto de regresión trata de describir no solamente el grado de relación entre dos variables, sino la naturaleza misma de la relación, de manera tal que podamos predecir una variable conociendo la otra (por ejemplo, el rendimiento académico a partir del resultado en un test, el ingreso a partir de la educación, etc.). Fíjense que aquí no estamos interesados en explicar por qué las variables se relacionan como se relacionan, sino simplemente a partir de la relación dada, predecir una variable a partir del conocimiento de los valores en la otra. Si la variable X es independiente de la variable Y (es decir si son estadísticamente independientes), no estamos en condiciones de

predecir Y a partir de X o viceversa, es decir nuestro conocimiento de X no mejora nuestra predicción de Y. Por razonamiento inverso, cuando las variables son dependientes -están correlacionadas, co-varían-, el conocimiento de X nos puede ayudar a predecir el comportamiento de Y y viceversa.

Esto se logra mediante lo que se llama ecuación de regresión de Y sobre X, que nos da la forma en cómo las medias aritméticas de los valores de Y se distribuyen según valores dados de X.

La operación de regresión contiene los siguientes supuestos:

- que la forma de la ecuación es linear.
- que la distribución de los valores de Y sobre cada valor de X es normal, y
- que las varianzas de las distribuciones de Y, son similares para cada valor de X.
- que el error es igual a 0

Cumplidas estas condiciones, la ecuación de la regresión es:

$Y = \alpha + \beta X$

donde α y β son constantes y se les da una interpretación geométrica.

Si X es igual a 0, entonces $Y=\alpha$

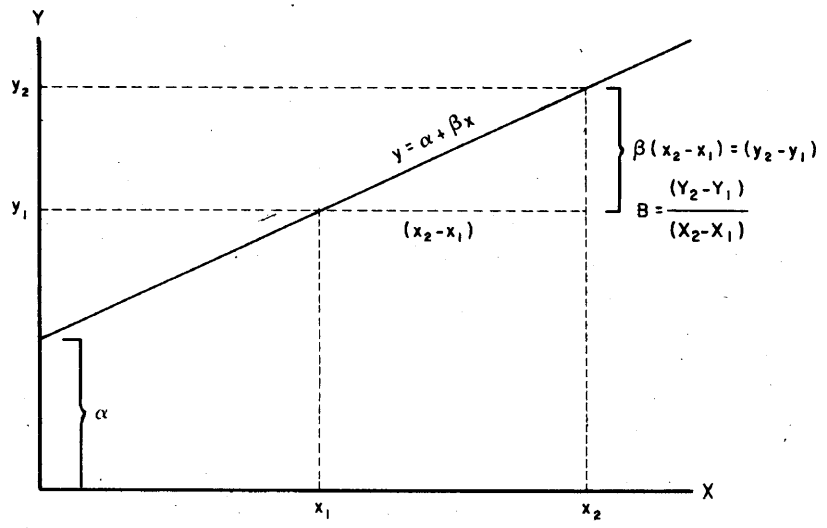
α representa entonces el punto donde la línea de regresión cruza el eje de Y.

La inclinación de la línea de regresión es dada por β, indicando la magnitud en el cambio de Y por cada unidad de cambio en X. Cuando β es igual a 1, y si las unidades de X y Y están indicadas por distancias idénticas a lo largo de sus ejes respectivos, la línea de regresión estará en un ángulo de 45° con respecto al eje de las X. A más grande el tamaño de β, mayor será el declive, es decir más grande el cambio en Y dados determinados valores de cambio en X.

H. Blalock* presenta la siguiente figura que aclara la interpretación geométrica del coeficiente de regresión.

---

* Hubert Blalock: Social Statistics (2a. ed.) McGraw-Hill, Kogahusha, Tokio, Japón, 1972.



Es decir que β mide la tangente del ángulo, con lo cual queda identificado el ángulo.

Correlación parcial

Todos los coeficientes de correlación y asociación examinados hasta ahora tomaban en cuenta la relación entre dos variables (con la excepción del coeficiente gamma).

La correlación parcial provee medidas del grado de relación entre una variable dependiente Y, y cualquiera de un conjunto de variables independientes, controladas por una o más de esas variables independientes. Es decir, describe la relación entre dos variables, controlando los efectos de una o más variables adicionales.

Es similar a lo que se hace en tabulaciones cruzadas, cuando se introducen variables de control. Sin embargo, ya habíamos visto que para controlar varias variables con varios valores, necesitábamos una muestra demasiado grande, además la inspección del efecto era de tipo literal.

Con correlaciones parciales, el control no solamente es estadístico, sino además la cantidad de casos no necesita ser muy grande.

$r_{ij.k}$ indica entonces a i y j variable independiente y dependiente (el orden es inmaterial, ya que la correlación entre ij y ji serán idénticas). La variable de control es indicada con k (*).

(*) Salvo en el caso de utilizar la correlación parcial para predic-

Desde la perspectiva de la teoría de la regresión, la correlación parcial entre i y j, controlando por k es la correlación entre los residuales de la regresión de i sobre k y de j sobre k, permitiéndonos establecer predicciones sobre las variables dependientes e independientes a partir del conocimiento del efecto que tiene la variable control sobre ellas.

El coeficiente de correlación parcial puede ser utilizado por el investigador para la comprensión y clarificación de las relaciones entre tres o más variables. Por ejemplo, puede ser utilizado para la determinación de espureidad, para la localización de variables intervinientes, y para la determinación de relaciones causales.

El coeficiente de correlación parcial para la determinación de espureidad en las relaciones: una relación espuria es aquella en la cual la correlación entre una variable X y una variable Y, es el resultado de los efectos de otra variable (Z) que es el verdadero predictor de Y. La correlación es espuria cuando, controlando por Z (esto es, a Z constante), los valores de X no varían con los valores de Y. Este es el caso en que los coeficientes de correlación parcial dan valores 0 o próximos a 0.

Supóngase una relación entre X y Y de .40. Computo un coeficiente de correlación parcial $r_{xy.z}$ y el resultado es .20. Este coeficiente de correlación parcial ya me está indicando que la variable Z explica parcialmente la relación original entre X y Y. Computo un coeficiente de correlación parcial de segundo orden, en el cual controlo por dos variables, Z y W. El coeficiente de correlación parcial es ahora $r_{xy.zw} = .06$, es decir que la relación desaparece, consecuentemente, la relación original era espuria.

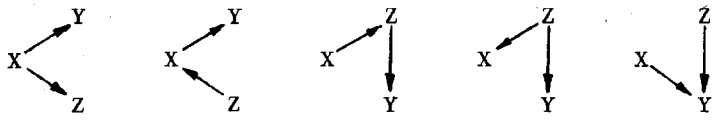
Para la localización de variables intervinientes, así como para la determinación de relaciones causales, el problema es de naturaleza más conceptual, esto es, hay que combinar valores de coeficientes de correlación parcial, con una serie de supuestos sobre las formas de las distribuciones y sobre la intervención de otras variables, además de las que se consideran en el modelo. Los supuestos no pueden ser verificados empíricamente por el análisis estadístico, sino que van a depender del razonamiento teórico.

En cualquiera de los siguientes casos, salvo en (e) la corre-

---

ciones en la utilización de regresiones en cuyo caso se acostumbra a interpretar $r_{12.3}$ denotando 1 la variable dependiente, 2 la variable independiente y 3 la variable de control.

lación parcial $r_{yz.x}$ debe ser próxima a cero (Y es la variable dependiente, es decir la que va a ocurrir al final en la secuencia temporal);



(d) es un caso típico de correlacion espuria. La relación entre X y Y se explica en función de las relaciones de X con Z y de Y con Z.

(c) este es un modelo donde en la que Z actúa como variable interviniente en la relación entre X y Y. La correlación parcial también dará 0. Pero hay que tener mucho cuidado en no interpretar los modelos (c) y (d) de la misma manera, y la correlación parcial tiene sentido solamente para probar que no hay relación entre X y Y, sino cuando interviene Z. El modelo (b) es similar, aunque ahora X es interviniente.

En el modelo (a) la relación X con Y, y la de X con Z son relaciones directas, mientras que no se postula relación entre Y y Z.

En los modelos (a) y (b), la correlación parcial entre X y Z, controlado por Y debe ser 0.

Similarmente, en los modelos (c) y (d), la correlación parcial entre X y Y, controlado por Z, debe ser 0.

Cuando el modelo es (e), la correlación parcial $r_{xy.2}$ dará valores más altos que la correlación entre X y Y. La correlación entre X y Z será 0.

El investigador debe informar a los programadores sobre la lista deseada de correlaciones parciales, en la que se especifiquen las combinaciones de variables (todas las combinaciones posibles o solamente algunas de las combinaciones). Por ejemplo: si se presentan variables: ingreso, educación, actitud frente al cambio y religiosidad, o se especifican las combinaciones deseadas o se deja que se correlacionen todas con todas en n combinaciones.

La palabra with especifica en el programa la combinación entre variables cuando la lista incluye solamente algunas combinaciones. Cuando el programa no incluye with se calculan todas las combinaciones posibles.

VIII. ANALISIS DE REGRESIONES MULTIPLES (SUBPROGRAM REGRESSION)

Este subprograma es considerablemente más complejo que los anteriores, y puede ser utilizado para una variedad bastante grande de análisis de variables múltiples: regresiones polinomiales, regresiones mudas (dummy), análisis de la varianza y análisis de la covarianza, predicciones, etc.

Por lo general, la regresión múltiple requiere variables medidas a nivel intervalar o racional y que las relaciones sean lineares y aditivas. Sin embargo, hay casos especiales en los cuales regresores mudos, medidos a nivel nominal pueden ser incorporados a la regresión, relaciones no lineares y no aditivas pueden ser manipuladas, etc.

Existen algunas diferencias entre análisis de correlaciones múltiples y análisis de regresiones múltiples, que conviene destacar.

Los análisis de correlaciones múltiples se utilizan para: a) la evaluación de la medida en que cada variable predictora o subconjunto de variables contribuye a la explicación de los puntajes de un criterio sobre una muestra; o b) para predecir los puntajes de un criterio en una muestra diferente en la cual existe información del mismo grupo de variables predictoras. Aquí no estamos interesados tanto en la relación entre la variable dependiente y cada una de las variables independientes tomadas separadamente, sino en el poder explicativo del conjunto de variables independientes en su totalidad. El coeficiente de correlación múltiple es expresado entonces como: $r_{1.2345}$ si son 4 las variables predictoras, o $r_{1.23}$ si son dos, etc.

Los modelos para el análisis de regresiones múltiples, a la vez que son más complejos en términos de cantidad de operaciones o

de derivaciones que a través de ellos se puedan realizar, son bastante más simples en términos de los supuestos y condiciones para su utilización.

Por ejemplo, los modelos correlacionales requieren que las variables y los parámetros observables tengan una distribución normal conjunta; los modelos de regresión múltiple requieren solamente que la distribución de las desviaciones de la función de regresión sea normal, no se supone que las variables predictoras provengan de una distribución normal multivariata, o a veces requiere que los datos estén contenidos en códigos binarios.

Nosotros vamos a dar algunos ejemplos de prueba de hipótesis a través de análisis de regresiones múltiples.

El análisis de las regresiones múltiples puede ser utilizado ya sea para la descripción de las relaciones entre variables o como instrumento para la inferencia estadística.

Como instrumento descriptivo la regresión múltiple es útil:

a) para encontrar la mejor ecuación linear de predicción y para evaluar su eficiencia predictiva;

b) para evaluar la contribución de una variable o un conjunto de variables;

c) para encontrar relaciones estructurales y proveer explicaciones para relaciones complejas de variables múltiples.

Habíamos visto que el coeficiente de regresión simple se expresaba en la fórmula:

$$Y = \alpha + \beta X$$

Un coeficiente de correlación parcial dijimos era una medida de la cantidad de variación explicada por una variable independiente, después que las otras variables han explicado todo lo que podían. En el coeficiente de correlación múltiple estamos interesados en el poder explicativo de un conjunto de variables independientes sobre la variable dependiente $(r_{1.2345})$

Para ambos casos, la ecuación de la regresión toma ahora la siguiente forma:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k$$

Esta es la ecuación más simple, y parte de los mismos supuestos delineados para la ecuación de regresión simple. En la medida en que nos movemos en espacios multidimensionales, la representación geométrica es imposible.

Los coeficientes β, se interpretan de manera distinta que en el caso de regresiones simples, ya que aquí las inclinaciones son varias, y que se obtienen cada una de ellas controlando por cada una de las variables independientes remanentes. Manteniendo X_2 en un valor fijo, β_1 representa la inclinación de la línea de regresión de Y sobre X_1, para el caso en que solamente se estén controlando dos variables. Y así sucesivamente.

Ejemplos:

a) Para encontrar la mejor ecuación linear de predicción y para evaluar la eficiencia predictiva. Jae-On Kim y Frank Kohout (*) presentan el problema de predecir la tolerancia política, a partir de educación, ocupación e ingreso. A través de técnicas de regresión múltiple, el investigador podría estar interesado en determinar el grado de dependencia linear de la tolerancia política sobre la base de la educación, la ocupación y el ingreso de una persona. Supóngase que la tabla de resultados sea la siguiente:

Correlación múltiple	:	.5312
R^2	:	.2822
Error estándar	:	.8604

Variables independientes	B	β par
Educación	.1296	.3889
Ocupación	.0089	.1778
Ingreso	.0018	.0556
(constante A)	2.9889	

(*) Jae-On Kim y Frank J. Kohout: "Multiple regression analysis: subprogram regression", en N. Nie, C.H. Hull, J. Jenkins, et al.: Statistical package for the social sciences, 2a. ed.: McGraw-Hill, N. Y. 1975.

La interpretación en esta caso podría ser la siguiente:

1. La cantidad de variación en tolerancia política, explicada por la operación conjunta de educación, ocupación, e ingreso es del 28.22% de la varianza total.

2. Si el investigador está interesado en predecir los puntajes que un sujeto va a obtener en tolerancia política a partir de las tres variables independientes, aplicará la ecuación de predicción señalada más arriba.

$$Y = 2.9889 + .1296\ (X_1) + .0089\ (X_2) + .0018\ (X_3)$$

Si el sujeto tiene 10 años de educación formal (X_1), un puntaje de 60 en prestigio ocupacional (X_2) y un ingreso de 100 ($10,000) ($X_3$) entonces

$$Y = 2.9889 + .1296\ (10) + .0089\ (60) + .0018\ (100) = 4.9989$$

El error estándar que figura en la tabla (.8604), predice que los puntajes precedidos en la escala de tolerancia política se van a desviar de los valores parámetros en .8604 unidades.

Los valores B en la tabla son coeficientes de regresión parcial, y pueden ser utilizados como medida de la influencia de cada variable independiente, sobre la tolerancia política cuando se controlan los efectos de las otras variables.

Obsérvese en el ejemplo que el coeficiente de correlación múltiple (R) es mayor en magnitud que cualquiera de los r, y esto es evidente desde el momento en que es imposible explicar menos variación agregando variables. El máximo valor relativo del coeficiente total ocurre cuando la intercorrelación entre las variables independientes es igual a 0, de manera que si queremos explicar la mayor cantidad de variación en la variable dependiente que sea posible, deberemos buscar por variables independientes que si bien tienen correlaciones moderadas con la variable dependiente, son relativamente independientes unas de las otras.

Relacionado a la intercorrelación entre variables independientes, es el problema de la multicolinearidad, esto es cuando las variables independientes están estrechamente intercorrelacionadas, tanto las correlaciones parciales como la estimación de los β se hacen muy sensitivas a los errores de muestreo y de medición. Cuando la multicolinearidad es extrema (intercorrelaciones del rango de .8 a 1.0) el análisis de regresión no es recomendable.

b) La regresión múltiple puede ser utilizada también para eva-

luar la contribución de una variable independiente en particular, cuando la influencia de otras variables independientes es controlada. Aquí utilizamos coeficientes de regresiones parciales. Hay dos coeficientes designados, la contribución de cada variable a la variación de la variable dependiente: coeficiente de correlación semi-parcial (part-correlation) y el coeficiente de correlación parcial. El primero se denota como $r_{y(1.2)}$ y el segundo como $r_{(y\ 1).2}$.

