UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Unidad Académica de los Ciclos
Profesional y de Posgrado
Colegio de Ciencias y
Humanidades

1989

III FORO DE *Estadística Aplicada*

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Unidad Académica de los Ciclos
Profesional y de Posgrado
Colegio de Ciencias y
Humanidades

1989

III FORO DE *Estadística Aplicada*

INDICE

P R E S E N T A C I O N

En estas Memorias se incluyen los trabajos seleccionados por el Comité Editorial del Tercer Foro de Estadística. Dicho evento académico se llevó a cabo en el mes de Septiembre de 1988 en la Ciudad de Guanajuato, Gto. La organización del Foro estuvo a cargo del Centro de Investigaciones en Matemáticas en Guanajuato y de la UNAM a través de la Facultad de Ciencias, el Instituto de Investigaciones en Matemáticas Aplicadas y la Unidad Académica de los Ciclos Profesionales y de Posgrado del CCH.

En su tercera celebración el Foro de Estadística logró conjuntar el trabajo de participantes de diversos puntos del país. De esta forma con estas memorias se tendrá una mejor idea del nivel de investigación y de aplicación de la Estadística que actualmente se desarrolla en México

Esperamos que en las futuras celebraciones del Foro de Estadística se confirme su carácter nacional y se convierta en el espacio de intercambio, de discusión y de reflexión de todos los especialistas en Estadística en México.

DR. VICTOR AGUIRRE TORRES DR. RUBEN HERNANDEZ CID

Organizadores del
Tercer Foro de Estadística

Septiembre de 1989.

EL PARADIGMA CUANTITATIVO VS. EL CUALITATIVO

EN LA INVESTIGACION

Dr. Ignacio Méndez Ramírez
Departamento de Estadística
IIMAS - UNAM

I. Introducción

En este ensayo se pretenden plantear y comentar las dos diferentes maneras de investigar que han surgido en el desarrollo de la metodología científica. Las ciencias naturales desarrollaron los métodos experimentales y una creciente matematización de la ciencia, lo que condujo junto con otros componentes al llamado **Paradigma Cuantitativo**. Las ciencias sociales, en cambio, ante la dificultad del control experimental y la complejidad de sus conceptos, hicieron un uso mayor de las técnicas observacionales, las que también con otros componentes formaron el **Paradigma Cualitativo.**

Se discuten ambos paradigmas, para concluir que nunca se ha tenido la adherencia total a uno de ellos, además que los modernos enfoques filosóficos de la ciencia, llevan a la consideración de la conveniencia de mezclar los dos paradigmas, para superar las limitaciones de cada uno por separado.

Se hace una brevísima exposición sobre los niveles de metodología científica. En la parte final, se discuten los conceptos de validez externa e interna, en los que aparece esa síntesis de los dos paradigmas.

Se postula que todo investigador debe conocer los paradigmas, para que en su práctica de investigación, tome los aspectos convenientes de cada uno de ellos.

II. Niveles de Metodología Científica

Se considera que hay tres niveles de conceptualización y enfoque de la ciencia. Estos son:

Nivel de Filosofía de la Ciencia.- Es el nivel más general y con menos apoyo empírico. En él se discuten conceptos tales como leyes, causalidad, conocimiento, objetividad, realidad, verdad, forma, contenido, relación, necesidad, inducción deducción, etcétera.

Nivel de Metodología General.- Se incluyen aquí conceptos y métodos que pueden haber surgido de una ciencia particular, pero que tienden a usarse o ya se usan, en cualquier otra ciencia. Aquí entran los métodos matemáticos y estadísticos, los conceptos de información, algoritmo, sistema, control, cibernética, probabilidad, etcétera.

Nivel Metodológico Específico.- Es el nivel menos general y con mayor soporte empírico, en el que se incluyen métodos y conceptos de utilidad en una o varias disciplinas científicas, pero no en todas. Como métodos tenemos los estudios de caso,

los estudios experimentales, los observacionales, etcétera; y los conceptos según la disciplina tales como célula, fisiología, evolución, competencia, fuerza, masa, campo, individuo, libertad, agresión, inteligencia, etcétera.

III. Nivel de Filosofía de la Ciencia

En general, las grandes escuelas filosóficas, han ido superando deficiencias, al modificar sus postulados, lo que se estimula con el propio avance del conocimiento.

El racionalismo de la Antigua Grecia, consideraba que con sólo razonar podríamos adquirir conocimiento sobre el mundo. Posteriormente con el Renacimiento surge el inductivismo, que nos lleva al otro extremo y postula que con la observación y la experimentación, sin hipótesis ni teorías, podremos conocer el mundo. Más adelante, el positivismo y neopositivismo mejoran esta última posición, al señalar que con el uso de hipótesis y teorías, que se pueden verificar, además de un acercamiento totalmente objetivo al mundo, se obtendrá conocimiento verdadero de la realidad. Los falsacionistas como Popper, consideran que no es factible probar o verificar hipótesis, sino sólo rechazarlas (falsarlas), aunque sostienen la objetividad.

Algo que modifica lo anterior, es la consideración de que no es posible ser totalmente objetivo, al estudiar la realidad del mundo, ya que el tipo de elementos que estudiemos, los métodos con que lo hagamos y la forma de interpretar los resultados, depende de nuestras teorías y expectativas. Se reconoce ahora que no hay hechos puros, sino que siempre tienen componentes subjetivos, derivados de esos aspectos señalados, que dependen de la teoría.

En la ciencia como en la "vida diaria", las cosas deben verse para creerse, así como, creerse para verse y las preguntas deben estar ya un poco contestadas, si van a ser preguntadas.

Otro aspecto fundamental, surgido de la física y la sociología, es que siempre hay en mayor o menor medida una interferencia o interacción entre el investigador y el fenómeno que estudia, lo que de nuevo nos lleva a la imposibilidad de tener hechos absolutamente independientes del investigador.

T.S. Kuhn considera que un paradigma es el conjunto de conceptos, teorías y métodos que comparten un conjunto de científicos para estudiar y explicar el mundo.

Se reconoce así, que siempre hay subjetividad y cierto grado de irracionalidad (al menos no hay una justificación racional clara) en el conocimiento científico. Esto ha sido postulado, entre otros, por Kuhn, Feyarabend y Lakatos.

Pero la ausencia de objetividad supone una cantidad de matices, desde la descripción errónea y la parcialidad; hasta el simple hecho de la preferencia por un tipo de problemas y técnicas. Además del uso, siempre presente en cualquier tipo de investigación, de los conocimientos teóricos que posee el investigador. Como una regla casi universal, se busca la mayor objetividad posible, procurando eliminar supuestos que entren en contradicción con hechos o teorías apoyadas empíricamente, y más que nada, la búsqueda de concenso entre los científicos de un área de conocimiento dada.

Una posición filosófica reciente es el realismo no representativo. Realismo en el sentido de materialismo y no representativo porque considera que el conocimiento no corresponde cabalmente con la realidad, en este sentido no es verdadero. Bajo esta

posición, al conocimiento le podremos considerar como un modelo de la realidad, que nos permite dar una explicación, hacer predicciones y modificar la propia realidad.

IV. Paradigmas Cualitativo y Cuantitativo

Así como hay diferentes escuelas en los niveles metodológico general y filosófico, las hay en el metodológico específico.

Para las ciencias sociales en particular, está ocurriendo un cambio de paradigma, que se refleja en el uso cada vez mayor de la estadística y las matemáticas en general, en áreas como ciencias políticas, sociología, sicología, medicina (que tiene un componente social), y ciencias afines. Por otro lado, en las llamadas ciencias naturales como la física, química y biología; el reconocimiento de la falibilidad de la observación (por su carga teórica) y de la interacción entre el observador y el objeto observado; ha llevado también a un cambio de conceptos y métodos. En estas últimas ciencias, se reconocen actualmente dos aspectos básicos; que no es posible ser totalmente objetivo y la naturaleza no representativa del conocimiento obtenido. Esto último choca con los métodos específicos derivados del neopositivismo, que preconiza la medición numérica objetiva que refleje la realidad, como única manera válida de estudiar el mundo. Se llegó a decir que no es ciencia la que no usa las matemáticas. Fue el positivismo el que dió origen a la estadística, inmersa en el paradigma cuantitativo de las ciencias naturales.