El coeficiente semi-parcial es la correlación simple entre el Y original y el residual de la variable independiente X_1 a la cual se le extraen los efectos de la variable independiente X_2, es decir que el efecto de X_2 es sacado solamente de la variable X_1, mediante una regresión linear simple de X_2 sobre X_1, entonces ese residual de X_1 es correlacionado con la variable dependiente Y. En el caso de la tolerancia política uno podría estar interesado en determinar de qué manera el ingreso contribuye a la variación de la tolerancia política aparte de lo que es explicado por educación y ocupación. La tabla siguiente permite calcular los valores de el coeficiente semi-parcial y el coeficiente parcial:

Regresión con dos variables independientes

A	B	C
ED (X_1) y OCU (X_2)	ED (X_1) e ING (X_3)	OCU (X_2) e ING (X_3)
Regress. Mult. (R) .5292	.5118	.4163
R^2 .2800	.2619	.1733

Regresión con tres variables independientes

ED (X_1) y OCU (X_2), e ING (X_3)	
Regresión múltiple (R) :	.5312
R^2 :	.2822

Coeficiente semiparcial:

Su cuadrado es igual a la diferencia entre un R^2 que incluye a las

tres variables independientes (.2822) y a un R^2 que incluye solamente ocupación y educación (.2800). En nuestro caso entonces $R^2_{y(3.12)}$ es igual a .0022, indicando que ingreso solamente contribuye a un .22% de incremento en la variación a lo que ya estaría explicado en términos de variación de tolerancia política por educación y ocupación, en otras palabras que el incremento es trivial, y que se puede ignorar ingreso. Para los casos del coeficiente semiparcial para educación y para ocupación los valores respectivos serían .1089 y .0203, es decir que educación explicaría aproximadamente un 11% de la variación y ocupación un 2%.

El coeficiente parcial

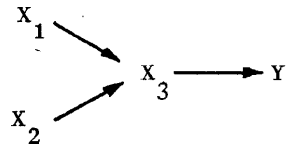
Es la correlación entre los dos residuales el residual de Y y el residual de X_1, para los cuales y en ambos se han extraído los efectos de X_2. El cuadrado de una correlación parcial es el incremento proporcional en la variación explicada debido a X_1, expresada como una proporción de la variación que no está explicada por X_2. El coeficiente de correlación parcial indicaría el grado en que una variable da cuenta del remanente de variación que no dan cuenta las otras variables independientes. En nuestro ejemplo de ingreso la correlación parcial es .0031, es decir, que solamente da cuenta del .31% de la variable dependiente.

c) El análisis de regresiones múltiples para la determinación de relaciones estructurales entre variables.

Se trata aquí de una conjunción de la técnica de regresión múltiple con la teoría causal. La teoría causal especificaría un ordenamiento de las variables que refleja una estructura de eslabones causa-efecto, la regresión múltiple determina la magnitud de las influencias directas e indirectas que cada variable tiene sobre las otras variables, de acuerdo al orden causal presumido. El método de Path Analysis es un método para descomponer e interpretar relaciones lineares entre conjuntos de variables, en los que se parte del supuesto que el sistema causal es cerrado, consistente de causas y efectos encadenados. Las relaciones causales (pathways) se representan con flechas que conectan la causa al efecto.

Cuando se relacionan tres variables, de las cuales una es dependiente (efecto) existen teóricamente seis maneras a partir de las cuales se puede establecer la relación (ver ejemplos en la sección de correlación parcial), con cuatro variables podemos producir 65 diferentes diagramas, etc. La tarea del investigador es seleccionar

de entre los diagramas posibles, aquellos que sean más significativos desde el punto de vista de la teoría sustantiva. Cualquier diagrama, por ejemplo:



Puede ser representado e interpretado en términos de ecuaciones estructurales: una variable a la que una o más flechas apuntan, es interpretada como una función de solamente aquellas variables de<u>s</u> de donde partan las flechas.

Uno de los supuestos principales del path análisis es que todas las relaciones son lineares, que las variables son aditivas y que las relaciones son unidireccionales. Cuando se cumplen esas condicio<u>n</u>es, la función linear toma la forma:

$$X_0 = C_{01} X_1$$

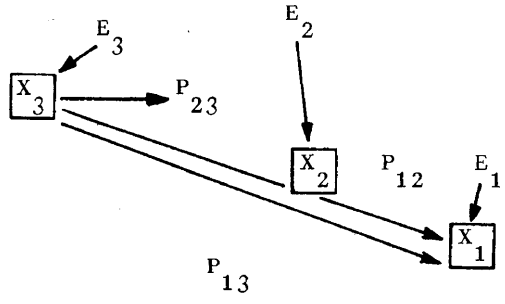
Donde X_1 es la variable independiente, o causa,

X_0 es la variable dependiente o efecto

C_{01} es una constante que expresa la magnitud del cambio en X_0 para cada unidad de cambio en X_1, este coeficiente mide el efecto causal linear, o simplemente el coeficiente efecto.

El path análisis no es una técnica para demostrar causalidad. Es un procedimiento para el análisis de las implicaciones de un conjunto de relaciones causales que el investigador impone, a partir de algunos supuestos técnicos, en el sistema de relaciones.

Consideremos ahora un path análisis de tres variables, X_3, X_2, X_1. Asumiendo que existe un orden en la relación entre las variables digamos $X_3 \geqslant X_2 \geqslant X_1$, y suponiendo que el sistema sea cerrado podemos representar la relación de la siguiente manera:



o por un sistema de ecuaciones lineares tal como:

$$X_3 = E_3$$

$$X_2 = P_{23}X_3 + E_2$$

$$X_1 = P_{13}X_3 + P_{12}X_2 + E_1$$

$$\text{cov}\ (E_3, E_2) = \text{cov}\ (E_3, E_1) = \text{cov}\ (E_2, E_1) = 0$$

Cada E_i representa todos los efectos residuales en las causas de cada X_i y se denominan errores independientes o perturbaciones independientes, o variables latentes. Cada una de estas variables latentes se estiman a partir de cada R^2 por medio de la fórmula $\sqrt{1-R^2}$, donde el coeficiente de correlación múltiple R, es la parte de la ecuación de la regresión en la cual X_i es la variable dependiente y todas las variables que la causan son usadas como predictores.

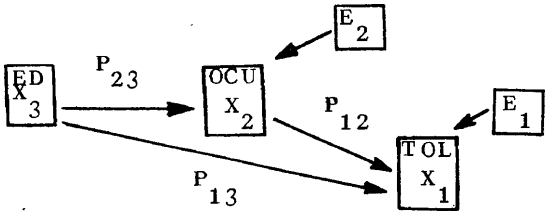
Cada P_{ij} indica un path y pueden ser estimados a partir de las regresiones de los X_i sobre los X_j. En el path que aparece más arriba P_{23} es estimado a partir de la regresión de X_2 sobre X_3, donde $X_2 = B_{23}X_3$.

Y donde P_{13} y P_{12} pueden ser estimados de las regresiones de X_1 sobre X_2 y X_3: $X_1 = B_{13}X_3 + B_{12}X_2$.

Por lo general, dadas n variables en orden $X_n \leqslant \ldots\ldots \leqslant X_3$, $\leqslant X_2, \leqslant X_1$, la estimación de todos los path coeficientes requerirá n-1 soluciones de regresión, en las que se toma cada una de las n-1 variables de orden menor en el diagrama como independientes en sucesión y todas las variables de orden mayor como sus predictores.

Sigamos el mismo ejemplo de Kim y Kohout (op.cit.): tenemos 3 variables: tolerancia (X_1), status ocupacional (X_2) y educación (X_3); si podemos sostener que el grado de tolerancia, probablemente va a estar afectado por el nivel educacional y por el nivel del status ocupacional, y que el status ocupacional del individuo probablemente va a estar influenciado por su nivel educacional, entonces podemos postular un ordenamiento causal débil del tipo $X_2 \geqslant X_2 \geqslant X_1$. Estamos afirmando un juicio de probabilidad en el que no sabemos cómo una variable afecta a la otra. Para continuar con el Path Análisis necesitamos otro supuesto, que el sistema causal es cerrado, que es más difícil de justificar, pero supongamos que esté justificado.

Tenemos entonces un diagrama de la siguiente forma:



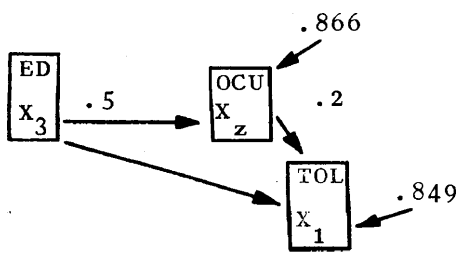
Calculamos los coeficientes de regresión simple para cada uno de los P_{ij} obteniendo los siguientes valores:

$$P_{23} = .5$$
$$P_{12} = .2$$
$$P_{13} = .4$$

Calculamos también los valores E_i que resultan ser:

$$E_2 = .866$$
$$E_1 = .8485$$

El diagrama o path análisis tiene entonces la siguiente forma ahora:



¿Cómo interpretamos este path?

a) Primero examinamos cada subsistema, a través de las variables latentes. Y vemos que el 75% de la variación en ocupación y el 72% de la variación en tolerancia, permanece sin explicar por las relaciones causales explicitadas en el modelo.

b) Identificamos los efectos de educación sobre ocupación: de ocupación sobre tolerancia; y de educación sobre tolerancia. El coeficiente C_{ij} mide los cambios que acompañan a X_i dada una unidad de cambio en X_j, estando controladas todas las causas extrañas. Los datos son los siguientes:

$$C_{23} = P_{23} = .5$$

$$C_{13} = (P_{23})\ (P_{12}) + P_{13} = .5$$

$$C_{12} = P_{12} = .2$$

c) La covariación total entre pares de variables, representadas por la correlación simple, puede ser descompuesta de la siguiente manera:

	OCU,ED (X_2,X_3)	TOL,ED (X_1,X_3)	TOL,OCU (X_1,X_2)
(A) Covariación Original (r_{ij})	.5	.5	.4
(B) b_1: Causal-directa	.5	.4	.2
b_2: Causal-indirecta	0	.1	0
Total Causal $(b_1) + (b_2) = C_{ij}$	.5	.5	.2
(C) No causal (A) - (B)= $r_{ij} - C_{ij}$	0	0	0

Para la relación entre ocupación y educación, el path análisis confirma los supuestos, todas las covariaciones entre los dos son tomadas como causales o genuinas.

La covariación entre educación y tolerancia es también tomada como causal, pero la covariación se descompone entre lo que es mediatizado por ocupación y entre lo que no lo es. Aquí parte de la rela

ción entre educación y tolerancia está mediatizada por una variable interviniente.

La relación entre tolerancia y ocupación, esto es la última columna, está descompuesta en componentes causales y componentes espurios.

Casos especiales en el Path Analysis

Hasta ahora consideramos modelos generales de Path Analysis en los cuales todas las relaciones bivariatas eran asumidas como teniendo una relación causal y donde el sistema como un todo era cerrado. Es posible introducir en el Path Analysis una cantidad de supuestos diferentes. Sin embargo, siempre hay que recordar que cada vez que incorporamos supuestos ambiguos, producimos como resultado un modelo que da lugar a interpretaciones también ambiguas. Hasta ahora representamos las relaciones bivariatas como X → Y, también podemos representar la relación entre las dos variables de las siguientes formas:

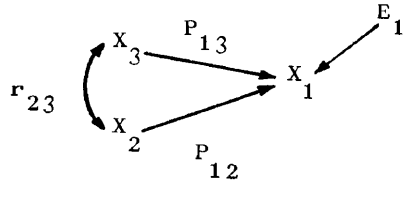
X ⇜⇝ Y : esto significa una correlación no analizada, por lo tanto la relación es ambigua, en el sentido en que la covariación puede ser causal o espuria, y la dirección de la relación puede ser de X a Y o de Y a X

X Y : la ausencia de flecha recta o curva significa que no existe convariación entre X y Y.

X ⌒ Y : la curva simple significa que la relación entre X y Y es completamente espuria o no causal.

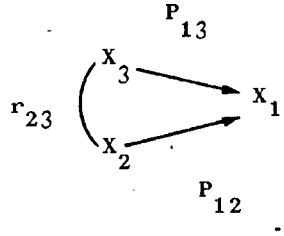
X ⌒→ Y : representa una relación que es parcialmente causal y parcialmente espuria.

La relación representada hasta ahora en los diagramas anteriores, era del tipo X → Y, esto es, asumíamos un orden causal entre las variables. Existen situaciones en las cuales no conocemos la verdadera naturaleza de la relación causal, entre algunas de nuestras variables aunque sí conocemos que existe correlación entre ellas. En este caso, el gráfico tendría la siguiente forma:



Es decir, postulamos relación causal entre X_3 y X_1 y entre X_2 y X_1, pero las variables independientes no están conectadas entre sí por una conexión causal, sino simplemente por su correlación. La estimación de P_{13} y de P_{12} se obtiene a partir de las regresiones en la que X_1 es la variable dependiente y X_2 y X_3 como variables independientes. Las relaciones entre X_2 y X_3 se expresan por un coeficiente de correlación simple. Nótese que el cambio total en la variable dependiente no está definido en el modelo, lo que dificulta la predicción.

Cuando existen suficientes elementos en la teoría que efectivamente permiten asegurar que la covariación entre las variables exógenas no es de naturaleza causal, el modelo puede ser representado de la siguiente forma:



Ahora sí todas las relaciones entre variables pueden interpretarse de manera causal, simplemente porque partimos del supuesto menos ambiguo que en el del diagrama anterior. Aquí en vez de plantear desconocimiento sobre la naturaleza de la covariación, planteamos que la covariación entre X_3 y X_2 es de naturaleza no causal, es decir que X_3 no causa variación en X_2 y viceversa. De esta manera es posible hacer predicciones en relación a los cambios que una unidad en X_3 o en X_2 producirán en X_1.

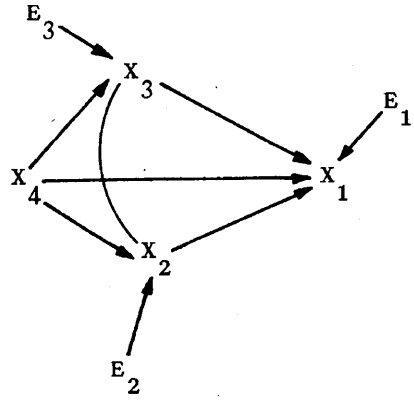
Kim y Kohout presentan en otro ejemplo con cuatro variables y un conjunto de supuestos fuertes, lo que da lugar a interpretaciones menos ambiguas:

Se trata de un esquema en el que se relaciona sexo (X_4) y accidentes de tráfico (X_1), controlando por cantidad de kilómetros conducidos al año (X_3) y frecuencia de conducción en horas de mucho tráfico (X_2).

Los supuestos causales es que el sexo puede afectar tanto la cantidad de kilometraje recorrido como las condiciones en las que se maneja, las cuales a su vez van a determinar las tasas individuales

diferenciales de accidentes de tráfico. El investigador no tiene ningún supuesto teórico que le permita relacionar en forma causal el kilometraje recorrido con las condiciones de manejo. Asimismo, ni el total de kilometraje recorrido, ni las condiciones de manejo, se considera sean totalmente explicadas por sexo, sino que a su vez existen una serie de factores que pueden causar ambos.

El modelo del path adquiere entonces la siguiente forma:



Donde las estimaciones de P_{34} y de P_{24} pueden realizarse a partir de coeficientes de correlación simple, y los P_{13}, P_{12}, y P_{14} por las regresiones de X_1, sobre X_2, X_3, y X_4.

El coeficiente de covariación residual entre X_2 y X_3 se obtiene a partir de : $r_{23} - (P_{34})(P_{24})$. Los coeficientes para los E_i se obtienen respectivamente de la siguiente forma:

$$E_1 = \sqrt{1 - R^2_{1.234}}$$

$$E_2 = \sqrt{1 - R^2_{2.34}}$$

$$E_3 = \sqrt{1 - R^2_{3.24}}$$

IX. REGRESIONES CON VARIABLES MUDAS (DUMMY VARIABLES)

Este es un caso especial de regresión, en el cual introducimos mediciones a nivel nominal en la ecuación de la regresión. Estas variables mudas se obtienen tratando a cada categoría de la variable nominal como si fuera una variable por separado, asignando puntajes arbitrarios según la presencia o ausencia del atributo en cuestión. Por ejemplo, si en afiliación política tenemos 3 partidos políticos: radical, demócrata cristiano y conservador, cada uno de esos partidos o categorías representa 1 de 3 variables dicotómicas, entonces los puntajes 1 a 0 pueden ser asignados a cada "variable". Si un sujeto tiene afiliación radical, entonces su puntaje en radical será 1, su puntaje en demócrata cristiano será 0, y su puntaje en conservador será 0. Los valores 0 y 1 son tratados como variables intervalares e incluidas así en la ecuación de la regresión. Sin embargo, por un problema de álgebra*, una de las variables mudas debe ser excluida de la ecuación de regresión. De hecho, la variable muda excluida, va a actuar ahora como punto de referencia a partir del cual los valores en cada una de las otras variables mudas va a ser interpretado. Cada categoría ahora es representada por una combinación de las i variables mudas. Supongamos en nuestro caso que la categoría de referencia sea otro partido, tendríamos entonces la siguiente distribución

---

* La inclusión de todas las variables mudas resultantes de categorías nominales hace que las ecuaciones normales no puedan ser resueltas, ya que la inclusión de las últimas categorías está completamente determinada por los valores de las primeras categorías ya incluidas en la ecuación.

de puntajes:

	X_1	X_2	X_3
Radical	1	0	0
Dem. cristiano	0	1	0
Conservador	0	0	1
Otro	0	0	0

Si otro ha sido elegido como categoría de referencia, la ecuación de la regresión puede ser escrita entonces como:

$$Y = A + B_1X_1 + B_2X_2 + B_3X_3$$

donde los casos de la categoría "otros" pueden ser predecidos mediante:

$$Y = A$$

los radicales por:

$Y = A + B_1X_1$: y en la medida que el valor de radicales X_1 es 1, entonces:

$$Y = A + B_1$$

Los valores esperados para cada una de nuestras categorías serán entonces:

	Y
Radical	$A + B_1$
D.cristiano	$A + B_2$
Conservador	$A + B_3$
Otro	A

Análisis de la varianza unidireccional con variables mudas

El análisis de la varianza unidireccional puede ser obtenido a través de diferentes subprogramas en el SPSS: los subprogramas Anova, Oneway y Breakdown (ver sección análisis de la varianza). En estos tres subprogramas las variables estran como variables nominales, no introduciéndose la creación de variables mudas.

Sin embargo, el investigador puede desear un análisis de la varianza unidireccional con el subprograma regression. Para ello debe

crear su conjunto de variables mudas según el sistema explicado más arriba e instruir al programador para que incluya las instrucciones pertinentes para la creación de variables mudas en el SPSS.

El out-put del subprograma regression en su porción referente al análisis de la varianza unidireccional, tiene la siguiente forma, en la cual introducimos cálculos ficticios para las tres variables usadas en nuestro ejemplo anterior con la variable dependiente actitud frente a la nacionalización del petróleo.

R múltiple	.5844	Análisis de var.	D.F.	Suma Cuadr.	F
R^2	.3416	Regresión	3	56.3529	16.5993
Error estándar	1.0638	Residual	96	108.6371	

Variable en la ecuación

Variable	B	Beta	Error estándar B	F
D.cristiano	1.3156	.4435	.4135	10.121
Radical	-.3961	-.1497	.3795	1.089
Conservador	-.9444	-.1441	.6393	2.183
(Constante)	2.444			

El valor F de 16.5993 tiene una probabilidad mayor que .001, es decir, que las diferencias son muy significativas para el conjunto de partidos. El R^2 es equivalente al coeficiente de correlación múltiple que se derivan del coeficiente de correlación eta (ver correlación simple), y su valor indica que el 34% de la actitud frente a la nacionalización del petróleo depende o se explica por la afiliación política.

Los promedios para cada categoría pueden ser obtenidos a partir de la columna B de "variables en la ecuación" que es el out-put de la regresión:

-Actitud frente a la nacionalización del petróelo: Y = 2.444

- radical: $Y = 2.444 + 1.3156 = 3.76$
- D. cristiano: $Y = 2.444 + (-.3961) = 2.05$
- conservador: $Y = 2.444 + (-.9444) = 1.50$

Regresiones con variables mudas para dos o más variables categorizadas

Las ecuaciones de predicción para dos variables nominales (representadas por dos conjuntos de variables mudas) es la siguiente:

$$Y = A + B_1X_1 + B_2X_2 + B_3X_3 + B_4E_1$$

Donde los X_1 representan una variable nominal con 4 categorías y E_1 una categoría de una variable nominal dicotómica.

El valor predictivo para cada celda de la matriz estará dada por la siguiente tabla, siguiendo nuestros ejemplos anteriores:

	Varón	Mujer
Radical	$A + B_1 + B_4$	$A + B_1$
D.Cristiano	$A + B_2 + B_4$	$A + B_2$
Conservador	$A + B_3 + B_4$	$A + B_3$
Otro	$A + B_4$	A

Lo que ocurre ahora es que las categorías mujer y otro actúan como categorías de referencia.

Análisis de la varianza multidireccional con variables mudas

La regresión múltiple con n variables mudas puede ser utilizado para computar análisis de la varianza. Cuando se desea computar análisis de la varianza con las variables nominales sin recurrir a variables mudas se recomienda el subprograma Anova (ver más adelante en "análisis de la varianza").

Cuando se utilizan variables mudas y queremos realizar análisis de la varianza con el subprograma Regression es necesario agregar para el caso de dos factores (afiliación política y sexo) los

efectos de interacción, es decir necesitamos crear tres nuevas variables mudas en nuestro ejemplo: (X_1E_1), (X_2E_1), (X_3E_1), y donde la ecuación de la regresión múltiple tendrá ahora la siguiente forma:

$$Y = A + B_1X_1 + B_2X_2 + B_3X_3 + B_4E_1 + B_5(X_1E_1) + B_6(X_2E_1) + B_7(X_3E_1)$$

Esta regresión representa el modelo saturado donde todos los términos de interacción posible están incluidos.

Los valores predictivos para el modelo saturado se obtienen de la siguiente tabla:

	Varón	Mujer
Radical	$A + B_1 + B_4 + B_5$	$A + B_1$
D. cristiano	$A + B_2 + B_4 + B_6$	$A + B_2$
Conservador	$A + B_3 + B_4 + B_7$	$A + B_3$
Otro	$A + B_4$	A

Para dos variables nominales A y B, la estrategia del análisis de la varianza sigue lo que se llama modelo clásico de análisis de la varianza en el cual ni los factores A y B (afiliación política y sexo) son ortogonales, esto es si las frecuencias en las celdas son proporcionales a las frecuencias marginales de afiliación política y sexo la suma de 2a y de 2b será simplemente la suma de los cuadrados debido a cada factor y será igual a la suma de los cuadrados debidos a efectos aditivos. Si A y B no son ortogonales, los efectos de A se confundirán con los efectos de B, y la suma de 2(a) y de 2(b) no será igual a la suma de los efectos aditivos. La siguiente tabla ilustra el modelo clásico:

Fuente de variación	Suma de cuadrados	Df	F
(1) Suma de cuadrados debido a A y B modelo saturado.	$SS_y(R^2_{A,B,AB})$	$K=k_1+k_2+k_1k_2$	$\frac{(1)/K}{(4)(N-K-1)}$
(2) Suma de cuadrados debidos a A y B modelo aditivo	$SS_y(R^2_{A,B})$	k_1+k_2	$\frac{(2)/(k1 + K2)}{(4)/(N-K-1)}$

(a) Suma de los cuadrados debidos a B ajustados por B	$SS_y(R^2_{A,B}-R^2_B)$	k_1	(2a)/k1 (4)/(N-K-1)
(b) Suma de los cuadrados debidos a B ajustados por A	$SS_y(R^2_{A,B}-R^2_A)$	k_2	(2b)/k2 (4)/(N-K-1)
(3) Suma de los cuadrados debida a la interacción	$SS_y(R^2_{A,B,AB}-R^2_{A,B})$	k_1k_2	(3)/k1k2 (4)/(N-K-1)
(4) Suma de los cuadrados residuales	$SS_y(1-R^2_{A,B,AB})$	N-K-1	

Los significados de este cuadro serán analizados en la sección siguiente que corresponde a análisis de la varianza.

X. ANALISIS DE LA VARIANZA Y DE LA COVARIANZA (SUBPROGRAMAS ANOVA Y ONEWAY)

El análisis de la varianza es una técnica estadística utilizada para la determinación de asociación entre dos o más variables. El análisis de la varianza simple (o unidireccional) se refiere a situaciones en las cuales el investigador está interesado en determinar los efectos de una variable o factor (medido a nivel nominal) sobre una variable dependiente (o variable criterio) que debe estar medida a nivel intervalar. Si el investigador está interesado en el efecto simultáneo de varios factores, entonces el análisis de la varianza es bivariato o n-maneras.

En el análisis de la covarianza el investigador está interesado en los efectos tanto de variables no métricas como de variables métricas.

Análisis de la varianza simple

Habíamos visto en la sección correspondiente a confiabilidad de la diferencia entre estadísticos algunas pruebas para la determinación de diferencias significativas en pares de medias muestrales (prueba t de student). El análisis de la varianza simple, sobre todo en lo referido a la prueba F de Fisher se emplea de manera similar, aunque ahora para decidir estadísticamente, si la serie de datos entre n pares de medias son lo suficientemente diferentes entre sí para permitirnos el rechazo de la hipótesis de que esas medias surgieron por efectos del azar, de una población única.

La varianza total en muestras combinadas tiene dos componentes: un componente representado por la varianza interserial (es decir, por la suma de los desvíos al cuadrado de las medias de las sub

muestras con respecto a la media total), y un componente representado por la varianza intraserial (es decir por la suma de los desvíos al cuadrado dentro de cada una de las series de datos).

Es decir que la varianza total se descompone en dos componentes: una intervarianza y una intravarianza: la prueba F es la razón entre la intervarianza y la intravarianza. $F = \frac{\text{intervarianza}}{\text{intravarianza}}$

La intervarianza se estima sobre la base de k medias, las que pueden ser consideradas como k datos independientes. Puesto que se quiere una estimación de la varianza de la población la suma de los desvíos al cuadrado se divide por los grados de libertad. Es decir, en el caso de la intervarianza sería k-1. La fórmula para el cálculo de la intervarianza será entonces:

$$\text{Intervarianza} = \frac{\Sigma n_s d_s^2}{k-1}$$

Donde:

n: cantidad de casos en cada una de las series

d: desvío de la media serial con respecto a la media total $(M_s - M_t)$

k: cantidad de series

La intravarianza se estima sobre la base de las medias muestrales o seriales, de donde:

$$\text{Intravarianza} = \frac{\Sigma x_s^2}{k(n-1)}$$

Donde:

x: desviación del puntaje de su media muestral

La confrontación de los valores F se hace con tablas especiales, que señalan el nivel de significación. Las F mayores o iguales a los distintos niveles indican el grado de confianza con el cual se puede rechazar la hipótesis nula. La prueba F indica solamente si existe o no una diferencia. Para encontrar dónde se encuentra ésta se hace necesaria una investigación y tests sucesivos, por ejemplo, pruebas t.

Un ejemplo de análisis de la varianza simple:

Se trata de saber si el nivel de ingresos difiere por tipo de rama ocupacional. La variable dependiente o criterio será "ingreso" en escudos (los datos corresponden a Chile en 1968), por lo tanto, la variable está medida a nivel racional. La variable independiente o factor es sector, distinguiendo tres categorías: industria, construcción y servicios.

El siguiente cuadro muestra la distribución:

Cuadro I. Distribución de ingreso por tipo de ocupación

	Industria	Construcción	Servicios
0- 100	7	1	4
100- 200	12	4	6
200- 300	34	12	19
300- 400	62	30	31
400 -500	51	24	24
500- 600	20	12	12
600- 700	22	15	8
700- 800	21	8	9
800- 900	17	8	7
900-1000	9	5	5
1000-1400	5	5	8
1400-1800	8	4	6

La hipótesis nula a probar es que no existen diferencias entre los distintos promedios de ingresos según las tres ramas ocupacionales.

Los resultados obtenidos son los siguientes:

- Suma de las desviaciones al cuadrado de los datos individuales respecto a la media total..... 53.471.950

- Suma de las desviaciones al cuadrado

de los datos individuales respecto a
la media serial..... 178.652

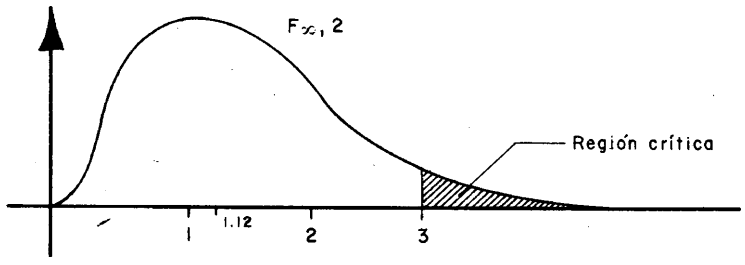
- Suma de los desviaciones al cuadrado
de las medias seriales respecto a la
media total 53.293.342

Estimación de la intervarianza: 100.175

Estimación de la intravarianza: 89.326

F = 1.12

La distribución para F con 2° de libertad en él numerados y 532 en el denominador nos dice que hay una probabilidad de 5% de obtener un valor de F más alto que 3. (3)



Como el valor obtenido para F cae fuera de la región crítica, no podemos rechazar la hipótesis nula de que las medias de las poblaciones son iguales. Por lo tanto, se concluye que los distintos sectores de la economía considerados no difieren en los niveles de ingreso de los trabajadores.

De haber sido F lo suficientemente grande como para caer en la región crítica, se hubiera rechazado la hipótesis nula, correspondiendo un análisis ahora de las diferencias entre pares de medias seriales.

El subprograma Breakdown computa análisis de la varianza simple o unidireccional, computando medias, desviaciones estándar, intervarianza, intravarianza y valores F.

Ya vimos más arriba que el subprograma Regression computaba también análisis de la varianza, aunque son variables mudas.

El subprograma Oneway computa también análisis de la varianza simple, esto es con una sola variable independiente o factor. La ventaja de este subprograma sobre los anteriores es que acepta hasta un máximo de 20 variables dependientes; claro que hay que especificar

una sola variable independiente, y los out-puts, son de dos en dos.

Análisis de la varianza n-dimensional

En el análisis de la varianza simple las series de datos se diferenciaban en base a un solo factor. En un problema de clasificación bidireccional existen dos bases distintas para la clasificación; en un análisis tridireccional habrá tres bases distintas para la clasificación, etc.

Un problema de análisis bidireccional típico puede ser la comparación de distintos métodos para la enseñanza de lectura y escritura en la escuela primaria, combinados con distintos tipos de maestros. A este respecto, la comparación de cuatro métodos (global, palabra generadora, silábico y "novo"), con 3 clases distintas de maestros (normales fiscales, normales privados, especiales), genera 12 combinaciones posibles de método-maestro. Si se quisiera incluir una variable o factor adicional, digamos sexo, las combinaciones posibles aumentarían a 24.

Las fuentes de la varianza en una clasificación bidireccional, son ahora:

a) una varianza relativa al tipo de método de enseñanza;

b) una varianza relativa al tipo de maestro;

c) una varianza de la interacción entre determinado tipo de maestro y determinado tipo de método de enseñanza;

d) y una varianza residual o intravarianza, que constituye la estimación básica o residual de la varianza una vez que las tres fuentes de variación han sido eliminadas. Esta varianza residual puede ser considerada entonces como una varianza del error, puesto que representa las influencias de los factores desconocidos o no controlados.

Los grados de libertad en clasificaciones bidireccionales dependen de la fuente u origen de la varianza y son:

Entre líneas	r-1
Entre columnas	k-1
Interacción	(r-1) (k-1)
Dentro de las submuestras	rk (n-1)
Total	N-1

El out put del subprograma Anova provee para una tabla como la ilustrada, los siguientes datos de análisis de la varianza:

Efectos principales	Suma de Cuadrados	DF	F	Nivel de Sign.
Métodos	--------	--	-	--------
Maestros	--------	--	-	--------
Método-maestro	--------	--	-	--------
Residual	--------	--	-	--------
Total	--------	--	-	--------

Para tablas más complejas (digamos tres variables factor), el programa proveerá los cálculos correspondientes a los efectos principales de cada uno de los efectos, 3 efectos de interacción bivariata por ejemplo, método-maestro; método-sexo, maestro-sexo; el efecto de interacción de los tres factores (método-maestro-sexo), la varianza residual y la varianza total.

Acorde con los valores, el investigador acepta o rechaza sus hipótesis nulas (relativas a diferencia entre métodos de enseñanza, tipo de maestros, sexo de los mismos; además de las interacciones método-maestro; método-maestro-sexo, etc.). Hay que recordar que para la selección de las combinaciones o de métodos o de maestros especiales, una vez que se encuentran valores F significativos, es necesario, aplicar pruebas t, a fin de seleccionar aquellos que son más afectivos para el aprendizaje de los niños.