Las ciencias sociales en general tienen, entre otras, las siguientes características: sus conceptos son difíciles de definir, por ejemplo cultura, democracia, agresión,

represión, libertad, competencia, etcétera; hay una gran dificultad para realizar experimentos controlados; existe una enorme interacción entre el observador y los fenómenos observados; finalmente, la teoría juega un papel muy importante en el proceso de observación. En virtud de estas características, las ciencias sociales en su origen, prácticamente no utilizan herramientas matemáticas y dentro de ellas la estadística. De este modo, sus métodos tradicionales son los estudios de caso único, sin controles, con métodos observacionales llamados "clínicos" o de observación participante. Sus métodos hacen un uso fuerte de las concepciones teóricas para discutir cada elemento observado; con un intento de llegar a explicaciones exhaustivas de lo estudiado y con posición antirreduccionista o integral. En general, el paradigma de investigación de las ciencias sociales es el llamado cualitativo.

Existen algunas ciencias como la agronomía, que inicialmente se consideran una rama aplicada de la biología y por tanto sus métodos de investigación se enmarcan dentro del paradigma cuantitativo. Modernamente, se define la agronomía como la ciencia que estudia la agricultura (en un sentido amplio comprende el uso de plantas y animales), y ésta como un proceso social de producción que usa plantas y animales para obtener bienes útiles al hombre. Resulta así, que la agronomía tiene ahora, componentes sociales y biológicas. Sin embargo, sus métodos de investigación aún tienen como parte principal el paradigma cuantitativo, entre otras cosas con excesivo énfasis en los experimentos. Quizá haya otros casos de ciencias, cuya concepción parte de las ciencias naturales y evolución hacia las sociales, como la epidemiología. También puede haber ciencias como la psicología, que inicialmente se originan en las ciencias sociales y cada

vez más involucran componentes naturales. Esto se refleja en un aumento del uso de la estadística y la matemática.

Podemos conceptualizar entonces, en su origen, una división fuerte de los métodos de investigación específicos: por un lado las ciencias sociales que utilizan el paradigma cualitativo, y por el otro "las ciencias naturales" que utilizan el paradigma cuantitativo.

Las características extremas de ambos paradigmas son las de la tabla siguiente; señalando que los atributos de uno y otro son independientes desde un punto de vista lógico.

ATRIBUTOS DE LOS PARADIGMAS*

PARADIGMA CUALITATIVO	PARADIGMA CUANTITATIVO
Fenómenologismo (Comprensión) "interesado en comprender la conducta humana desde el propio marco de referencia de quien actúa".	Positivismo lógico; "busca los hechos o causas de los fenómenos sociales, prestando escasa atención a los estados subjetivos e los individuos".
Observación naturalista y sin control.	Medición penetrante y controlada.
Subjetivo.	Objetivo.
Próximo a los datos; perspectiva "desde adentro".	Al margen de los datos; perspectiva "desde fuera".
Fundamentado en la realidad, orientado a los descubrimientos; exploratorio, expansionista, descriptivo e inductivo.	No fundamentado en la realidad, orientado a la comprobación, confirmatorio, reduccionista, inferencial e hipotético deductivo.
Orientado al proceso.	Orientado al resultado.
Válido: datos "reales", "ricos" y "profundos".	Confiable, datos "sólidos", repetibles ("hard data").
No generalizable: estudios de casos aislados.	Generalizable: estudios de casos múltiples.
Observacional.	Experimental.
Holista (integral) Sistémico.	Particularista, Analítico, Reduccionista.
Supone una realidad dinámica.	Supone una realidad estable.
Etnometodología.	Encuestas.
Conceptos por definirse en la propia investigación.	Conceptos definidos apriori y luego medir indicadores.
Descubrimiento de Teoría.	Comprobación de teoría.

* Tomado con modificaciones de Cook T.D. y Ch.S. Reichard "Métodos Cualitativos y Cuantitativos en Investigación Evaluativa "Ediciones Morata. Madrid, 1986.

La presentación anterior de los paradigmas es una idealización, podemos decir que lo que se ha dado es una preponderancia fuerte del paradigma cualitativo en ciencias sociales y del cuantitativo en las naturales. Es muy difícil encontrar una investigación enmarcada totalmente en uno cualquiera de los paradigmas. En realidad lo que se observa es un acercamiento cada vez mayor de los dos paradigmas, para emerger como uno sólo, de la síntesis de los anteriores. Es decir se procede actualmente a un acercamiento en las dos direcciones. Las ciencias sociales incorporando cada vez más atributos del paradigma cuantitativo, como la estadística, los experimentos y los controles, a sus métodos tradicionales. También las ciencias naturales tomando aspectos del paradigma cualitativo como: la definición verbal de conceptos complejos, la extrapolación de resultados de un contexto a otro en base a la teoría, el enfoque sistémico antirreduccionista, la incorporación de relaciones de causalidad no determinística y los estudios observacionales, entre otros aspectos.

V. Síntesis de los Paradigmas

Como se señaló, los métodos de investigación actuales, tienden a la síntesis de ambos paradigmas, el apoyo en uno u otro paradigma no implica la exclusión de atributos del otro. Veamos algunas ideas que resaltan esta síntesis, al hacer referencia al caracter no absoluto de los atributos de cada paradigma, señalados en el cuadro anterior.

Un investigador considerado positivista lógico, puede utilizar aspectos cuantitativos y estudiar a profundidad, con investigación no estructurada, algunos

casos, al mismo tiempo que hace mediciones cuantitativas en todos o una muestra aleatoria de sus casos. Por otro lado, un investigador fenomenologista puede usar experimentos y controles.

Si subjetivo es lo "influido por el juicio humano", entonces todas las investigaciones son subjetivas, ya sea que se apeguen a uno u otro paradigma; ya que todos los hechos se estudian y evalúan a la luz de la teoría, por lo que al menos son parcialmente subjetivos. La asignación de números y el empleo de métodos matemáticos (estadísticos) no garantiza nunca la objetividad, así un experimento en un ambiente controlado, lleva la subjetividad, por lo menos en la elección de las variantes en estudio (tratamientos) y en las características del ambiente controlado. Si se toma por subjetivo, la medición de creencias y sentimientos, entonces una encuesta de opinión, enmarcada tradicionalmente en el paradigma cuantitativo, resulta subjetiva.

Los métodos cuantitativos no excluyen un acercamiento e interiorización, por parte del investigador a los fenómenos estudiados, como cuando se deja parte de la investigación sin estructurar y el investigador observa y mide lo que surja, según su acercamiento, sin una hipótesis especificada previamente.

Una misma realidad se puede estudiar tanto desde el punto de vista sistémico o expansionista, como del reduccionista, y esto se podrá hacer por el mismo investigador en una o más investigaciones.

Un proceso se puede estudiar con experimentos, con énfasis en la medición longitudinal del desarrollo y no sólo medir el resultado final. Así también un estudio de caso puede centrarse en el resultado final, vg. efecto de la erupción del volcán

"Chichonal" en la producción agrícola.

Los métodos cualitativos pueden ser confiables además de válidos; confiable en el sentido de que en varias circunstancias (pueden ser estudios de caso) y por varios investigadores se obtengan resultados semejantes. Asimismo, un experimento o un muestreo aleatorio puede ser válido, además de confiable; válido en el sentido de medir y evaluar a profundidad realmente lo que se pretende medir. En ambos casos todo depende de los conceptos que sirvan de base a la investigación. Ni la confiabilidad, ni la validez, dependen totalmente del instrumento de medición y el método particular de investigación.