La tabla que figura en las páginas 49 y 50, resume las distintas fuentes de variación para problemas n-direccionales.

En el tipo de ejemplo que describimos (métodos-maestros), la matriz de datos a partir de la cual se realiza el análisis de la varianza se representa en el cuadro de la página siguiente.

El subprograma Anova permite cálculos de análisis de la varianza para un máximo de 5 factores o variables independientes en cada diseño. Como opciones permite análisis de la covarianza (hasta 5 covariaciones) y una tabla de análisis de clasificaciones múltiples.

El subprograma Anova es aplicable tanto para diseños ortogonales (igual número de frecuencias en cada una de las celdas) como para diseños no ortogonales (distintos números de frecuencias en las celdas). Incluso el programa puede considerar algunas celdas vacías.

El subprograma Anova también puede producir una tabla de análisis de clasificación múltiple, por medio de la cual los resultados

Líneas (maestros)	Sujetos	Columnas (métodos) 1	2	3	4	Suma de las líneas	Medias de las líneas
A	1	X_{a1}	X_{a2}	X_{a3}	X_{a4}		
	2	.	.	.	.		
	3	.	.	.	.		
	4	.	.	.	.		
	.	.	.	.	.		
	.	.	.	.	.		
	Σ	ΣX_{a1}	ΣX_{a2}	ΣX_{a3}	ΣX_{a4}	ΣX_a	
	M	M_{a1}	M_{a2}	M_{a3}	M_{a4}		ΣM_a
B	1	X_{b1}	X_{b2}	X_{b3}	X_{b4}		
	2	.	.	.	.		
	3	.	.	.	.		
	4	.	.	.	.		
	.	.	.	.	.		
	.	.	.	.	.		
	Σ	ΣX_{b1}	ΣX_{b2}	ΣX_{b3}	ΣX_{b4}	ΣX_b	ΣM_b
	M	M_{b1}	M_{b2}	M_{b3}	M_{b4}		
C	1	X_{c1}	X_{c2}	X_{c3}	X_{c4}		
	2	.	.	.	.		
	3	.	.	.	.		
	4	.	.	.	.		
	.	.	.	.	.		
	.	.	.	.	.		
	M	X_{c1} M_{c1}	X_{c2} M_{c2}	X_{c3} M_{c3}	X_{c4} M_{c4}	ΣX_c	ΣM_c
Suma de las columnas (ΣX_k)		ΣX_1	ΣX_2	ΣX_3	ΣX_4	ΣX_{ij}	M_T
Medias de las columnas (M_k)		M_1	M_2	M_3	M_4		Media Total

del análisis de la varianza son expuestos de manera más específica. Este método es particularmente útil cuando los efectos de interacción no son significativos y cuando los factores son variables nominales o atributos que no han sido manipulados experimentalmente, y por lo tanto, pueden estar intercorrelacionados. Para dos o más factores interrelacionados puede ser importante conocer el efecto neto de cada variable, cuando se controlan los otros factores. Seguimos el ejemplo de Kim y Kohout (op.cit.) para ilustrar la interpretación de una tabla completa de análisis de clasificación múltiple:

La variable dependiente es el salario semanal de los empleados de una industria. Los factores son sexo y raza. En la medida que se sospecha cierto grado de discriminación social, el investigador está interesado en los efectos de los factores raza y sexo. Se sabe que dos variables adicionales: nivel de educación y duración en el empleo determinan el nivel de los salarios, así que ambas son introducidas como variables. Los resultados obtenidos son los siguientes:

Clasificación de análisis múltiple: salario por sexo, raza, con educación y duración en el empleo.

Media total: 100 dólares

	Desviaciones de la media total		
Variables	No ajustada	Ajustada por independientes	Ajustada por independientes y covariables
Raza			
1- Blancos	+ 10	6	4
2- No blancos	- 40	-24	-16
(eta y beta)	(.632)	(.384)	(.253)
Sexo			
1- Varones	12	8	6
2- Mujeres	-18	-12	-9
(eta y beta)	(.465)	(.310)	(.232)
R Múltiple	....	.648	.866
R^2	....	.420	.750

La primera columna expresa las medias en cada categoría, como desviaciones de la media total.

Las medias en la segunda columna, expresan las medias de cada

categoría (también como desviaciones de la media total), pero ahora ajustadas. Las disminuciones en los valores indican que en el contexto del empleo, sexo y raza están relacionados (cada factor disminuye cuando lo ajustamos por el otro factor). Los valores indican que los empleados varones tienden a ser blancos, mientras que las empleadas mujeres tienden a ser no blancas.

La tercera columna indica que cuando se introduce educación y duración en el empleo, la influencia del sexo y de la raza disminuye, aunque aún persiste la discriminación.

Los beta y eta que figuran en la tabla ayudan a una mejor interpretación de la tabla. Los valores de la primera columna son valores eta, mientras que los de la segunda y la tercera son beta parciales. Comparándolos vemos que tanto para el caso de sexo como en raza (en este último con mayor intensidad), la relación decrece a medida que se introducen más variables de control.

La correlación múltiple indica la relación entre salario y como variable dependiente y los efectos aditivos de sexo y raza en la segunda columna, y de sexo, raza, educación y antigüedad en la última columna.

XI. ANALISIS FACTORIAL (SUBPROGRAMA FACTOR)

El análisis factorial es una técnica matemática cuyo objetivo más amplio es el descubrimiento de las dimensiones de variabilidad común existentes en un campo de fenómenos. Cada una de estas dimensiones de variabilidad común recibe el nombre de factor. El razonamiento subyacente, es el siguiente: si tenemos un conjunto de fenómenos, y si cada fenómeno varía independientemente de los demás, entonces habrá tantas dimensiones de variación como fenómenos. Por el contrario, si los fenómenos no varían independientemente, sino que hay ciertas dependencias entre ellos, entonces encontraremos que las dimensiones de variación serán menores que los fenómenos. El análisis factorial, a través de una serie de procedimientos, nos permite detectar la existencia de ciertos patrones subyacentes en los datos de manera que éstos puedan ser reagrupados en un conjunto menor de factores o componentes.

Hay cuatro pasos fundamentales en el análisis factorial:

A) preparación;

B) factorización;

C) rotación;

D) interpretación

Dentro de cada uno de los pasos existen procedimientos u opciones que se irán detallando en la medida en que desarrollemos cada uno de esos pasos.

A) Preparación

Consiste tanto en el problema a tratar, cuanto a la formulación de hipótesis y recolección de datos. El tipo de variables que el investigador utilice tendrá importancia fundamental tanto en lo referente a los factores como a la interpretación. La mayoría de las técnicas analíticas requieren de variables intervalares al menos, aunque es posible utilizar algunas de las medidas de asociación que discutimos en otras secciones de este trabajo. Lo importante es que el resultado de este primer paso de preparación es una matriz de correlaciones, que adquiere ya sea la forma de un triángulo o de un cuadrado:

Variables	1	2	3	4	5	6	7	8	n
1	r_{11}	r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	r_{17}	r_{18}	r_{1n}
2		r_{22}	r_{23}	r_{24}	r_{25}	r_{26}	r_{27}	r_{28}	r_{2n}
3			r_{33}	r_{34}	r_{35}	r_{36}	r_{37}	r_{38}	r_{3n}
4				r_{44}	r_{45}	r_{46}	r_{47}	r_{48}	r_{4n}
5					r_{55}	r_{56}	r_{57}	r_{58}	r_{5n}
6						r_{66}	r_{67}	r_{68}	r_{6n}
7							r_{77}	r_{78}	r_{7n}
8								r_{88}	r_{8n}
.								.....	.
.									.
.									r_{nn}

El otro lado del triángulo para completar el cuadrado representa los mismos números o correlaciones, ya que la correlación r_{12} es idéntica a r_{21} etc. La diagonal designa la correlación de la variable consigo misma.

El investigador tiene una opción en términos de preparación de la matriz de correlaciones: se trata de lo que se da en denominar Q-factor analysis o r-factor análisis.

Si el análisis factorial se aplica a la matriz de correlaciones de unidades, donde por unidad se entiende el objeto, persona, etc.,

que detenta una característica o conjunto de características (es decir, donde correlacionamos pares de unidades), entonces se trata de Q-factor analysis.

Si las correlaciones se hacen entre variables entre cada par de características o atributos, la técnica se denomina R-factor analysis.

En términos de preparación de matriz los usuarios del SPSS, en realidad no tienen esta opción ya que el subprograma factor únicamente opera con el análisis de tipo R. Lo importante a destacar en esta sección es que el subprograma acepta como imput, datos brutos, matrices de correlaciones o matriz factorial.

B) Factorización

La factorización trata de poner de manifiesto por métodos matemáticos cuántos factores comunes es preciso admitir para explicar los datos originales o la matriz de intercorrelaciones.

Por este procedimiento surgen "nuevas variables" o factores que pueden ser definidos como transformaciones matemáticas exactas de los datos originales (análisis de componentes principales), o a través de supuestos inferenciales acerca de la estructura de las variables y de su fuente de variación (análisis factorial clásico o de factores inferidos). Ya sea que los factores sean definidos o inferidos, los factores iniciales son extraídos de tal manera que sean independientes los unos de los otros, esto es, factores que sean ortogonales. En este segundo estadio importa más la reducción de la matriz o de dimensiones que la localización de dimensiones significativas.

El análisis de los componentes principales no requiere ningún supuesto acerca de la estructura subyacente al conjunto de variables. Simplemente, trata de encontrar la mejor combinación linear de variables, tal que dará cuenta de una mayor proporción de la varianza que cualquier otra combinación linear posible. El primer componente principal es entonces, el mejor conjunto de relaciones lineares entre los datos; el segundo componente es la segunda combinación linear tal que no está correlacionada con el primer componente (es decir es ortogonal al primer componente); el segundo factor da cuenta de la varianza residual no explicada por el primer factor; el resto de los componentes es definido en forma similar, siendo los componentes tantos hasta cuando se haya explicado totalmente la varianza.

El modelo del componente principal puede ser expresado como:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + a_{j3}F_3 + ... + a_{jn}F_n$$

Donde:

a_{ji}: coeficiente de regresión múltiple estandarizado de la variable j sobre el factor i;

F_i: factores definidos

El análisis de factores inferidos está basado en el supuesto de que las correlaciones empíricas son el resultado de alguna regularidad subyacente a los datos. Se supone que cada variable está influenciada por varios determinantes, algunos de los cuales son compartidos por otras variables (determinante común) y por otros determinantes que no son compartidos por ninguna de las otras variables en el modelo (determinante único). Se supone entonces, que las correlaciones entre variables son el resultado de variables compartiendo determinantes comunes. Se espera por lo consiguiente, que el número de determinantes sea menor que el número de variables.

El modelo se expresa ahora de la siguiente forma:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + ... + a_{jm}F_m + d_jU_j$$

Donde:

U_j: factor único para la variable j

d_j: coeficiente de regresión estandarizado de la variable j sobre el factor único j.

Se asume en el modelo que $r_{(F_i, U_j)} = 0$, y que $r_{(U_j, U_k)} = 0$ es decir que el factor único es ortogonal a todos los factores comunes y a todos los factores únicos asociados a otras variables.

Si se denota la existencia de una correlación entre cualesquiera dos variables, esta correlación es asumida como producto de factores comunes; en otras palabras, que la correlación parcial entre las dos variables, controlando por el factor común, dará por resultado 0.

Por la técnica del análisis factorial buscamos especificar un número hipotético mínimo de factores, tal que, todas las correlaciones parciales entre el resto de las variables devenga 0.

La varianza residual -es decir, la varianza que no se explica por los factores comunes-, y la determinación de las comunalidades es uno de los problemas más complejos del análisis factorial.

Los diferentes métodos para la factorización, están basados en distintos procedimientos para la estimación de las comunalidades, y el investigador debe estar consciente de las ventajas o desventajas de uno u otro para su diseño de investigación en particular.

Métodos de factorización en el SPSS

Existen disponibles 5 métodos de factorización: a) factorización principal sin interacción (PA1); b) factorización principal con interacción (PA2); c) factorización canónica de Rao (RAO); d) alfa factorización (ALPHA) y e) imagen factorización (IMAGE).

Los cinco métodos tienen de común las siguientes características:

1) todos los factores son ortogonales;

2) los factores son colocados en orden según su importancia;

3) el primer factor es comúnmente el factor general (es decir, tiene un factor de carga significante en cada variable); el resto de los factores tienden a ser bipolares (algunos factores de carga son positivos y otros negativos).

a) Factorización principal sin interacción

Se compone de dos métodos separados, según el usuario decida por:

1) reemplazar la diagonal principal de la matriz de correlaciones por estimaciones de comunalidad; o

2) la diagonal principal de la matriz no se altera.

Que la diagonal se reemplace o no, depende que el investigador haya extraído sus factores iniciales, ya sea por el método de factores definidos o por el método de factores inferidos.

Cuando se utilizan estimaciones de comunalidad, estamos asumiendo la existencia de un factor único (U_j) y al reemplazar la diagonal estamos extrayendo los factores únicos de cada variable, analizando solamente las porciones remanentes de las mismas.

Cuando la diagonal de la matriz no se altera, los factores ini

ciales son definidos, consecuentemente los factores principales son calculados según los métodos especificados más arriba en la sección de factores definidos. En la matriz de componentes principales, los valores de peso asociados a cada componente representan la cantidad de varianza total que es explicada por el componente o factor. Esta estadística es calculada por el programa. En términos de fórmula uno puede representar entonces la varianza total y cada uno de los factores según la siguiente fórmula:

$$\sigma_T^2 = \sigma_a^2 + \sigma_b^2 + \ldots\ldots + \sigma_n^2$$

donde: σ_T^2 : varianza total

σ_i^2 : varianza explicada por el factor i.

A menos que haya indicación del usuario, el programa imprime y retiene únicamente componentes cuyo valor de peso sea igual o mayor que 1.0. El número de componentes significativos que van a ser retenidos para la rotación final, son entonces determinados especificando un criterio mínimo de valor de peso.

b) Factorización principal con interacción

Es una modificación del primer método, solamente que aquí hay dos procedimientos que se hacen automáticamente: a) la diagonal principal es reemplazada con estimaciones de comunalidad; y b) las estimaciones de comunalidad, son corregidas por un proceso de interacción en el que se van reemplazando los elementos en la diagonal principal de manera tal que las diferencias entre dos comunalidades sucesivas sea negligible. Este méotodo es el más recomendable para usuarios no familiarizados con los métodos de factorización.

c) Factorización canónica de Rao

Parte de los mismos supuestos del método clásico de factorización, y centra el problema alrededor de la estimación de la varianza única mediante una estimación de parámetros poblacionales a partir de datos muestrales.

Este tipo de factorización aplica un test de significación para el número de factores requeridos tal que la cantidad de factores requeridos por los datos y los factores hipotetizados no se desvíen significativamente del azar.

d) Alfa-factorización

Las variables incluidas en el modelo son consideradas como una muestra del universo de variables. El propósito de esta factorización es definir los factores de manera tal que tengan un máximo de generalidad. Su aplicación tiene que ver con inferencia teórica más que estadística.

e) Imagen-factorización

Es un método bastante complicado desarrollado por Guttman para la determinación de comunalidades verdaderas. Por este método se obtiene una aproximación sobre la exacta proporción de la parte de la variable explicada por factores comunes y la parte de la variable explicada al factor único.

C) Rotación

La rotación es un procedimiento por el cual se trata de encontrar una estructura tal que un vector aparezca como una función de un mínimo número de factores.

Este tercer último paso en la computación de un análisis factorial, contiene diversas soluciones para la búsqueda de la mejor configuración, soluciones que dependen de los intereses teóricos y pragmáticos del investigador.

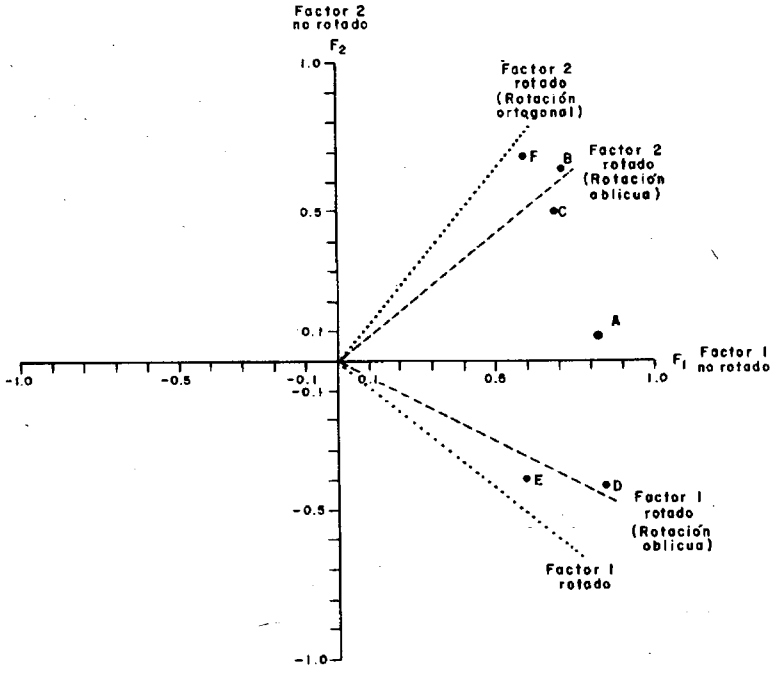
Para comenzar éste debe decidirse por un método de rotación ortogonal o por un método de rotación oblicuo. Los métodos ortogonales proporcionan factores terminales no correlacionados, mientras que en los métodos oblicuos estos pueden estar correlacionados. En términos gráficos una rotación ortogonal es una en la cual los ángulos entre los ejes se mantienen a 90°. En las rotaciones oblicuas los ángulos pueden ser agudos u obtusos.

La extracción de factores, tal como fueron descritas en la sección de factorización da lugar a una solución inicial que puede o no resultar en una estructura con significado. Por medio de la rotación de los ejes y la solución de ecuaciones lineares simultáneas es posible interpretar de forma más adecuada la configuración de los resultados.

Una representación gráfica, en la forma de un espacio de coordenadas cartesianas nos ayudará a entender en forma más clara el significado de las rotaciones. Vamos a representar solamente dos factores, ya que se trata de ejes cartesianos. Es posible representar es

pacialmente 3 factores (a través de planos), pero es imposible representar gráficamente más de tres.

F_1 y F_2 están representados por los ejes de referencia. Los factores de carga se representaban en el modelo por puntos (si el factor de carga en F_1 de una variable es .20 y en F_2 de .75, se avanza por el eje de las F_1 hasta la posición .20 y por el eje de las F_2 hasta .75; la línea de intersección señalará el punto para esa variable. Y así sucesivamente con las otras.



Los valores que figuran en el gráfico corresponden a la siguiente matriz con factores de carga no rotados.

Variables	F_1	F_2
A	.83	.10
B	.75	.63
C	.70	.50
D	.85	-.42
E	.60	-.40
F	.69	.57

La interpretación más elemental es que cuando más cerca están

los puntos, más relacionadas están las variables. Sin embargo, la interpretación a partir de estos dos factores es algo ambigua, por ejemplo, vemos que si bien las variables están "cargadas" en el factor$_1$ el segundo factor es bipolar. Además si bien la matriz principal provee información sobre factores de carga y comunalidades, no nos dan una información precisa sobre la estructura de las relaciones. De por sí la matriz original (no rotada) es arbitraria, en el sentido que se pueden trazar infinito número de posiciones. La rotación de los ejes, es entonces un medio para buscar la mejor manera de acomodar los datos en un espacio n-dimensional.

Una razón adicional para la rotación, como lo señala Kim (op. cit.) es que los factores de carga en la solución no rotada, dependen muy fuertemente en el número de variables. Los factores rotados son más estables. En fin el objetivo de la rotación es la obtención de factores teóricamente significativos.

Las líneas de puntos en el gráfico, señalan una rotación ortogonal para los datos que aparecen en la tabla anterior. Las líneas segmentales señalan una rotación oblicua.

Por lo general las tablas rotadas proporcionan patrones mucho más claros. Siguiendo el ejemplo hipotético presentado en la tabla, los factores rotados y no rotados serían ahora (rotación ortogonal):

	Factores no rotados		Factores rotados	
Variables	F_1	F_2	F_1	F_2
A	.83	.10	.60	.58
B	.75	.63	.14	.95
C	.70	.50	.22	.83
D	.85	-.42	.94	.09
E	.60	-.40	.72	.08
F	.69	.57	.14	.90

La estructura de los datos es ahora mucho más clara, aunque la variable A sigue teniendo carga en los dos factores. Por lo demás queda claro que en el factor 1 intervienen las variables D y E, mientras que en el factor 2 la carga se concentra en las variables D, C y F.

Las rotaciones, cualesquiera sea su tipo, tienen como objeto

hacer que los valores en la horizontal o en la vertical de los ejes, se aproximen lo máximo posible a 0. Al hacer esto con un factor, maximizamos a la vez el valor del otro factor.

Los métodos de rotación incluidos en el SPSS son:

I) rotación ortogonal quartimax;

II) rotación ortogonal varimax;

III) rotación ortogonal equimax;

IV) rotación oblicua

I) Rotación ortogonal quartimax:

Esta rotación sigue el principio de reducción del máximo de complejidad en una variable, mediante la rotación de los factores iniciales de tal manera que el factor de carga se concentre en un factor, haciendo que el peso de los otros factores se acerque lo máximo al valor 0. Este método enfatiza la simplificación de las líneas, por lo tanto, el primer factor rotado tiende a ser un factor general (muchas variables tienden a concentrar su peso en él). Los siguientes factores tienden a ser subclusters de variables.

II) Rotación ortogonal varimax:

El método varimax se concentra en la simplificación de las columnas de la matriz inicial. Es el método de uso más generalizado.

III) Rotación ortogonal equimax:

Es un método intermedio a los dos anteriores. En vez de concentrarse en la simplificación de las líneas (quartimax), o en la simplificación de las columnas (varimax), equimax trata de lograr algo de cada una de esas simplificaciones.

IV) Rotación oblicua

En las rotaciones oblicuas se acepta por principio que los factores están intercorrelacionados. A partir de esto, los ejes son rotados libremente, de manera tal que los hiperplanos se coloquen oblicuos los unos a los otros. Hay diversos métodos de rotación oblicua (a partir de los gráficos de las rotaciones ortogonales, el método del plano único de Thurstone, de rotación directa hacia estructuras primarias de Harris, etc.,) El método utilizado en el subprograma, es

un método de rotación oblicua objetivo, utiliza un método directo para la simplificación de los factores primarios de carga. Los factores son intercorrelacionados, si tales intercorrelaciones existen, sin embargo; algunos métodos tienden a hacer que los factores resultantes estén más correlacionados que otros métodos. En el método cional oblicuo, los valores δ son colocados en 0 (son los que por lo general tienden a reproducir las mejores soluciones oblicuas). Sin embargo, el usuario puede alterar los valores δ para lograr menor o mayor oblicuidad. Cuando varios valores δ son especificados, el programa calculará una rotación oblicua para cada valor especificado.

D) Interpretación

Es la tarea teórica de identificar el contenido y la naturaleza de los factores. Esto se hace mediante procesos inferenciales acerca de que tienen en común las variables con alta carga, con las variables de carga moderada, o con las variables con factores de carga próximos a cero. Esas inferencias son probadas a posteriores en otros diseños, con las hipótesis necesarias, etc.

Out-put del programa factor

Una tabla de análisis factorial completo provee la siguiente información:

1.- Una matriz de correlaciones de las variables en el modelo tal como aparece en la figura.

2.- Factores de carga iniciales: esta tabla contiene factores ortogonales proporcionando el patrón y la estructura de la matriz. Ya sea que los factores sean inferidos o definidos, se ordenan por importancia decreciente.

Cuando existe un solo factor se dice que el conjunto de variables es puro, o saturado, o cargado con el factor; cuando hay más de un factor, se dice que el conjunto es factorialmente complejo. Esta tabla nos informa tanto sobre el número de factores como de la magnitud de la carga o saturación de cada variable en cada uno de los factores iniciales. Los factores de carga varían de -1.0, pasando por 0, hasta +1.0 y se interpretan de la misma manera que un coeficiente de correlación (de hecho los factores de carga expresan la correlación entre las distintas variables y los factores). Las comunalidades son la suma de los cuadrados de los factores de carga en una variable y expresan el factor de varianza común: $h^2 = (F_1^2) + (F_2^2) + \ldots (F_n^2)$ Con base en la matriz inicial, el investigador puede decidir

sobre la cantidad final de factores a retener.

3.- Pesos para estimar variables a partir de factores (factor pattern matrix). Contiene los pesos de regresiones de los factores comunes y nos informa sobre la composición de una variable en términos de factores hipotéticos. Esta matriz es rotada y nos permite expresar la variable como una combinación de variables independientes, sean éstas definidas o inferidas.

4.- Pesos para estimar factores a partir de variables (factor-estimate o factor-score matrix) Provee medios para estimar puntajes en los factores a partir de las variables observadas.

5.- Correlación entre factores y variables (factor-structure matrix). Proporciona el coeficiente de correlación entre cada variable y cada factor. La solución es rotada, siendo esta tabla idéntica a la factor-pattern matrix.

6.- Matriz de correlación para los factores terminales. La interpretación de las tablas es diferente, según la solución haya sido ortogonal u oblicua. Las matrices básicas para los dos tipos de soluciones son:

	Solución ortogonal	Solución oblicua
Datos básicos	Matriz de correlación idéntica	
Factores iniciales	Idéntica matriz factorial ortogonal	
Factores terminales	Factor-matrix	a) Pattern-matrix b) Structure-matrix
Correlación entre factores		Factor-correlation matrix
Factor-estimate matrix	Factor-estimate matrix	Factor estimate matrix

I- Soluciones terminales para factores rotados ortogonalmente

Los coeficientes en la tabla representan tanto pesos de regresión (pattern-matrix), como coeficientes de correlación (structure matrix). Esto es porque la solución es ortogonal. El ejemplo de Kim nos ayudará a interpretar la tabla:

Matriz factorial final con rotación varimax

Variables	Factor 1	Factor 2	Factor 3
A	.88920	.07829	.03230
B	.78523	.14023	.05768
C	.10210	.67352	.06342
D	.07237	.85632	.09643
E	.08390	.09470	.76480
F	.12345	.00320	.69532
G	.32460	.34210	.04274

Examinando cada variable (es decir, cada línea), vemos que el determinante más importante en la variable A es el factor 1, lo mismo que en la variable B. Para las variables C y D el determinante más importante es el factor 2. Para las variables E y F el factor más importante es el 3. Todas estas variables tienen por lo consiguiente una complejidad de 1. La variable G, por el contrario, tiene como determinantes principales a los factores 1 y 2, por lo tanto su complejidad factorial es de 2, es decir que ésta no es una variable que se explique por una dimensión, sino que mide dos dimensiones; y así sucesivamente para complejidades factoriales 3, 4 ... n.

Leyendo ahora las columnas, podemos determinar cómo los factores hipotéticos (cada uno de los factores que aparecen en la tabla y que representan las variables independientes que explican cada una de las variables en la tabla), dan cuenta de cierta proporción de la varianza en la variable dependiente.

Por ejemplo, la varianza de la variable A explicada por el factor 1 es $(.8892)^2 = .79067$, es decir, 79% de la varianza de A es explicada por el factor 1.

Siguiendo el mismo razonamiento, la proporción de la varianza explicada por todos los factores, en el caso de la variable A, es:
$h_A^2 = (.88920)^2 + (.07829)^2 + (.03230)^2 = .79783$

Esto es lo que llamamos más arriba comunalidad, y es claro que en el caso de la variable A, la contribución de los factores 2 y 3 es mínima (exactamente .00716).

El complemento de la comunidad $(1-h_j^2)$ representa la proporción

de la varianza única, es decir, la proporción de la varianza no explicada por los factores comunes, o por ninguna variable en el conjunto.

Con los datos de la matriz, el investigador también puede calcular los coeficientes de correlaciones entre cualquier par de líneas, determinando así las fuentes de variación común a ambas variables. Por ejemplo, la correlación entre la variable A y la variable B, en los tres factores es:

$$R_{12} = r_{1F_1} r_{2F_1} + r_{1F_2} r_{2F_2} + r_{1F_3} r_{2F_3} = (.88920)(.78523) + (.07829)$$

$$(.14023) + (.03230)(.05768) = .71105$$

La correlación entre A y B es debida básicamente al factor 1. Mediante la fórmula se pueden calcular las intercorrelaciones para todos los pares de variables.

II. Soluciones terminales para factores rotados oblicuamente

En una solución oblicua habrá dos matrices separadas, una para la pattern-matrix y otra para la structure-matrix. La pattern-matrix delinea más claramente la agrupación de variables. El cuadrado de un pattern-coeficiente representa la contribución d recta de un factor determinado a la varianza de una variable. En la medida en que un factor puede contribuir a la varianza de una variable a través de otros factores correlacionados (es decir una contribución indirecta), el total de la varianza del que da cuenta un factor no es igual a suma de las contribuciones directas.

La matriz de estructuras (structure-matrix) consiste de coeficientes de correlación. La contribución entre la variable A y el factor 1, en el ejemplo que sigue es .97652 y su elevación al cuadrado dará cuenta de la cantidad de varianza en la variable A explicada por el factor 1.

Matrices para factores oblicuos

	Patterns-matrix		Structure-matrix	
Variables	Factor 1	Factor 2	Factor 1	Factor 2
A	.99978	-.11012	.97652	.08240
B	.88724	-.08012	.87234	.08902
C	.76098	.34380	.82600	.43000
D	-.08202	.99870	.11721	.99668
E	.05368	.96723	.24231	.97823

El subprograma factor provee una representación gráfica de los tres métodos de rotaciones ortogonales. Ya que la representación se hace en ejes cartesianos, los factores son tomados de dos en dos, es decir, si hay tres factores habrá tres gráficos (F_1 con F_2; con F_1 con F_3; y F_2 con F_3). Por este gráfico, como ya señalamos más arriba, el usuario tiene una idea más clara sobre el agrupamiento de variables, sus valores, etc. Pudiendo utilizarlos para decisiones sobre la rotación oblicua por ejemplo.

El programa factor también construye índices compuestos que representan las dimensiones teoréticas asociadas a los respectivos factores.

XII. ANALISIS DISCRIMINANTE (SUBPROGRAM DISCRIMINANT)

Es una técnica estadística para la clasificación, predicción y análisis en problemas de grupos o clases de objetos. Puede ser utilizado tanto para la determinación de las diferencias entre dos o más grupos, así como -a partir de esas diferencias- construir esquemas clasificatorios de manaera tal que sea posible clasificar cualquier caso cuya pertenencia a un grupo específico nos es desconocida.

Los supuestos que subyacen en el análisis discriminante son:

- los grupos son discretos e identificables;
- cada observación en los grupos puede ser descrita por un conjunto de mediciones de m características o variables;
- las m variables tienen una distribución normal multivariata en cada población.

La distinción entre los grupos se realiza a partir de un conjunto de variables discriminatorias, esto es, variables que el investigador sospecha miden características sobre las cuales los m grupos difieren. Por medio del análisis discriminante esas variables son combinadas linearmente de manera tal que se maximice la distinción entre los grupos.

Las combinaciones entre las funciones discriminatorias toman la forma de una ecuación donde las funciones discriminantes son:

$$D_i = d_{i1}Z_1 + d_{i2}Z_2 + \ldots\ldots\ldots + d_{im}Z_m$$

Donde:

D_i: puntaje de la función discriminante i

d_i: coeficientes de carga

Z_i: valor estándar de las m variables discriminantes utilizadas.

Una vez que se determinan las funciones discriminantes, éstas pueden ser utilizadas para propósitos de clasificación y de análisis.

Como técnica clasificatoria, el análisis discriminante puede ser utilizado para clasificar cualquier caso cuya pertenencia a un grupo específico nos es desconocida. Es decir, se construyen un conjunto de reglas para clasificar las observaciones en el grupo más apropiado. Por ejemplo, una vez determinadas las características que diferencian a conservadores de radicales, podemos utilizar las combinaciones lineares de las variables discriminatorias en la determinación de la probabilidad que un miembro cualquiera cuya afiliación desconocemos se adhiera a conservadores o radicales.

Como técnica analítica el análisis discriminante permite detectar en qué medida las variables discriminatorias efectivamente discriminan cuando se combinan en funciones discriminantes. También es posible reducir el número de funciones discriminantes, siguiendo el mismo tipo de razonamiento utilizado en el análisis factorial. Por medio de la técnica es posible el estudio de las relaciones espaciales entre los grupos, como así también identificar las variables que contribuyen de manera más significativa a la diferenciación entre los m grupos.

Los pasos en el análisis discriminante son los siguientes:

1.- El investigador selecciona un conjunto de variables que sospecha van a diferenciar entre los m grupos. Las variables pueden ser tantas como el investigador considere necesarias.