Los métodos cualitativos, se señala en el cuadro, no son generalizables, esto es relativo, ya que aún en el estudio de un sólo caso, se generaliza en base a la teoría, además de que se pueden tener otros casos que bajo algunos supuestos, se consideren semejantes al estudiado. La generalización depende entonces de la teoría, además del tamaño de muestra. Por ejemplo, una investigación del Instituto Nacional de Siquiatría, sobre migración y salud mental, se concentra en el pueblo de S. Vicente Chicoloapan, México y en él se hace un muestreo probabilístico. Se está mezclando el estudio de caso del paradigma cualitativo con el muestreo del cuantitativo.

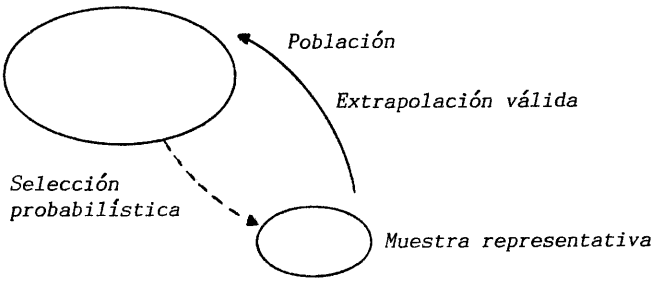
Vemos así, que los paradigmas no determinan totalmente la elección de métodos y que un buen investigador, eligirá atributos de ambos paradigmas, dependiendo de sus propósitos, sus medios, y objeto de estudio; para determinar así su método particular de investigación.

La elección de un modelo estadístico que encaje con los datos, la interpretación

de los resultados del análisis y la generalización de los descubrimientos a otros entornos, se hallan basadas en un conocimiento cualitativo. Simplemente, los investigadores no pueden beneficiarse del empleo de números si no conocen, en términos de sentido común, lo que éstos significan.

VI. Validez Externa

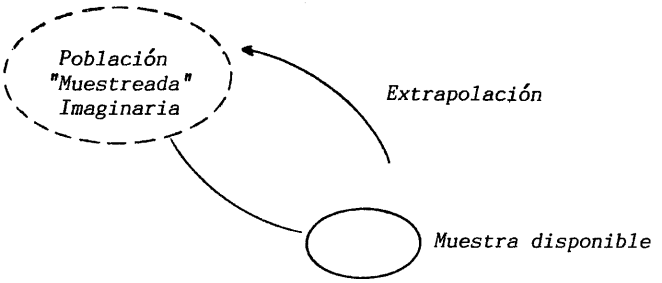
Típicamente en el paradigma cuantitativo, pero también implícitamente en el cualitativo, se presenta el problema de la generalización, extrapolación de los resultados obtenidos en un grupo de elementos en un lugar y época a un grupo mayor de ellos (la población objetivo) en otros lugares y/o épocas. Para que esta extrapolación sea correcta, el estudio debe tener Validez Externa. Esto consiste en el hecho de que los elementos estudiados sean representativos de la población objetivo, porque en ellos las características relevantes (según la teoría) al proceso estudiado, sean semejantes a las de toda la población.



La recomendación más usual es que se defina primero la población, y de ella se tome la muestra en forma aleatoria (con probabilidades de selección conocidas) para

que resulte representativa. Esto es lo común en el paradigma cuantitativo.

En ambos paradigmas es frecuente que por razones prácticas se disponga de uno o más elementos, sin que provengan de selección aleatoria de una población. En este caso la población muestreada, se construye hipotéticamente como la constituida por elementos semejantes a los estudiados. Esa semejanza se refiere a las características relevantes al fenómeno estudiado. En este caso la validez externa existe para esa población imaginaria.



La determinación de qué características son relevantes al fenómeno y hasta qué grado las variaciones en ellas se consideran "semejantes" a las de la muestra disponible, se determinan en base a los conocimientos teóricos del investigador, por lo que en parte son subjetivos.

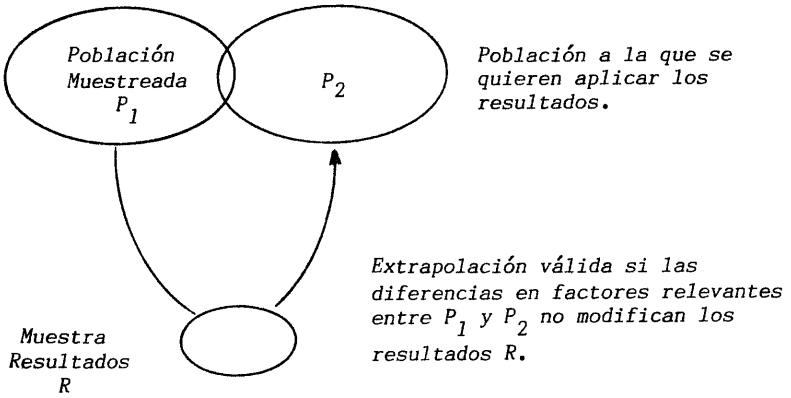
A menudo, ya sea que la población muestrada sea real (primer caso anterior) o imaginaria (último caso anterior), es necesario en ambos paradigmas aplicar el conocimiento obtenido en los elementos de estudio a otra población diferente a la muestreada. Para poder extrapolar los conocimientos de los elementos estudiados a una población de elementos, para la cual los estudiados no son representativos, se

requieren dos consideraciones:

1. ¿ En qué difieren (factores) y en qué magnitud la población muestreada y aquella a la que se pretenden aplicar los resultados?

2. ¿ Las diferencias anteriores invalidan los resultados obtenidos en los elementos estudiados, en el sentido de que en la nueva población ya no se espera que se mantengan los resultados?

Si la segunda pregunta se contesta negativamente, se podrá hacer la extrapolación de la muestra, a la población para la cual ella no es representativa. En caso contrario, no es válida la extrapolación.



En este proceso, tan común en la aplicación de desarrollos científicos y tecnológicos a ambientes distintos al que los generó, se requieren conocimientos teóricos sólidos. Esos conocimientos son los que permiten contestar las preguntas 1 y 2 anteriores, por lo general con bastante subjetividad.

Este proceso de razonamiento es el que permite utilizar los conocimientos obtenidos en los estudios de caso único del paradigma cualitativo. Como se puede ver, para la extrapolación señalada no es tan importante el tamaño de la muestra.

En realidad, al efectuar un experimento (o un muestreo aleatorio), típicos del paradigma cuantitativo, se está seleccionando un caso único, un elemento de la superpoblación de condiciones experimentales (o poblaciones muestreadas). Por esto comparte características del paradigma cualitativo. De nuevo la extrapolación a otras condiciones experimentales se apoya en la teoría disponible.

VII. Validez Interna

En el paradigma cuantitativo se utilizan los estudios comparativos (experimentos o pseudoexperimentos) en los que se somete a prueba una deducción de una hipótesis que relaciona un factor X (como causa o factor asociado) con un factor Y (como efecto o factor asociado).

Se deben estudiar grupos de elementos con algunas variantes del factor X, para evaluar la presencia y magnitud del factor Y. Sin embargo, estos grupos no deben diferir en otras características relevantes a la asociación entre X y Y, a esto se le denomina que hay comparabilidad o validez interna, o bien que se han eliminado (o controlado) factores de confusión.

Esto también ocurre en los estudios de caso único, en general en observación naturista, del paradigma cualitativo, en el momento de construir teorías que relacionen

un factor X con uno Y.

Un factor de confusión es aquel que cumple los dos requisitos siguientes:

a. Se encuentra presente de modo diferente en los grupos en que X varía. Es decir tiende a variar en forma concomitante con X.

b. modifica la relación de X con Y; al aumentar o disminuir la influencia o asociación de X con Y (es decir interactúa con X en la producción de Y), o bien tiene un efecto independiente de X, sobre Y.