2.- Las variables seleccionadas en el primer paso pueden a su vez ser seleccionadas o no para su inclusión en el análisis discriminante. Si el investigador decide incluir todas sus variables, se deben dar instrucciones (rutina Method= direct) por medio de las cuales las funciones discriminantes se crearán independientemente del poder discriminatorio de cada una de las variables independientes.

Si por él contrario, el investigador decide seleccionar sus variables en base al poder discriminatorio, es decir si decide introducir un criterio estadístico adicional a los teóricos, existen 5 criterios de selección disponibles en el SPSS. Los procedimientos de selección operan de tal manera que seleccionan primero la variable que tiene el valor más alto en el criterio de selección; luego esa

variable es apareada con cada una de las otras variables hasta seleccionar una segunda variable, que combinada con la primera, mejora el criterio de selección; luego se aparean estas dos variables con cada una de las variables que quedan hasta seleccionar una tercera variable que combinada con las dos anteriores mejoran aún más el criterio; y así sucesivamente hasta que la inclusión de una variable adicional no provea un mejoramiento en la discriminación entre los grupos.

Estos métodos de ubicación de las variables por rango y de evaluación de su poder discriminatorio, no tienen la precisión de los análisis de regresión ya que no existe una prueba clara para determinar la significancia de un coeficiente en particular en una función discriminativa dada. De allí que los 5 métodos que se mencionan enfatizan distintos aspectos de la separación.

Método de Wilks (Method= Wilks) está basado en una prueba de significación de las diferencias entre grupos, mediante el test F de las diferencias entre grupos centroides. El método toma en consideración de la diferencia entre todos los centroides y la cohesión entre los grupos. Las variables pueden ser ordenadas según los valores de los coeficientes lambdas de Wilks de manera que se de un rangueamiento en referencia a su poder discriminatorio relativo. A más bajo el valor de lambda, más alto el poder discriminatorio de la variable. El problema de este método es que deja de considerar las correlaciones entre el conjunto de variables que se está utilizando, de manera tal que solamente en los casos en que las variables sean independientes (no correlacionadas) este método permitirá un ordenamiento y comparación válido.

El método Mahal (Method= Mahal) maximiza las distancias mahalonobis entre los dos grupos más próximos.

El método Miniresid (Method= Miniresid) separa los grupos de manera tal que la variación residual sea mínima. Tomando en cuenta la correlación múltiple entre el conjunto de variables discriminantes y una variable muda que identifica al par de grupos correspondiente, su objetivo es el de minimizar R, esto es, la variación residual.

El método Maxminf (Method= Maxminf) maximiza la distancia entre los grupos de manera tal que se seleccione la razón R más pequeña entre dos pares de grupos más próximos.

El método de Rao (Method= Rao) utiliza una medida de distancia V. La variable seleccionada es aquella que contribuye al aumento más grande en V cuando se agrega a las otras variables, de manera tal que se obtenga una separación máxima entre los grupos. Estas decisiones tienen que ver con la comparación del poder discriminatorio en diferentes conjuntos de variables (diferentes en relación a su tamaño). La solución de Rao tiene que ver con la significancia que pue

de tener la agregación o no de una variable en particular.

3.- Determinación del número de funciones discriminantes. Para cualquier cantidad de grupos o cualquier cantidad de variables discriminantes, el máximo número a derivar será igual a la cantidad de variables discriminantes o a la cantidad de grupos menos 1, cualesquiera sea menor. Es decir, que tres funciones pueden ser suficientes para describir a dos grupos. El subprograma Discriminant provee dos medidas para juzgar la importancia de las funciones discriminantes: porcentaje relativo del eigenvalue asociada a la función y la significación estadística de la información discriminatoria.

El eigenvalue es una medida especial que representa los valores característicos de una matriz cuadrada, y es una medida relativa de la importancia de la función, en la medida en que la suma de los eigenvalues es una medida de la variancia total que existe entre las funciones discriminantes; así un eigenvalue en particular es expresado como un porcentaje de esa suma. En la medida que las funciones discriminantes son derivadas por orden de importancia, el proceso de derivación es para cuando el porcentaje relativo (es decir el valor del eigenvalue) es demasiado pequeño.

El segundo criterio para juzgar la importancia de las funciones discriminantes es el test de la significación estadística de la información discriminante todavía no contemplada por las funciones ya determinadas. El método de cálculo utilizado es el lambda de Wilks.

4.- Interpretación de los coeficientes de función discriminante. Habíamos visto más arriba que la ecuación de las funciones discriminante era:

$$D_i = d_{i1}Z_1 + d_{i2}Z_2 + \ldots\ldots\ldots + d_{im}Z_m$$

los coeficientes de la función discriminante corresponden a los d_{ij} y son utilizados para computar el puntaje discriminante que es el resultado de aplicar la fórmula que aparece más arriba. Habrá por lo tanto un puntaje separado para cada caso en cada función. En la medida que los puntajes Z son estándar, su media es 0 y su valor estándar es 1. Por lo tanto, cualquier puntaje singular representa una desviación de la media de todos los casos sobre una función discriminante dada.

Computando el promedio de los puntajes para un grupo en particular, tenemos calculada la media del grupo en la función discriminante respectiva. Para cada grupo, las medias de todas las funciones se denominan grupo centroide, y señala la localización más típica de los casos en ese grupo con referencia al espacio de la función discriminante. Una comparación de las medias de los distintos grupos en ca-

da función nos indica entonces cómo se distribuyen los grupos a lo largo de una dimensión.

Los coeficientes estandarizados de las funciones discriminantes se pueden interpretar en forma similar a los coeficientes beta en las regresiones múltiples. El signo del coeficiente nos indica si la contribución de la variable es positiva o negativa. El valor del coeficiente indica la contribución relativa de la variable a la función. Los coeficientes estandarizados pueden ser utilizados -como en el análisis factorial- para identificar las características dominantes que ellos miden, nombrándolos así según características teóricas.

El programa contiene una opción para el cálculo de coeficientes no estandarizados, útiles para propósitos computacionales, pero que no nos informan sobre la importancia relativa de las variables.

5.- Distribución gráfica (plots) de los puntajes discriminantes. El programa imprime una representación espacial de la distribución de los puntajes discriminantes a lo largo del continuo de dos primeras funciones discriminantes. Es posible obtener ya sea la distribución de todos los casos en un gráfico, o gráficos separados para cada grupo. La representación espacial es particularmente útil para el estudio de los grupos centroides y su localización relativa, así como también para un análisis del grado en que los grupos se superimponen. Cuando solamente existe una función discriminante, la distribución toma la forma de un histograma.

6.- Rotación de los ejes de las funciones discriminantes. Tal como en el análisis factorial, es posible rotar la orientación espacial de los ejes manteniendo constante la localización relativa de los casos y de los centroides. Un criterio puede ser la solución varimax, es hacer rotar los ejes de manera que las variables discriminantes se aproximen a 1.0 o a 0.0. Mediante esto si bien es posible mejorar la interpretación de la distribución de las variables principales, hay pérdida de información referente a la importancia relativa de cada función.

7.- Clasificación de casos. Por este proceso es posible identificar la pertenencia de un caso a un grupo determinado, cuando solamente conocemos los valores del caso en las variables discriminatorias. La clasificación se logra mediante el uso de una serie de funciones de clasificación, una para cada grupo. Se computan para cada caso tantos puntajes como grupos existen y el caso es clasificado en el grupo con el puntaje más alto. Los puntajes pueden ser asimismo en probabilidades de pertenencia a grupo asignándose el caso al grupo cuya probabilidad de pertenencia es más alto.

El sistema de clasificaciones de probabilidad es útil no sola-

mente para adjudicar casos en un grupo, sino además para controlar cuán efectivas son las variables discriminantes. De allí que sus valores se utilicen aun para los casos de selección de variables y de funciones. Si existe un número de casos cuya afiliación conocemos, pero que están mal clasificados, entonces las variables seleccionadas son muy pobres en el proceso de discriminación.

Algunos ejemplos de análisis discriminante.

a) Ejemplo en dos grupos

El ejemplo es de una aplicación de Heyck y Klecka, y aparece en el Manual del SPSS. Se trata de una análisis del Parlamento Británico durante el periodo 1874-1895, en el que el Partido Liberal estaba fraccionado entre radicales y no-radicales. Algunos miembros del Parlamento fueron clasificados en uno u otro grupo según documentos históricos. Sin embargo quedaron sin clasificar un conjunto de miembros para los cuales no se disponía de suficiente información o para los cuales la información era contradictoria.

Las variables discriminantes seleccionadas fueron votos en el Parlamento, los votos en asuntos particularmente relacionados al programa radical. Se seleccionaron 17 de estas variables, siendo los asuntos a votar los siguientes:

Fecha	
25/marzo /74	1.- horas de votación
17/abril /74	2.- gastos de la Corona
10/junio /74	3.- educación no sectaria
17/junio /74	4.- temperance reform
9/junio /75	5.- escolarización compulsoria
17/junio /75	6.- parlamentos trienales
14/junio /75	7.- asignaciones de tierras
15/julio /75	8.- gastos de la Corona
15/julio /75	9.- extensión de la
5/abril /76	10.- educación gratuita
30/mayo /76	11.- extensión de la
10/julio /76	12.- control del estado en educación
13/marzo /77	13.- temperance reform
23/marzo /77	14.- reformas en Turquía

13/mayo /80	15.- prerrogativas de la Corona en política externa
24/febrero/80	16.- parlamentos quinquenales
5/marzo /80	17.- temperance reform

los votos desde el punto de vista radical fueron clasificados con +1, los votos en contra con -1 y las abstenciones con 0.

Antes de calcular el análisis discriminante las variables fueron seleccionadas para su inclusión mediante el método V de Rao, así solamente se seleccionaron 11 de las 17 ocasiones de voto. Esos 11 votos proveían un alto grado de separación entre los grupos (el lambda de Wilks dio como valor .19264 con una correlación canónica de .899 para la función discriminante). Las variables eliminadas fueron la 6, 7, 8, 9, 10 y 17.

Los valores obtenidos después del análisis discriminante para las 11 variables discriminatorias fueron los siguientes:

Paso Número	Variable	F para incluir o extraer del análisis	Número incluido	Lambda de Wilks	Rao V	Cambio en el V de Rao
1	Var013	39.95802	1	0.66125	39.97799	39.95799
2	Var011	33.86707	2	0.45926	91.83968	51.88168
3	Var014	14.96441	3	0.38370	125.28114	33.44147
4	Var005	10.68981	4	0.33584	154.25507	28.97392
5	Var015	8.99831	5	0.29943	182.49681	28.24174
6	Var003	8.65288	6	0.26770	213.37398	30.87717
7	Var002	6.10148	7	0.24678	238.06580	24.69182
8	Var001	5.55072	8	0.22889	262.77515	24.70935
9	Var016	6.10082	9	0.21054	292.47461	29.69946
10	Var012	3.38349	10	0.20070	310.63940	18.16479
11	Var004	2.84360	11	0.19264	326.89038	16.25098

Clasificación de coeficientes de función

	Grupo Radical	Grupo no radical	Coeficientes Discriminantes
Var001	4.78972	1.72740	-0.340
Var002	3.77127	0.07671	-0.410
Var003	3.41047	1.11915	-0.254
Var004	0.40823	-1.01541	-0.158
Var005	5.96921	2,52984	-0.382
Var011	3.94122	1.46608	-0.275

	Grupo radical	Grupo no radical	Coeficientes discriminantes
Var012	3.61321	0.97632	-0.293
Var013	3.12955	0.53350	-0.288
Var014	1.31335	-0.79479	-0.234
Var015	1.21796	-0.86230	-0.231
Var016	2.82108	-0.42259	-0.360
Constant	-12.65345	-1.92266	1.167

Funciones discriminantes	Eigen value	Porcentaje relativo	Correlación canónica	Funciones derivadas	Lambda de Wilks	Ji cuadrado	DF
1	4.19092	100.00	0.899	0	0.1926	119.401	11

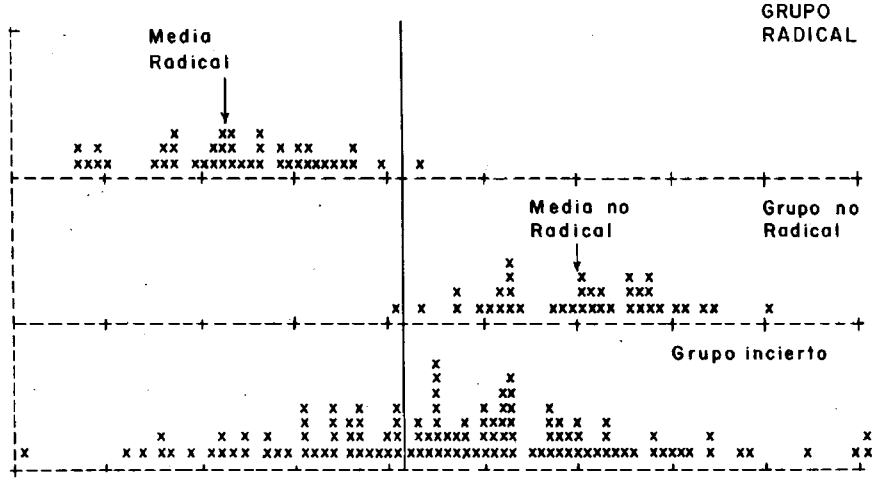
Coeficientes estandarizados de función discriminante

Var001	-0.17241
Var002	-0.20581
Var003	-0.15982
Var004	-0.10921
Var005	-0.16783
Var011	-0.13131
Var012	-0.13519
Var013	-0.20311
Var014	-0.10893
Var015	-0.16580
Var016	-0.17411

Centroides de grupo en espacio reducido

Grupo 1	-0.83134
Grupo 2	0.91885

Distribución gráfica (plots) de los puntajes discriminantes para miembros del Parlamento Inglés; 1874-1895.



Los coeficientes discriminantes no estandarizados son utilizados para calcular el puntaje discriminante para la función. Esto se obtiene multiplicando cada coeficiente por el valor en la variable (voto) y luego sumando los productos más la constante. Supónganse los votos de dos miembros del parlamento, Sir Wilfrid Lawson y Lord Hartington.

En la votación de los diferentes asuntos los miembros lo hicieron de la siguiente forma (+1 indica voto radical, -1 en contra y 0 abstención)

	Lawson	Hartington
Var. 1	+1	0
Var. 2	0	0
Var. 3	+1	+1
Var. 4	+1	0
Var. 5	+1	0
Var. 11	+1	+1
Var. 12	+1	0
Var. 13	+1	-1
Var. 14	+1	0
Var. 15	+1	-1
Var. 16	+1	0

Los puntajes para los dos miembros serán entonces:

Lawson

$$D = (-.340)(+1) + (-410)(0) + (-254)(+1) + (-.158)(+1) + (-.382)(+1) + (-.275)(+1) + (-.293)(+1) + (-.288)(+1) + (-.234)(+1) + (-231)(+1) + (-.360)(+1) + (1.167) = -1.648$$

Hartington

$$D = (-.340)(0) + (-410)(0) + (-254)(+1) + (-.158)(0) + (-.382)(0) + (-.275)(+1) + (-.293)(0) + (-.288)(-1) + (-.234)(0) + (-.231)(-1) + (-.360)(-1) = 1.157$$

Los puntajes calculados para todos los miembros aparecen en la distribución gráfica. Donde 0 es la media total de todos los radicales y no-radicales conocidos.

El siguiente paso es la clasificación de los miembros no identificados claramente como radicales o no-radicales. Esto se hace calculando para cada uno de ellos el puntaje discriminante, usando los coeficientes derivados de los radicales y no-radicales conocidos. La lógica de la clasificación se basa en la comparación de las pautas de votos de radicales conocidos con las pautas de votos del miembro incierto, de manera tal de clasificarlo en el grupo más similar en cuanto a pauta. Para ello se utilizan una serie de ecuaciones (una para cada grupo), cada ecuación dará lugar a una probabilidad, y el miembro será colocado en el grupo para el cual obtiene una probabilidad más grande. Este sistema tiene algunos problemas, sobre todo cuando las probabilidades de pertenencia a uno u otro grupo son del tipo .53, .47. Sucede en el caso del ejemplo que consideramos que algunos miembros no muestran una pauta consistente de apoyo a uno u otro grupo.

Ejemplo con varios grupos

El ejemplo es de Gansner, Seegrist y Walton*, en el que se trata de

---

(*) D. Gansner, D. Seegrist, G. Walton: A technique for defining subareas for regional analysis; en Growth and Change, October, 1971.

seleccionar subregiones en el Estado de Pennsylvania (USA) mediante una combinación de análisis discriminante y técnicas de agrupamiento (Clustering techniques). La regionalización está relacionada a un análisis de la eficiencia económica de la industria de la madera, con el objetivo de desarrollar un sistema más eficiente de tala y entrega de derivados de la madera. Se busca que las subregiones sean homogéneas y compactas en términos de producción de madera y de mercado potencial. Se busca además que cada subregión tenga fronteras equidistantes de sus centros de consumo.

Las variables discriminantes seleccionadas, fueron las siguientes:

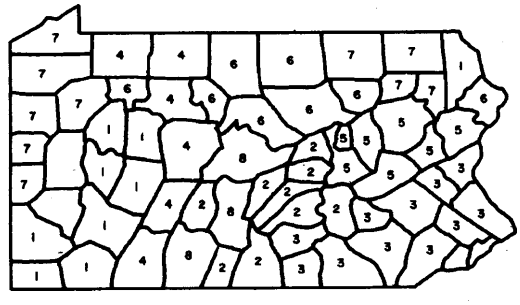
Variables que reflejan existencia de madera y accesibilidad.	Variables que reflejan actividad de mercado y demanda.
Var.1: Bosques comercializables como % del total de tierras.	Var.9: Promedio anual de producción de pulpa de madera (round pulpwood).
Var.2: Volumen de stock creciente por acre de bosque comercializable.	Var10: Madera blanda (softwood) como % de la producción de pulpa de madera.
Var.3: Volumen de stock creciente en bosques comercializables	Var11: Número de fábricas que usan pulpa de madera dentro de un radio de 100 millas de radio del centro del municipio.
Var.4: Madera blanda (softwood) como % del total del volumen de stock creciente.	Var12: Capacidad total productiva de las fábricas que usan pulpa de madera.
Var.5: % de bosques en propiedad pública.	Var13: Número de campos madereros y firmas contractantes.
Var.6: Diferencia de elevación máxima.	Var14: Empleo en campos madereros y firmas contractantes.
Var.7: % de superficie con pendiente suave.	Var15: Número total de industrias madereras y de productos derivados de la madera.
Var.8: Millas de carreteras por milla cuadrada de superficie total.	

Los grupos son todos los municipios de Pennsylvania, excepto Delaware y Philadelphia, esto es 65 municipios.

El segundo paso utilizado por los autores, fue el de agrupamiento de municipios. La técnica utilizada fue la del análisis discriminante stepwise que corresponde a nuestro segundo paso (selección de variables para su inclusión en el análisis). Resultaron seleccionados 8 grupos que forman subregiones compactas, tal como aparecen en la figura que sigue:

FIGURA 1

8 grupos de municipios en Pennsylvania formados por análisis discriminantes



A continuación se calcularon las medidas de distancia generalizada (D^2) o medida de similaridad entre pares de grupos, basada en valores F, cuyos resultados fueron los siguientes:

TABLA 2 VALORES D^2

Grupo	1	2	3	4	5	6	7
2	1,279.6	...	...	...	...	...	...
3	1,884.3	287.4	...	...	...	...	...
4	506.1	494.2	1,219.3	...	...	...	...
5	735.3	703.1	2,334.3	1,497.7	...	...	...
6	600.9	161.2	1,335.8	251.0	664.5	...	...
7	158.9	783.9	1,112.2	492.8	396.8	394.8	...
8	2,478.4	210.4	1,204.6	398.3	1,612.5	627.8	1,698.5

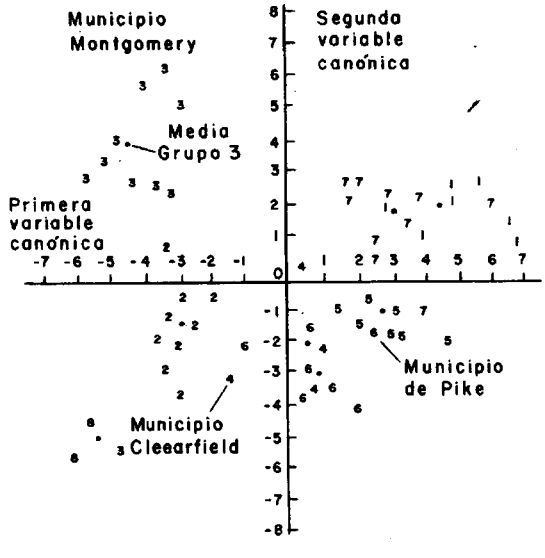
Como puede verse en la tabla, los grupos 1 y 7 son muy similares (tienen los valores D^2 más bajos). Los grupos 2 y 3; 2 y 6; 2 y 8; 4 y 6 también son similares. Por el contrario, los grupos 1 y 3; 1 y 8; 3 y 5; 5 y 8; 7 y 8 son muy diferentes.

El programa suministra una distribución gráfica de los valores de las dos primeras variables canónicas que muestra gráficamente la distribución en los grupos. Las medias de los grupos aparecen identificadas con asteriscos.

Los autores no proporcionan todos los valores, pero en el gráfico que figura a continuación puede verse que la media del grupo 3 está localizada a -4.211 en la primera variable canónica y a + 4.016 en la segunda. La municipalidad de Montgomery que pertenece al grupo 3 tiene como coordenadas (-3.237 y + 5.959). Las variables canónicas están relacionadas a los valores D^2 y reflejando la similaridad entre los grupos. Puede verse en el gráfico que los grupos 1 y 7 están próximos, mientras que 1 y 8 están alejados.

FIGURA 2

Municipios agrupados en 8 grupos.



El tercer círculo realizado, se refiere a una matriz de clasificación para evaluar las probabilidades de clasificar los municipios en grupos. La tabla siguiente, muestra una matriz casi perfecta, con la excepción de tres elementos, lo que indica que la clasificación es casi perfecta.

TABLA 3

Matriz para la evaluación de asignación de municipios en 8 grupos

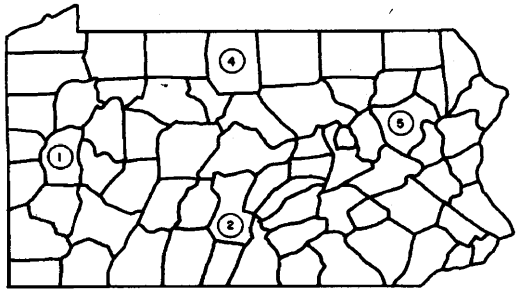
Subregión	Número de municipios por subregiones							
	1	2	3	4	5	6	7	8
1	9	0	0	0	0	0	1	0
2	0	9	0	0	0	0	0	0
3	0	0	11	0	0	0	0	0
4	0	0	0	6	0	0	0	0
5	0	0	0	0	7	0	0	0
6	0	0	0	0	0	7	1	0
7	1	0	0	0	0	0	10	0
8	0	0	0	0	0	0	0	3

En base a la similaridad de valores D^2 se forman ahora 5 subregiones:

La combinación de los grupos 1 y el noroeste de la región 7 produce una unidad. Los grupos 2 y 8 otra unidad compacta. El grupo 3 queda como está, se combina el grupo 5 con el nordeste del grupo 7, más 1 municipio del grupo 6. Parte del grupo 4 se combina con parte del grupo 6. Dos municipios del grupo 4 se agregan a la combinación del grupo 2-8. Quedan así 5 subregiones como aparecen en el siguiente mapa.

FIGURA 3

5 subregiones de municipios en Pennsylvania formados por análisis discriminante



Esta reclasificación es sometida nuevamente a un análisis discriminante para controlar la reclasificación, produciendo los resultados, la siguiente matriz que muestra una clasificación perfecta.

TABLA 4

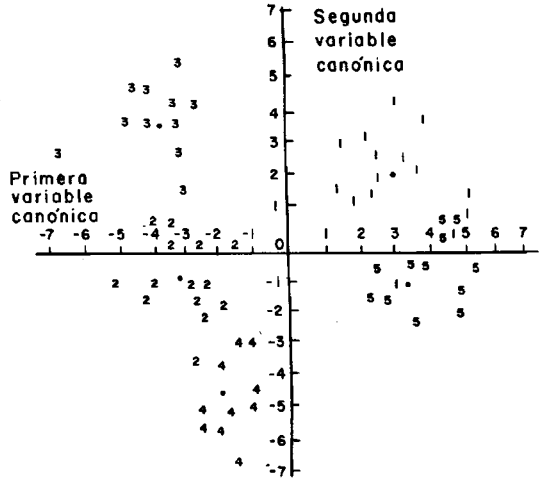
Matriz para la evaluación de asignación de municipios a 5 subregiones

	Número de municipios por subregiones				
Subregión	1	2	3	4	5
1	16	0	0	0	0
2	0	14	0	0	0
3	0	0	11	0	0
4	0	0	0	11	0
5	0	0	0	0	13

Al mismo tiempo la distribución gráfica de las nuevas variables canónicas muestra una completa separación entre los grupos, excepto para los grupos 1 y 5. Las siguientes variables canónicas separan a estos grupos.

FIGURA 4

Municipios agrupados en 5 subregiones



Aunque resultante de una combinación de análisis discriminantes con técnicas de cluster analysis, se ve claramente cómo la técnica puede ser útil para aplicar criterios objetivos en la determinación de subregiones.

La técnica del análisis discriminante encuentra numerosas aplicaciones en el campo de la medicina (para la determinación de síndromes y eventualmente para diagnósticos más objetivos), en psicología, economía, etc.

XIII. ANALISIS DE ESCALOGRAMA GUTTMAN (SUBPROGRAMA GUTTMAN SCALE)

La escala Guttman, conocida como "método del escalograma" o "análisis de escalograma", tiene como objetivo el definir lo más claramente posible qué es lo que está midiendo la escala, entendido esto como un problema de unidimensionalidad. Por el tipo especial de tratamiento al que se somete la escala, se busca la eliminación de factores extraños a las características o dimensión que se pretende medir.

La segunda característica de una escala Guttman, es su propiedad de escala acumulativa, es decir, que la respuesta positiva a un ítem supone que los ítems anteriores han sido respondidos en forma afirmativa. Se busca pues una coherencia en las pautas de respuesta de los sujetos, y esa coherencia es garantizada por medio de un coeficiente de reproducibilidad. El tamaño del coeficiente (su valor máximo es 1.00) indica el grado en que la escala es acumulativa. En una escala cuya reproducibilidad es perfecta (1.00), las respuestas de los sujetos a todos los ítems pueden ser reproducidas con el solo conocimiento de la posición de rango. El escalograma Guttman combina aspectos de construcción utilizados por las escalas Lickert y Thurstone*. A partir de una serie de items que son administrados a una muestra de sujetos que van a actuar como jueces, se procede a un análisis de los items en su conjunto, buscando la producción de una serie acumulativa de ítems. Los jueces se ordenan en término de los puntajes obtenidos en la escala, así como los ítems por grado de di-

---

(*) Para mayores detalles, puede verse: Jorge Padua: Escalas para la medición de actitudes, CES, El Colegio de México, México, 1975.

ficultad. Cuando la serie acumulativa de ítems es perfecta, todas las celdas cruzadas en el escalograma estarán en una posición sobre la diagonal que corre desde el ángulo superior izquierdo hasta el ángulo inferior derecho de la matriz. A más alto el número de desviaciones de esta diagonal, mayor el número de errores, es decir menor la reproducibilidad.

La idea pues, es producir cortes y seleccionar ítems de manera tal que la reproducibilidad se maximice.

La tabla que sigue muestra un análisis de escalograma con las respuestas de 20 jueces a 6 ítems, en términos de acuerdo-desacuerdo. Los acuerdos aparecen señalados con una x y los desacuerdos con un 0. Los valores encerrados en círculos son errores.

Análisis de escalograma.

Respuestas de 20 jueces a 6 items en términos de acuerdo-desacuerdo

Rango del Juez	4	2	1	6	5	3	Puntaje
1	x	x	x	x	x	x	6
2	x	x	x	x	x	x	6
3	x	x	x	x	x	x	6
4	0	x	x	x	x	x	5
5	0	x	(0)	x	x	x	4
6	0	0	x	x	x	x	4
7	0	0	x	x	(0)	x	3
8	0	0	x	x	x	(0)	3
9	0	0	0	x	x	x	3
10	0	0	0	x	x	x	3
11	0	0	0	x	x	x	3
12	0	0	(x)	0	x	x	3
13	0	(x)	0	0	x	x	3
14	0	0	0	0	x	x	2
15	0	0	0	0	0	x	1
16	0	0	0	0	0	x	1
17	0	0	0	0	(x)	(0)	1
18	0	0	0	0	0	x	1
19	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0
	3	6	8	11	14	16	

La técnica usada para la determinación de los cutting points, esto es los puntos que determinan el ordenamiento de los ítems son básicamente dos (técnica de Cornell y técnica de Goodenough). La técnica de Cornell se realiza estableciendo puntos de separación en el orden de rango de los jueces, tales como se definirían éstos si la escala fuese perfecta. La técnica de Goodenough se basa en el cálculo de errores en base a pautas marginales.

Las posibilidades de establecimiento de cutting-points en escalas más complejas, esto es con ítems con mayores categorías de respuesta que los presentados en el ejemplo anterior, obligan la mayoría de las veces a la reclasificación de categorías, simplificándolas. La tabla que sigue muestra un ejemplo algo más complejo.

Como puede observarse, existen bastantes sujetos "fuera de posición" (en el ítem 1, hay varios sujetos que respondieron 4, y están por debajo de sujetos que respondieron 3, lo mismo hay sujetos que respondieron 1 en el ítem y tienen puntajes totales por encima de sujetos que dieron respuesta 2, 3 y 4. Los únicos sujetos que no tienen errores en este ítem son los que dieron respuesta 2. Se trata entonces de recombinar las categorías de respuesta de manera de minimizar los errores. Supongamos que las reclasificaciones propuestas para cada ítem, con sus respectivos pesos son ahora las siguientes:

Item	Combinaciones	Nuevos pesos
1	4 3 2,1,0	2 1 0
2	4,3 2,1,0	2 0
3	4,3 2,1,0	2 0
4	4,3,2 1,0	2 0
5	4,3 2,1,0	2 0
6	4,3,2 1,0	2 0

Naturalmente, los puntajes totales de algunos sujetos variarán, alternando consecuentemente su orden de rango. El lector puede reconstituir la tabla.

Análisis de escalograma

Ordenación de puntajes y valores de respuesta en 6 ítems tipo escala Lickert

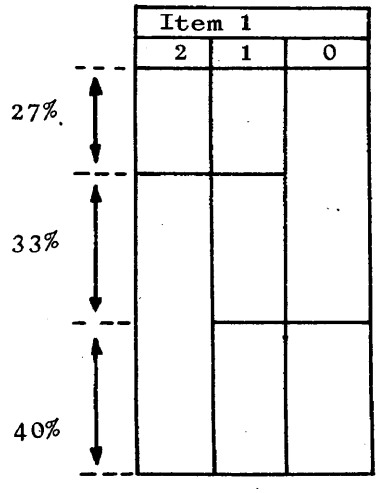
Puntaje	Item 1					Item 2					Item 3					Item 4					Item 5					Item 6				
	4	3	2	1	0	4	3	2	1	0	4	3	2	1	0	4	3	2	1	0	4	3	2	1	0	4	3	2	1	0
24	X					X					X					X					X					X				
23	X					X						X				X						X					X			
21	X					X					X							X				X				X				
20		X					X				X						X					X				X				
19	X						X				X					X							X					X		
19	X						X				X								X			X				X				
19		X				X						X						X			X						X			
18	X						X				X								X			X					X			
18		X					X					X					X					X					X			
18		X					X					X					X					X					X			
18	X						X					X					X						X				X			
17		X				X							X					X				X					X			
16		X					X				X								X					X		X				
15		X								X	X							X				X					X			
15				X		X						X								X		X				X				
13		X							X			X								X				X		X				
12		X						X						X						X			X				X			
12		X							X						X				X			X					X			
11	X								X						X					X	X							X		
10			X					X					X							X					X			X		
8					X					X			X						X					X	X					
8				X						X				X					X				X				X			
8					X				X					X						X		X					X			
8			X							X					X			X					X			X				
7			X					X						X						X			X					X		
6					X			X				X								X			X						X	
4			X					X							X					X				X					X	
4																														
3																														
2																														

Los cutting-point en la técnica de Goodenough se determinan según la distribución de las respuestas de los jueces a cada una de las alternativas en los distintos ítems. Los cálculos para la tabla III, nos darían los siguientes valores:

Items

	1			2		3		4		5		6	
categorías	2	1	0	2	0	2	0	2	0	2	0	2	0
frecuencias	8	10	12	14	16	15	15	14	16	16	14	24	6
porcentajes	27	33	40	47	53	50	50	47	53	47	53	80	20

Los cutting-points para cada uno de los ítems deben seguir entonces para cada ítem, los porcentajes respectivos. En el caso del ítem 1, los cortes serían entonces:



Es decir, que en ítem 1, el primer corte o cutting-point caería entre el último sujeto con puntaje total 10 y el primero con puntaje total 9. Y así sucesivamente para el resto de los ítems.