La validez interna consiste en diseñar la investigación, de manera que a) no se cumpla.

En ambos paradigmas, en base a conocimientos teóricos, el investigador propone las variables X y Y; y además determina los posibles factores de confusión y la forma en que pueden afectar la relación de X con Y. De nuevo aquí hay subjetividad.

Las formas de controlar o tomar en cuenta la influencia de los factores de confusión, son las siguientes:

C1.- Homogenización de los sujetos de estudio en el factor de confusión, de modo que éste no varíe. Con esto se reduce la validez externa.

C2.- Bloques, igualación de atributos o estratificación, que supone además independencia o aditividad del efecto de X y del factor de confusión sobre Y.

C3.- Aleatorización de los elementos de estudio a las variantes del factor causal (X), de nuevo se supone que el factor de confusión no interactúa con X en la producción de Y, sino que tiene efectos aditivos independientes.

C4.- Si el factor de confusión interactúa con X en la producción de Y, lo conveniente es considerar por separado la relación entre X y Y para cada variante del factor de confusión.

C5.- Mediante modelos estadísticos se pueden suponer varias formas de acción del factor de confusión sobre Y, con o sin interacción, para evaluar cuál es la más probable. Con el modelo de mejor ajuste, se pueden ahora explicar las relaciones entre esos tres factores: X, Y y el de confusión.

Para determinar cuáles son los posibles factores de confusión, su forma probable de acción y la de su control, el investigador debe recurrir a sus conocimientos y experiencia, con lo que una vez más debe admitirse la subjetividad de algunos procesos en el trabajo ubicado dentro del paradigma cuantitativo, típico de la estadística.

Determinantes del Empleo Médico en México:
Una Aplicación del Análisis de Causalidad Nominal

Alagón J.[*], *Frenk J.*[**], *Muñoz del Río, A.*[**]

(*) Instituto Tecnológico Autónomo de México
Río Hondo # 1, Tizapán San Angel, México, DF 01000

(**) Instituto Nacional de Salud Pública
Dr Fco P Miranda # 177, Merced Gómez, México, DF 01480

Resumen

El artículo presenta una aplicación de la Estadística a las Ciencias Sociales. Las diversas formas en que los médicos en México se incorporan al mercado laboral urbano, son relacionadas con factores sociodemográficos a través de la metodología estadística del análisis de causalidad. La hipótesis central del trabajo es que variables como el grupo socioeconómico, tipo de escuela en donde se realizaron los estudios, tipo e institución de especialidad y clase generacional, son determinantes de los diferentes patrones de empleo médico.

El análisis está basado en información de la "Encuesta Nacional de Empleo Médico" de 1986. Dado que las variables bajo estudio son de naturaleza nominal, las técnicas clásicas de análisis de causalidad ya no pueden ser aplicadas. Es por ello que se utilizan técnicas especiales basadas en modelos logito. Las relaciones entre las variables de estudio son sintetizadas en diagramas y tablas causales.

INTRODUCCIÓN

Desde hace dos décadas, grandes cambios en el sistema educativo médico en México han ocasionado serias distorsiones en el mercado laboral médico. Quizás la evidencia más devastadora de ello pueda verse en la siguiente paradoja: mientras que el 14% de los médicos disponibles se encuentran desempleados o subempleados, más de diez millones de mexicanos carecen de acceso a los servicios de salud (Frenk *et al.* 1988). Resulta por ende imperativo un entendimiento cabal de los mecanismos subyacentes a este fenómeno; solamente así, podrá llevarse a cabo una toma de decisiones racionales, conducentes a mejorar esta situación de ineficiencia e injusticia social.

Una de los factores que más ha afectado la oferta de recursos humanos médicos es que las escuelas de medicina han sido el escenario de dos tendencias antagónicas. La primera de ellas comienza hacia el final de los sesentas, y se caracteriza por el aumento masivo tanto del número de escuelas, como de la matrícula en ellas, esta última alcanzando su máximo en 1978. La segunda tendencia comienza en 1979, caracterizándose por un decremento importante en la matrícula escolar; después de siete años (duración promedio de la carrera de Medicina), este decremento conlleva una disminución sustancial en el número de egresados. El exceso de oferta generado por la primera tendencia aún no ha sido compensado por la reducción en el número de graduados generada por la segunda tendencia. Esto explica, al menos en parte, el por qué el porcentaje de desempleo y subempleo ha permanecido alto durante los últimos años.

No obstante que se han realizado muchas investigaciones intentando explicar las causas del desempleo y subempleo médico, sólo recientemente se han dirigido los esfuerzos a intentar responder preguntas tales como:

¿Qué características tienen los médicos que ocupan los sitios privilegiados del mercado de trabajo médico?

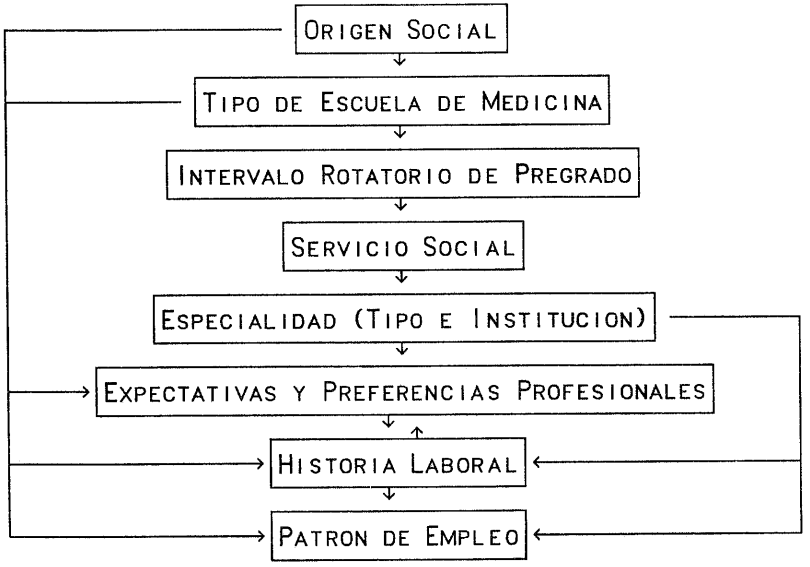
¿Qué tipo de educación permite a una persona tener acceso a estas posiciones?

¿Qué papel juega el origen social en este problema?

Los diversos patrones de empleo médico han sido conformados a partir de información empírica mediante el uso del análisis de conglomerados (Alagón *et al.* 1987). Estos patrones representan las distintas formas de inserción de los médicos al mercado laboral. En este artículo se analizan los determinantes de dichos patrones. Estos determinantes pueden ser personales o ambientales (Fig. 1). La presente investigación forma parte del proyecto "Patrones de Empleo Médico en las Areas Urbanas de México", comenzado en 1985 en el Instituto Nacional de Salud Pública.

La metodología escogida para estudiar la estructura e interrelaciones que rigen a los factores ilustrados en la Figura 1, es el análisis causal a partir de tablas de contingencia desarrollado y aplicado en las Ciencias Sociales primeramente por Goodman (1972, 1973); remitimos al lector interesado en este tema a estos artículos; para aquéllos interesados en cuestiones más generales de causalidad, las compilaciones hechas por Blalock (1964) y Aigner y Zellner (1988) resultan imprescindibles.

Figura 1. Posibles determinantes de la situación laboral de los médicos



METODOLOGÍA Y VARIABLES

El modelo representado en la Fig. 1 resultaría muy elaborado en términos del numero de casos así como del tiempo requerido para estudiar su adecuación. Es por ello que en este artículo se considera una versión reducida del mismo. De este modo, el modelo ahora consta de cinco variables: origen socioeconómico, escuela de medicina, tipo de especialidad, institución de especialidad, y patrón de empleo médico. Todas las variables son categóricas.

Origen socioeconómico se construyó a partir de tres características de la persona que sostuvo al médico mientras estudiaba la carrera: máximo grado académico obtenido, ocupación, y posición en el trabajo (empleado, patrón, práctica independiente, etc.). Las tres categorías resultantes son: bajo, medio y alto, de acuerdo al bienestar económico inferido.

Escuela de medicina mide dos atributos de la escuela donde el médico recibió su educación: tamaño, medido como matrícula total en 1978, y calidad, medida como el porcentaje de aprobados en un examen estándar al final de la carrera. Cada indicador fue discretizado en categorías: chica y grande, para tamaño, y mala, regular, y buena, para calidad. Las escuelas, sin embargo, se agruparon en solamente cuatro de las seis combinaciones posibles de tamaño-calidad: chicas malas, chicas regulares, grandes regulares, y buenas[1].

[1] La última categoría (buenas) incorpora solamente escuelas chicas, con excepción de una grande. Abrirle una categoría a esta escuela hubiera afectado los resultados obtenidos en la siguiente sección.

Tipo de especialidad e **institución de especialidad** se ocupan de dos dimensiones del proceso mediante el cual un médico general se convierte en un especialista. La primera de ellas, tipo, registra la dificultad relativa de cada especialidad, mientras institución mide algunas características estructurales del lugar donde se realizó la especialidad. Las categorías resultantes para cada variable son como sigue: sin especialidad, medicina familiar, especialidades básicas (también llamadas troncales), y subespecialidades, para tipo; sin especialidad, asistencia pública, seguridad social, y otras (privada, extranjero, etc.) para institución. Dado que la inclusión simultánea de ambas variables lleva a un modelo relativamente complicado, se estimaron dos modelos por separado, ajustando uno para cada variable.

Patrón de empleo médico es la variable dependiente. Su construcción está descrita en Alagón *et al.* (1987). Para fines de este artículo, baste decir que es una variable construída a partir de varias características del empleo que posee el médico (tales como el número de trabajos, la productividad, el ingreso, la posición en el trabajo y el nivel de ocupación, entre otras). La clasificación resultante obtenida en el artículo mencionado (Fig. 2) no es apropiada para efectos de la presente investigación, pues el número de observaciones es pequeño relativo al número de celdas. Es por ello que una primera solución a este problema fuera el dicotomizar los patrones de empleo en deseables e indeseables. El análisis que se presenta en este artículo está basado en esta variable dependiente en forma dicotómica. De esta manera, resulta natural el utilizar modelos logito para explicar las relaciones entre las variables independientes y la variable dependiente con dos categorías.

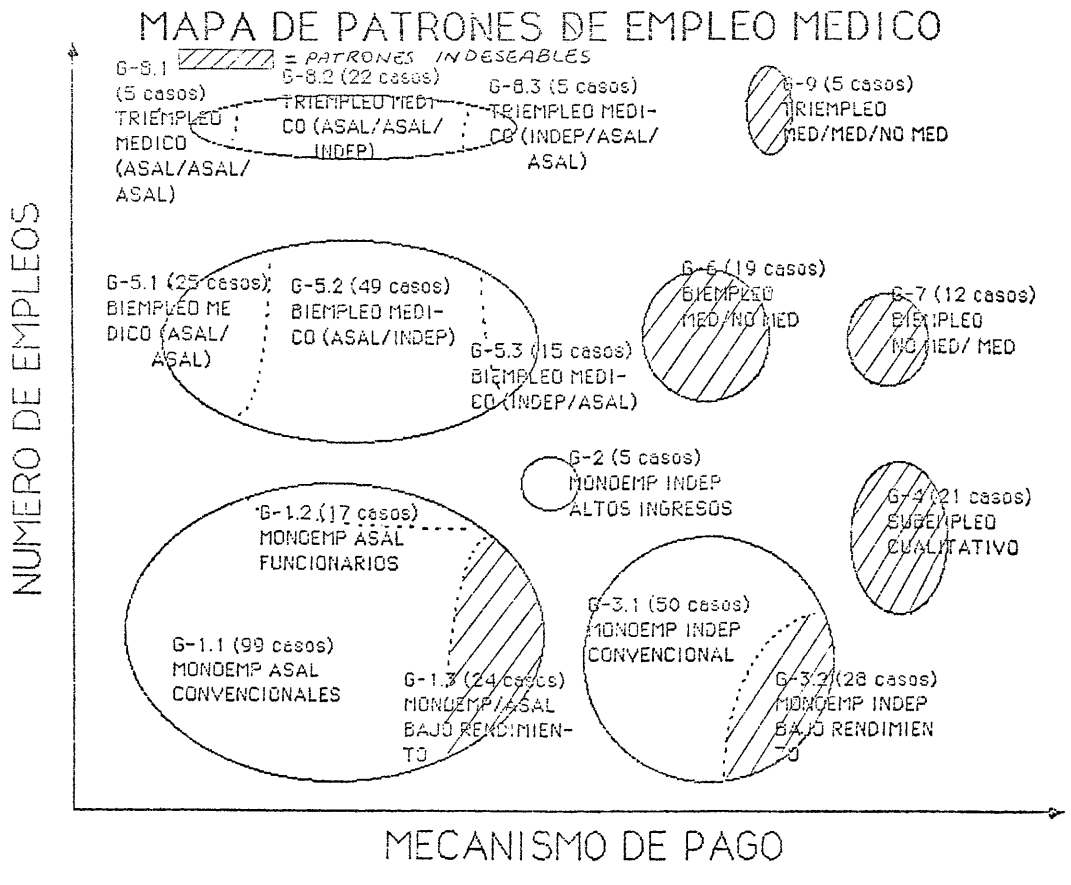
Todos los datos proceden de la Encuesta Nacional de Empleo Médico (ENEM), levantada en el primer semestre de 1986 por el Instituto Nacional de Estadística, Geografía e Informática. La muestra de médicos se obtuvo como sigue: la Encuesta Nacional de Empleo Urbano (ENEU), una encuesta periódica de panel rotatorio, se levanta cada trimestre en dieciséis de las principales áreas urbanas del país; en un trimestre particular, en 620 de los 41,000 hogares encuestados hubo al menos una persona que declaró haber terminado cuando menos el quinto año de la carrera de medicina. Estos 620 casos constituyen la muestra seleccionada, aunque sólo 604 fueron de hecho entrevistados.

Análisis Causal

En esta sección se describe en forma muy general las técnicas de análisis causal para datos nominales propuestas por Goodman. El principal objetivo de estas técnicas es explicar el comportamiento de una variable dependiente (patrón de empleo médico, en nuestro caso) a partir de variables que la preceden causalmente (origen socioeconómico, escuela de medicina, tipo e institución de especialidad).