Mediante estos cutting-points, se van a determinar pautas de respuestas correspondientes a cada corte. La pauta de respuesta es la manera "correcta" en que deberían distribuirse para cada puntaje total de cada juez si la escala tuviera perfecta escalabilidad. Cada respuesta que no sigue la pauta de respuesta ideal se considera un "error". En nuestro ejemplo, un juez con puntaje de 8 debe seguir una pauta: 0-0-2-2-2-2. Un sujeto con puntaje 6, debe consecuentemente tener una pauta de respuesta 0-0-0-2-2-2. Es decir, que primero hay que ordenar a los ítems en términos de escalabilidad, para luego determinar las pautas correspondientes a cada valor. El principio de la técnica es bastante simple. Para 4 ítems dicotomizados y con valores de 1 y 0 para alternativas de respuesta "de acuerdo" y "en desacuerdo" respectivamente, un puntaje total de 3 puede seguir cuatro tipos de pautas diferentes:

a) 0 1 1 1

b) 1 0 1 1

c) 1 1 0 1

d) 1 1 1 0

De estas pautas, solamente la primera es correcta, para ítems

que están ordenados en forma acumulativa. Cada una de las pautas b), c) y d) tiene respectivamente 2 errores (uno por tener un 1 donde debería haber un 0, y otro por tener un 0 donde debería tener un 1). Para hacer más claro el método, vamos a dar un ejemplo más simple, en el que presentamos 4 ítems dicotomizados, las pautas de respuesta y el cálculo de los errores.

Items								Puntaje total	Pauta de respuesta				Errores
1		2		3		4							
1	0	1	0	1	0	1	0						
x		x		x		x		4 →	1	1	1	1	0
	x	x		x		x		3					0
x			x	x		x		3					2
x		x		x			x	3	0	1	1	1	2
	x	x		x		x		3					0
	x	x		x		x		3					0
	x	x		x			x	2					2
	x		x	x		x		2	0	0	1	1	0
	x		x	x		x		2					0
	x		x		x	x		1					0
	x		x		x	x		1	0	0	0	1	0
	x		x		x	x		1					0
	x		x		x		x	0					0
	x		x		x		x	0	0	0	0	0	0
	x		x		x		x	0					0
								Total errores					6

El coeficiente de reproductibilidad, se calcula según la fórmula ya conocida de:

$$r_p = 1 - \frac{\text{Cantidad de errores}}{\text{número total de respuestas}}$$

El ejemplo arriba citado: $r_p = 1 - \frac{6}{60} = .90$

El coeficiente de reproductibilidad nos indica la proporción de respuestas a los ítems que pueden ser correctamente reproducidas.

Presentamos ahora dos ejemplos, a partir de los cuales, presentaremos criterios adicionales al coeficiente de reproductibilidad para la determinación del universo de contenido en la escala Guttman.

Tabla IV: Escalograma Guttman para la medición de nivel de vida determinada en función de materiales empleados en la construcción de viviendas y de las instalaciones sanitarias. Muestra de población de Colombres en el Departamento de Santa Cruz en Tucumán-Argentina. Respuestas de 33 jefes de familia.

Ejemplo A Vivienda							Ejemplo B Instalaciones Sanitarias				
Sujetos	Items						Sujetos	Items			
	1	2	3	4	5	6		1	2	3	4
1	x	x	x	x	x	x	3	x	x	x	x
33	x	x	x	x	x	x	15	x	x	x	x
31	x	x	x	x	x	x	17	x	x	x	x
22	x	x	x	x	x	x	7	x	x	x	x
19	x	x	x	x	x	x	11	x	x	x	x
16	x	x	x	x	x	x	30	x	x	x	x
4	x	x	x	x	x	x	2	x	x	x	x
6	x	x	x		x	x	1	x	x	x	x
3		x	x	x	x	x	4	x	x	x	x
15		x	x	x	x	x	5	x	x	x	x
17		x	x	x	x	x	8		x	x	x
14		x	x	x	x	x	6		x	x	x
7		x	x	x	x	x	33		x	x	x
2		x	x	x	x	x	31		x	x	x
18		x	x	x	x	x	22		x	x	x
5		x	x	x		x	19		x	x	x
8		x	x		x	x	16		x	x	x
11		x	x	x		x	14		x	x	x
30		x	x	x		x	9				x
29			x		x	x	18				x
9				x	x	x	13				
13				x	x		25				
10				x	x		24				
26					x	x	29				
28					x	x	27				
24					x		26				

Sujetos	Items						Sujetos	Items			
	1	2	3	4	5	6		1	2	3	4
27						x	32				
20						x	12				
21						x	20				
25							28				
12							23				
23							10				
32							21				

Cálculo universo de contenido para la tabla IV:

	Ejemplo A	Ejemplo B
Coeficiente de reproductibilidad:	.949	1.00
Rango marginal mínimo	.72	.60
Alcance de distribución marginal:	.229	.40

Referencias:

Ejemplo A

Item 1: Alejamiento de aguas servidas a pozo ciego y cámara séptica.

Item 2: Agua corriente

Item 3: Eliminador de residuos

Item 4: Techo de zinc, o loza o teja con aislación

Item 5: Piso de cemento o mejor

Item 6: Luz eléctrica

Item 1: Pileta de lavado

Item 2: Inodoro

Item 3: Ducha

Item 4: Revestimiento de cemento o superior en baño o letrina.

La x indica presencia del ítem en ambos ejemplos.

El coeficiente de reproductibilidad en el ejemplo B, cuyo valor es 1.00, no permite decir, sin tener en cuenta la distribución gráfica que el sujeto 18 tiene en su casa únicamente revestimiento de cemento o mejor en su baño; mientras que el sujeto 33, cuyo puntaje es 3, tiene revestimiento, inodoro y ducha, pero no pileta de lavado.

Es muy difícil lograr escalabilidad perfecta, y consecuentemente, existen errores que van a ser interpretados como errores de reproductibilidad. Guttman aconseja que los coeficientes de reproductibilidad no sean menores de .90

El coeficiente de reproductibilidad (r_p) es un criterio necesario, pero no suficiente para la determinación de la escalabilidad de los ítems. Deben tomarse en cuenta otros factores. Stouffer et.al.* señalan cuatro criterios adicionales:

a) alcance de la distribución marginal;

b) pauta de errores;

c) número de ítems en la escala; y

d) número de categorías de respuestas.

a) Alcance de la distribución marginal

Es el más importante de los criterios adicionales, y debe acompañar al coeficiente de reproductibilidad. El criterio de distribución marginal es determinado por el rango marginal mínimo (M.M.R.) que consiste en el r_p menos el promedio de los modos de las frecuencias relativas de las distribuciones de los items (r_p - MMR).

Para algunos, los valores de este criterio adicional deben variar entre .15 y .35; para otros el mínimo debe ser mayor que .10. Estos valores indican la escalabilidad de los ítems, dato que no es proporcionado por el r_p de manera completa (es decir, es posible alcanzar valores altos de r_p -digamos .90- y resultar una escalabilidad inaceptable. Este es el caso en el cual los cuttings points están muy próximos entre sí, con el resultado de discriminar solamente en los extremos de la escala y no a lo largo de la misma. En nuestro ejemplo, los valores de r_p son altos y muy aceptables; los alcances de la distribución marginal, en cambio, son aceptables para el ejemplo A, y demasiado alto para el ejemplo B.

(*) Stouffer, S. et.al.: Measurement and prediction. Studies in social psychology in World War II, Vol. IV Princeton University.

b) Pauta de errores

Cuando el r_p es menor que .90, pero es escalable, es decir que tiene un r_p M.M.R. mayor que .10 estamos en presencia de más de una variable, mejor dicho, de una variable dominante y de otra u otras menores; en el área a través de la cual se ordenan los sujetos, este tipo de escalograma es denominado cuasi-escala. Este no es el caso de los dos ejemplos que presentamos.

c) Número de ítems en la escala

A mayor el número de ítems, mayor la seguridad de que el universo, del cual estos ítems son una muestra, es escalable.

Es por esto que cuando los ítems están dicotomizados, como es el caso en nuestros ejemplos, su número es aconsejable que sea mayor que 10. Pero puede usarse un número menor de ítems si las frecuencias marginales se colocan en un rango con recorridos del 30% al 70%.

En los ejemplos dados por nosotros, el rango de frecuencias es:

Ejemplo A	Ejemplo B
Item 1 ---- 24%	Item 1 ---- 30%
Item 2 ---- 57%	Item 2 ---- 35%
Item 3 ---- 60%	Item 3 ---- 55%
Item 4 ---- 69%	Item 4 ---- 60%
Item 5 ---- 78%	
Item 6 ---- 87%	

Tenemos alguna seguridad, de acuerdo al requisito citado más arriba, que el universo se comporta como la muestra.

d) Número de categorías de respuestas

Es otro criterio para asegurar la escalabilidad; cuanto mayor el número de categorías, mayor la seguridad de que el universo es escalable. Por ella, a pesar de la necesidad de reducir las categorías por razones prácticas (disminución del número de errores), hay que asegurarse de que tal reducción no es la resultante de obtener frecuencias marginales extremas (.90-.10) que, como vimos más arriba, no permiten errores, pero artificialmente.

Si mantenemos el número de alternativas de respuestas, a pesar de que aumentará el número de errores, disminuimos la posibilidad de que aparezca una pauta escalable cuando de hecho el universo no lo es.

El out put del subprograma Guttman scale, tiene la siguiente forma (suministramos los datos correspondientes al texto en el SPSS op.cit.):

Se trata de 3 ítems:

1.- Miembro de organizaciones orientadas hacia el servicio.

2.- Miembros de organizaciones ocupacionales.

3.- Miembro de grupos recreacionales.

	Recre. (3)		Org. Oc. (2)		M. O. S. (1)		
	0	1	0	1	0	1	
	Err		Err		Err		
3	0	13	0	13	0	13	13
2	32	17	10	39	7	42	49
1	108	12	66	54	66	54	120
0	168	0	168	0	168	0	168
Sumas	380	42	244	106	241	109	350
%	88	12	70	30	69	31	
Errores	0	29	10	54	73	0	166

Coeficiente de reproductibilidad : .8419

Reproductibilidad marginal mínima: .7552

Porcentaje de mejoramiento : .0867

Coeficiente de escalabilidad : .3541

Coeficientes de correlación

	Var.1	Var.2	Var.3
Var. 1	1.000	.3496	.5151
Var. 2	.3496	1.000	.4024
Var. 3	.5151	.4024	1.0000
Escala Item	.2953	.2565	.3490

Los cutting points se realizaron con una técnica similar a la Goodenough algo más simple que la especificada más arriba.

El cuadro se puede leer de la misma forma que la primera tabla que presentamos más arriba. La diferencia es que aquí se representan en una doble columna los valores afirmativos (la x se reemplaza con un 1); y los negativos (se representan con un 0). En la primera columna de la tabla figuran los valores 3, 2, 1, 0 que representan los posibles puntajes y que nos sirven para la determinación de los cutting point y por lo tanto para contar los errores. Un puntaje tres representa una pauta de respuesta 1 1 1; un puntaje dos representa una pauta 0 1 1 (siendo error por ejemplo 101 y 110); una pauta de respuesta de 1 es del tipo 0 0 1 (y no 100, 010). Examinando ahora el ítem que quedó ordenado como en primer lugar y que es el que más discrimina (solamente el 12% de los sujetos dieron respuesta afirmativa), vemos que para el valor 0 del ítem no existe error pero si hay 29 errores por debajo del cutting point (sujetos que han contestado afirmativamente el ítem y cuyo puntaje era menor que 3). En el ítem 2 existen 64 errores, 10 errores en 0 y 54 errores en 1; y en el ítem 3 hay 73 errores. El total de sujetos fue de 350, y es el resultado de la suma de la última columna, cuyos parciales indican la cantidad de sujetos que obtuvieron puntaje 3, 2, 1 y 0 respectivamente.

El coeficiente de reproductibilidad en el ejemplo es bastante bajo, y de haberse incluido más items en la escala, podría haberse eliminado algunos a los fines de aumentar la reproductibilidad, ya que la idea general es precisamente utilizar el escalograma para seleccionar ítems escalables.

El rango marginal mínimo, nos da un valor que restado del coeficiente de reproductibilidad proporciona el porcentaje de mejoramiento o como preferimos llamarle, alcance de la distribución marginal. En este caso el valor es de .1731 muy aceptable.

El coeficiente de escalabilidad corresponde a nuestra pauta de errores y en el ejemplo que presentamos representa un valor de .3541 muy bajo e indica que la escala no es unidimensional (los coeficientes deben ser mayores que .70).

Finalmente, las correlaciones que aparecen al fondo de la tabla son coeficientes Q de Yule para la correlación inter ítems, y coeficientes biseriales para la intercorrelación de cada uno de los ítems con la suma del resto de los otros ítems. Con esto podemos analizar los ítems que no se correlacionan, ya sea a los otros ítems, ya sea a los valores de escala.

XIV. BIBLIOGRAFIA

H. Blalock: Social Statistics; McGraw Hill, Kogakusha, Tokio, 1972, (2a. ed.)

T. Cacoullos (ed): Discriminant analysis and applications: Academic Press, New York and London, 1973.

Eisenbeis, R. y Avery, R.: Discriminant analysis and classification procedures; Lexington Books, Mass., 1972.

Fruchter, B.: Introduction to factor analysis; P. van Nostrand, Princeton, N. J., 1954.

Gansner, D., Seegrist, D., Walton, G.: "Technique for defining subareas for regional analysis", Growth and Change, October, 1971.

Guilford, J. P.: Psychometric methods; McGraw-Hill, New York, 1954.

Gould, P.R.: "On the geographical interpretation of eigenvalue", en Institute of British Geographers, No. 42, December, 1967 (pp. 53-85)

Keerlinger, F.: Foundations of behavioral research; Holt, Rinehart and Winston, New York, 1973, (2a. ed.)

N. Nie, C.H, J. Jenkins, K, Steinbrenner, D. Bent: Statistical package for the social sciences; McGraw-Hill, New York, 1975, (2a. ed.)

Padua, Jorge: Escalas para la medición de actitudes, CES, El Colegio de México, 1974.

Se terminó de imprimir en el mes de
octubre de 1975 en Imprenta Madero,
S. A., Avena 102, México 13, D. F.
Se tiraron 1 000 ejemplares.

Cuadernos del Centro de Estudios Sociológicos

1.	Sistemas de relaciones obrero-patronales en América Latina, por Rodolfo Stavenhagen y Francisco Zapata	$ 5.00	Dls. 0.50
2.	Las migraciones rural-urbanas, por Claudio Stern	5.00	0.50
3.	Control político, estabilidad y desarrollo en México, por José Luis Reyna	8.00	0.80
4.	Las relaciones entre el movimiento y el gobierno de Salvador Allende, 1970-1973, por Francisco Zapata	5.00	0.50
5.	Aspectos psicológicos del rendimiento escolar, por Jorge Padua	5.00	0.50
6.	Estado y sociedad civil: patrón de emergencia y desarrollo del Estado argentino, 1810-1936, por Leopoldo Allub	8.00	0.80
7.	El proceso chileno de transformación y los problemas de dirección política, 1970-1973, por Hugo Zemelman	5.00	0.50
8.	Organización de las sociedades de crédito ejidal de La Laguna, por Silvia Gómez Tagle	8.00	0.80
9.	Espaldas mojadas: materia prima para la expansión del capital norteamericano, por Jorge A. Bustamante	12.00	1.10
10.	Agricultura capitalista y agricultura campesina en México (diferencias regionales en base al análisis de datos censales), por Kirsten A. de Appendini y Vania Almeida de Salles	12.00	1.10
11.	Tensiones estructurales y diferenciación en las organizaciones: ¿Un caso de acumulación teórica?, por Viviane B. de Márquez	12.00	1.10

Pedidos a: El Colegio de México
Departamento de Publicaciones
Guanajuato 125, México 7, D. F. Tels.: 584-05-85 y 584-86-63