Dado que el nivel de medición de las variables bajo estudio es nominal, éstas pueden ser modeladas convenientemente a través de modelos log-lineales. Un caso particular de estos modelos, los modelos *logito*, surgen cuando una de las variables puede ser considerada dependiente de las restantes (ver, por ejemplo,

FIGURA 2


MAPA DE PATRONES DE EMPLEO MEDICO
= PATRONES INDESEABLES
G-8.1 (5 casos) TRIEMPLEO MEDICO (ASAL/ASAL/ASAL)
G-8.2 (22 casos) TRIEMPLEO MEDICO (ASAL/ASAL/INDEP)
G-8.3 (5 casos) TRIEMPLEO MEDICO (INDEP/ASAL/ASAL)
G-9 (5 casos) TRIEMPLEO MED/MED/NO MED
G-5.1 (25 casos) BIEMPLEO MEDICO (ASAL/ASAL)
G-5.2 (49 casos) BIEMPLEO MEDICO (ASAL/INDEP)
G-5.3 (15 casos) BIEMPLEO MEDICO (INDEP/ASAL)
G-6 (19 casos) BIEMPLEO MED/NO MED
G-7 (12 casos) BIEMPLEO NO MED/ MED
G-2 (5 casos) MONOEMP INDEP ALTOS INGRESOS
G-4 (21 casos) SUBEMPLEO CUALITATIVO
G-1.2 (17 casos) MONOEMP ASAL FUNCIONARIOS
G-1.1 (99 casos) MONOEMP ASAL CONVENCIONALES
G-1.3 (24 casos) MONOEMP ASAL BAJO RENDIMIENTO
G-3.1 (50 casos) MONOEMP INDEP CONVENCIONAL
G-3.2 (28 casos) MONOEMP INDEP BAJO RENDIMIENTO
NUMERO DE EMPLEOS
MECANISMO DE PAGO

Fienberg, 1980). Se puede resumir en forma esquemática el problema abordado en este artículo de la siguiente forma:

$$\text{Patrón de Empleo Médico} = F\left(\begin{array}{c}\text{Origen}\\ \text{socioeconómico,}\end{array}\ \begin{array}{c}\text{Escuela}\\ \text{de Medicina,}\end{array}\ \begin{array}{c}\text{Tipo o Institución}\\ \text{de especialidad}\end{array}\right)$$

Los modelos logito se basan en los *momios* de la variable dependiente. Estos se definen como la frecuencia (o probabilidad) de una categoría comparada con la frecuencia (o probabilidad) de otra. En nuestro caso, la probabilidad de tener un patrón de empleo deseable se compara con la probabilidad de tener uno indeseable. Dichas comparaciones se llevan a cabo por medio de cocientes. Las razones de momios son la mejor forma de comprender los efectos que tienen las variables dependientes en modelos log-lineales.

En nuestro caso, los modelos logito pueden expresarse matemáticamente como

$$\omega_{ijk} = \mu\ \alpha_i\ \beta_j\ \gamma_k$$

donde $i=1,2,3$, $j=1,2,3,4$, $k=1,2,3,4$ son los niveles de las variables origen socieconómico (OS), escuela de medicina (EscMed), y tipo de especialidad (TipEsp) o, alternativamente, institución de especialidad (InstEsp). Los parámetros α_i, β_j and γ_k miden los efectos principales de cada nivel de cada una de las variables explicatorias sobre los momios de la variable dependiente (momios de poseer un patrón de empleo deseable).

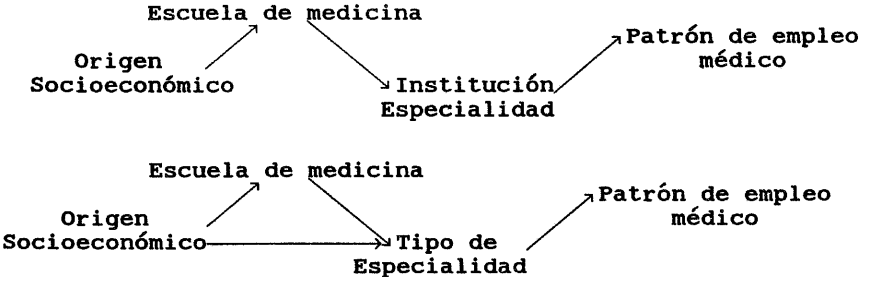
A partir de la información recabada en la Encuesta Nacional de Empleo Médico se ajustaron modelos logito con el propósito de explicar a la variable dependiente en términos de las independientes. Dado que las variables bajo estudio tienen un ordenamiento cronológico natural, el análisis causal se basa en ajustar modelos logito *recursivos*, modelos en los cuales se desea explicar una variable a partir de variables temporalmente precedentes. Bajo esta estrategia, el modelo logito de mejor ajuste[2] se encuentra para la tabla con escuela de medicina como variable respuesta explicada por el origen socioeconómico; para la siguiente tabla, estas dos variables (OS y EM) son ahora independientes, y la nueva variable dependiente es especialidad (tipo o institución). Finalmente, el patrón de empleo médico (PEM) dicotomizado juega el papel de variable dependiente, con las tres variables temporalmente precedentes como explicatorias.

RESULTADOS

Una vez que se han ajustado los modelos logito puede obtenerse un "mapa causal". Los mapas causales son diagramas que revelan los nexos causales entre las variables. Los mapas

[2] Por "modelo logito de mejor ajuste" entendemos un modelo logito parsimonioso (es decir, que usa el mínimo número posible de parámetros estimados) con ajuste estadístico satisfactorio. En otras palabras, intentamos explicar el máximo de variación posible con la representación más económica. Véase, por ejemplo, Benedetti y Brown, 1978, o Goodman, 1971, 1972.

causales obtenidos finalmente se muestran a continuación:



Para lograr una interpretación sustantiva de estos diagramas es necesario recurrir a los parámetros ajustados en los modelos logito. Estos parámetros representan una medida de los "riesgos relativos" de pertenecer a una de las categorías de la variable dependiente, dados los niveles de las variables independientes. En nuestro caso, los "riesgos" se refieren a los momios de pertenecer a un patrón deseable. Valores significativos al nivel de $\alpha = 0.05$ se representan con asterisco (*).

TABLA DE RIESGOS AJUSTADOS EN MODELOS LOGITO

OS		EscMed	
Bajo	*ChMala*	*1.606*	
Bajo	*GdeReg*	*1.251*	
Bajo	*ChReg*	*1.692*	
Bajo	*Buena*	*0.294**	
Medio	*ChMala*	*0.859*	
Medio	*GdeReg*	*0.893*	
Medio	*ChReg*	*0.803*	
Medio	*Buena*	*1.621**	
Alto	*ChMala*	*0.724*	
Alto	*GdeReg*	*0.895*	
Alto	*ChReg*	*0.735*	
Alto	*Buena*	*2.097**	

En la tabla anterior, los momios mayores a uno indican que médicos con ese origen socioeconómico son más susceptibles de estudiar en el tipo de escuela que corresponde al renglón donde el momio aparece con respecto a otros tipos de escuela. Si tomamos, por ejemplo, la última entrada de la tabla, el 2.097 nos dice que la probabilidad de que un médico de origen social alto realice sus estudios de medicina en una escuela buena son más o menos el doble que las que entre a cualquier otro tipo de escuela.

Los momios ajustados entre escuela de medicina por un lado, e institución y tipo por el otro, se muestran abajo. Los momios pueden interpretarse fácilmente, como se hizo para la tabla anterior.

EscMed		InstEsp		TipoEsp	
ChMala	*SinEsp*	*3.347**	*SinEsp*	*2.534**
ChMala	*AsPubl*	*1.073*	*MedFam*	*1.327*
ChMala	*SegSoc*	*1.019*	*BasSp*	*0.665*
ChMala	*Otras*	*0.273**	*Subesp*	*0.447**
GdeReg	*SinEsp*	*1.048*	*SinEsp*	*1.441*
GdeReg	*AsPubl*	*1.034*	*MedFam*	*0.312**
GdeReg	*SegSoc*	*0.609*	*BasSp*	*1.688*
GdeReg	*Otras*	*1.230*	*Subesp*	*1.315*
ChReg	*SinEsp*	*0.935*	*SinEsp*	*0.868*
ChReg	*AsPubl*	*0.966*	*MedFam*	*1.835*
ChReg	*SegSoc*	*0.889*	*BasSp*	*0.560*
ChReg	*Otras*	*1.244*	*Subesp*	*1.121*
Buena	*SinEsp*	*0.304**	*SinEsp*	*0.315**
Buena	*AsPubl*	*0.931*	*MedFam*	*1.314*
Buena	*SegSoc*	*1.812**	*BasSp*	*1.591*
Buena	*Otras*	*1.942**	*Subesp*	*1.516*

Para el modelo que incluye TipoEsp, fue necesario incluir el término de interacción entre origen socioeconómico y tipo de especialidad. Esto explica el que aparezca la flecha que une a ambas variables en el diagrama causal. Los riesgos correspondientes son como sigue:

OS		TipoEsp
Bajo	*SinEsp*	*1.267*
Bajo	*MedFam*	*2.255**
Bajo	*EspBas*	*0.379**
Bajo	*Subesp*	*0.922*
Medio	*SinEsp*	*0.763*
Medio	*MedFam*	*0.923*
Medio	*EspBas*	*1.567*
Medio	*Subesp*	*0.906*
Alto	*SinEsp*	*1.033*
Alto	*MedFam*	*0.481**
Alto	*EspBas*	*1.682**
Alto	*Subesp*	*1.197*

Hasta ahora se han dado todos los momios correspondientes a las relaciones ilustradas en los mapas causales, con excepción de los aparejados al último nexo causal, entre especialidad y patrón de empleo. Estos momios se dan a continuación:

InstEsp		PEM		TipoEsp		PEM	
SinEsp	*Deseable*	*0.229**		*SinEsp*	*Deseable*	*0.195**
AsPubl	*Deseable*	*1.380*		*MedFam*	*Deseable*	*3.571**
SegSoc	*Deseable*	*1.981**		*BasSp*	*Deseable*	*1.235*
Otras	*Deseable*	*0.807*		*Subesp*	*Deseable*	*1.161*

Para ambas modelos existen momios base que reflejan los momios *a priori* de poseer un patrón de empleo deseable, independientemente de los niveles de las variables restantes.

Esto es, los momios dados hasta ahora deben interpretarse en términos relativos; para que cobren sentido en términos absolutos la gran media también debe contemplarse. Los valores obtenidos son como sigue:

PEM (gran media)

Deseable **(InstEsp)** *4.670* *Deseable* **(TipoEsp)** *5.600*

Estos valores reflejan que la probabilidad de tener un patrón de empleo médico deseable son aproximadamente cinco veces las de tener uno indeseable, sin importar el origen socioeconómico, la escuela de medicina, ni el tipo o institución de especialidad.

Puede obtenerse un riesgo "total" que acumule los efectos parciales de los riesgos dados arriba. El cálculo de dicho riesgo total es sencillo, pues lo único que se debe hacer es multiplicar los riesgos parciales asociados a cada combinación de variables del individuo bajo consideración. Supongamos, por ejemplo, que un médico proviniente de un estrato socioeconómico bajo que estudió en una escuela grande regular y realizó una especialidad básica en una institución de asistencia pública, entomces el riesgo de que su patrón de empleo médico sea deseable se calcula como:

usando InstEsp:

1.251 x 1.034 x 1.380 x 4.670 = 8.336

Bajo→GdeReg *GdeReg→AsPubl* *AsPubl→Deseable* *Deseable*

usando TipoEsp:

1.251 x 1.688 x 0.379 x 1.235 x 5.600 = 5.535

Bajo→GdeReg GdeReg→Basica Bajo→Básica Básica→Deseable Deseable

El cálculo para cada combinación posible produce las Tablas 2 y 3.

Se analizará primero la tabla que incluye la variable institución de especialidad. Los momios de tener un patrón de empleo deseable muestran un pequeño incremento del origen social bajo, por un lado, a los estratos alto y medio, por el otro (de 3 a 4, aproximadamente). Esto significa que origen socioeconómico, por sí mismo, no determina el que un médico se ubique o no en un patrón de empleo deseable.

Riesgo de tener un Patrón de Empleo Médico Deseable

Tabla 2

Origen Socioecon.	Escuela Médica	Institucion de Especialidad				Media geom.
		SinEsp	*AsPubl*	*SegSoc*	*Otras*	
Bajo	*ChMala*	5.74	11.10	15.14	1.65	6.32
	GdeReg	1.40	8.34	7.05	5.80	4.67
	ChReg	1.69	8.19	1.39	1.74	2.41
	Buena	0.09	1.76	4.93	2.15	1.14
Media geométrica		1.05	6.04	5.20	2.45	3.00
Medio	*ChMala*	3.07	5.94	8.11	0.88	3.38
	GdeReg	1.00	5.95	5.03	4.14	3.34
	ChReg	1.69	4.60	6.61	3.76	3.73
	Buena	0.53	10.09	27.17	11.86	6.44
Media geométrica		1.29	6.36	9.25	3.57	4.06
Alto	*ChMala*	2.59	5.01	6.83	0.75	2.86
	GdeReg	1.00	5.96	5.04	4.15	3.34
	ChReg	0.73	4.21	6.05	3.45	2.83
	Buena	0.68	13.05	35.17	15.36	8.32
Media geométrica		1.06	6.36	9.25	3.58	3.87

Tabla 3

Origen Socioecon.	Escuela Médica	Tipo de Especialidad				Media geom.
		SinEsp	*MedFam*	*EspBas*	*Subesp*	
Bajo	*ChMala*	5.64	96.06	2.80	4.31	8.99
	GdeReg	2.50	17.62	5.54	9.86	7.00
	ChReg	2.04	42.71	2.47	11.40	7.04
	Buena	0.13	17.46	1.23	2.67	1.65
Media geométrica		1.39	33.52	2.62	6.00	5.20
Medio	*ChMala*	1.82	21.04	6.12	2.26	4.80
	GdeReg	1.08	5.15	16.34	6.92	5.01
	ChReg	0.58	8.29	4.87	5.30	3.34
	Buena	0.43	39.30	27.96	14.47	9.09
Media geométrica		0.83	13.71	10.80	5.88	5.20
Alto	*ChMala*	2.07	9.24	5.61	2.52	4.06
	GdeReg	1.46	2.69	17.57	9.15	5.01
	ChReg	0.71	3.96	4.79	18.30	3.96
	Buena	0.74	26.50	38.84	24.74	11.72
Media geométrica		1.12	7.15	11.64	10.11	5.54

En la tabla 2 también puede apreciarse que no existen diferencias marcadas entre los momios de tener un PEM deseable debidas a las distintas escuelas de medicina a las que asisten los médicos; al igual que en el párrafo anterior, los momios fluctúan entre 3 y 4. Esto va en contra de lo que se hubiera esperado: mejores escuelas debían incrementar las posibilidades de ubicarse en PEMs deseables. De hecho esta relación llega a cumplirse parcialmente para los médicos provinientes de los estratos medio y alto. Los médicos de origen social bajo, en cambio, tienen momios

decrecientes de pertenecer al sector deseable del mercado laboral a medida que asisten a mejores escuelas (véase la columna de hasta la derecha de la Tabla 2).

El factor que verdaderamente discrimina entre tener un patrón deseable *vs.* uno indeseable resulta ser la institución de especialidad. De hecho, los momios marginales se incrementan de 1.13 para médicos sin especialidad a 6.25, 7.63 y 3.15 para médicos con especialidad hecha en asistencia pública, seguridad social, y otras instituciones, respectivamente. El aumento en las probabilidades de pertenecer a un patrón deseable que resultan de tener una especialidad es más o menos constante a través de los distintos estratos socioeconómicos, excepto posiblemente que, médicos de clase media y alta con seguridad social tienen aproximadamente el doble de oportunidades que los médicos de origen bajo con la especialidad realizada en la misma institución.

La interpretación del modelo que incorpora al tipo de especialidad es muy similar a la que ya se ha hecho para el modelo con institución: los momios marginales para tanto origen socioeconómico como escuela de medicina no muestran diferencias de consideración entre sus categorías. Quizá la diferencia más evidente entre ambas tablas sea que el tipo de especialidad parece aumentar las posibilidades de poseer un patrón deseable aún más dramáticamente que lo que la institución lo hacía: de 1.09 para aquellos sin especialidad, a 14.87, 6.91, y 7.09 para aquellos con especialidad en medicina familiar, básica, y subespecialidad, respectivamente.

Finalmente, si se consideran los dos mapas causales, otra diferencia entre ambos modelos es evidente. La flecha que une al origen socioeconómico con institución de especialidad no fue incluída, mientras que la que va de origen social al tipo de especialidad sí lo fue. Su efecto puede apreciarse en la Tabla 3, donde el efecto de la especialidad varía a lo largo de los distintos estratos sociales. Una asociación positiva entre origen socioeconómico y tipo de especialidad ocasiona momios más altos para especialidades más difíciles a medida que aumenta el estrato social.

Los resultados obtenidos en este artículo son tan solo una aproximación al mecanismo subyacente determinante de los patrones de empleo médico. Existen diversas posibilidades metodológicas de análisis que pueden ser desarrolladas a partir de este trabajo para poder lograr un mejor entendimiento de dicho mecanismo. La importancia del problema lo amerita.

BIBLIOGRAFÍA

AIGNER, DJ y ZELLNER A eds. (1988) "Causality" Número especial del *Journal of Econometrics* **39**(1/2).

ALAGÓN, J, *et al.* (1987). "Análisis de conglomerados: una aplicación al problema de empleo médico", en **Memorias del 2 °Foro de Estadística Aplicada.** Mexico D.F.: UNAM.

BENEDETTI, JM and BROWN, MB (1978). "Strategies for the Selection of Log-Linear Models" *Biometrics* **16**(1):45-77.

BLALOCK, HM Jr (ed.) (1985). **Causal Models in the Social Sciences**, 2ª ed. Hawthorne, NY: Aldine Publishing Company.

FIENBERG, SE (1980). **The Analysis of Cross-Classified Contingency Data**, 2nd ed. Cambridge, MA: MIT Press.

FRENK, JJ, *et al.* (1988) "Subempleo y desempleo entre los médicos de las áreas urbanas de México" *Salud Publica de Mexico* **30**(5).

GOODMAN, LA (1971). "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models" *Technometrics* **13**(1):33-61.

———— (1972). "A General Model for the Analysis of Surveys" *American Journal of Socilogy* 77(6):1035-1086.

———— (1973). "The Analysis of Multidimensional Contingency Tables When Some Variables are Posterior to Others" *Biometrika* **60**(1):179-192.

AJUSTE DE UN MODELO LOG-LINEAL CON VARIABLES ORDINALES A DATOS DE SERVICIOS DEL SECTOR SALUD

Jorge M. Olguín U.

Depto. de Estadística

IIMAS-UNAM

En el análisis de tablas de contingencia multidimensionales frecuentemente se presentan variables cuyas categorías representan puntos en una escala ordinal. Esta información puede incorporarse en los modelos log-lineales dando como resultado una gran variedad de modelos interesantes diferentes a los que consideran a todas las variables como nominales. En este trabajo se presenta el ajuste de uno de estos modelos a una tabla de contingencia formada por dos variables ordinales y una nominal.

Los datos que se analizan fueron tomados de una muestra nacional llamada "Encuesta de Morbilidad Atendida" realizada por la Dirección General de Información y Estadística de la Secretaría de Salud.

De un estrato de esta muestra se consideraron los menores de edad con diagnóstico de estado nutricional y se clasificaron por edad y sexo con lo que se formó la tabla de tres dimensiones que se presenta en el cuadro 1.

Se tienen entonces las siguientes variables, su escala de medición y sus niveles o categorías:

1. Edad (E) es una variable ordinal con niveles 0, 1, 2, 3 y 4 o más.

2. Estado nutricional o desnutrición (D) también es una variable ordinal con

niveles: no, leve y moderada.

3. Sexo (S) que es nominal con niveles masculino y femenino.

CUADRO 1

			Edad			
Desnutri.	0	1	2	3	4+	total
No	37	9	7	8	16	77
Leve	7	1	4	0	2	14
Moderada	6	4	1	4	2	17
Total	50	14	12	12	20	108
			Edad			
Desnutri.	0	1	2	3	4+	total
No	24	12	9	7	4	56
Leve	6	4	4	1	4	19
Moderada	7	3	1	6	7	24
Total	37	19	14	14	15	99

Si se quiere ajustar un modelo log-lineal "estándar" (esto es, que considera a todas las variables como nominales) se tendrían que explorar casos particulares del modelo saturado para tres dimensiones (ver Bishop et al (1975))

$$log \quad m_{ikj} = \mu + \lambda_i^E + \lambda_j^D + \lambda_k^S + \lambda_{ij}^{ED} + \lambda_{ik}^{ES} + \lambda_{jk}^{DS} + \lambda_{ijk}^{EDS} \qquad (1)$$

$$i = 1,2,3,4,5; \quad j = 1,2,3; \quad k = 1,2$$

Donde m_{ijk} es el valor esperado en la celda ijk de la tabla bajo la suposición de que el modelo es correcto.

Los primeros cuatro términos de (1) corresponden al modelo de completa independencia entre las tres variables; cada uno de los términos con dos subíndices representa medidas de correlación entre las variables que aparecen en el superíndice y el último término corresponde a medidas de correlación entre las tres variables.

Un modelo que utiliza la información ordinal de las variables Edad y Desnutrición y que considera relaciones de primer orden entre todos los pares de variables sin considerar relación conjunta entre las tres variables es

$$log \quad m_{ijk} = \mu + \lambda_i^E + \lambda_j^D + \lambda_k^S + \beta^{ED}[u_i - \overline{u}][v_j - \overline{v}] + \tau_k^{ES}[u_i - \overline{u}] + \tau_k^{DS}[v_j - \overline{v}] \quad (2)$$

$$i = 1,2,3,4,5; \quad j = 1,2,3; \quad k = 1,2; \quad u_1 < u_2 < \ldots < u_5; \quad v_1 < v_2 < v_3$$

Para los datos del cuadro 1 consideramos casos particulares de este modelo con el propósito de seleccionar el más adecuado. Como se puede apreciar, este modelo contiene un número de parámetros considerablemente menor que el caso particular de (1) correspondiente y, como se verá a continuación, los parámetros en (2) tienen una interpretación sencilla en función de las razones de productos cruzados o momios.

Si $[u_i = i]$ y $[v_j = j]$ se tiene que

$$log\theta_{ij(k)} = \beta^{ED}$$

$$log\theta_{i(j)k} = \tau^{ES}_{k+1} - \tau^{ES}_{k}$$

$$log\theta_{(i)jk} = \tau^{DS}_{k+1} - \tau^{ES}_{k}$$

donde las θ's son los momios que se obtienen al considerar niveles adyacentes de las variables correspondientes a los subíndices que no se encuentran entre paréntesis, dado el nivel de la variable del subíndice que se encuentra entre paréntesis. Por ejemplo:

$$\theta_{ij(k)} = \frac{m_{ijk}m_{i+1j+1k}}{m_{i+1jk}m_{ij+1k}}$$

es un medida de la relación "local" entre las variables Edad y Desnutrición en sus niveles i y j respectivamente dado el nivel k de sexo.

Como se puede ver, en el model (2) esta medida es la misma independientemente de los niveles (adyacentes) de las variables Edad y Desnutrición y es igual al parámetro β^{ED}, por lo que se podría decir que señala una asociación uniforme entre las variables Edad y Desnutrición.

Así, si β^{ED} es diferente de cero, el término correspondiente en el modelo (2) refleja desviaciones de $logm_{ikj}$ del modelo de independencia; si es mayor que cero se esperan mayores frecuencias en valores altos (bajos) de E y D que bajo la situación de independencia; si es menor que cero, se esperarán valores mayores en los niveles bajos de una variable y altos de la otra que bajo la situación de independencia.

Los términos con τ's reflejan, para un nivel k de la variable Sexo, desviaciones

de los $logm_{ijk}$ del modelo de independencia que son funciones lineales de las variables ordinales.

Para el ajuste del modelo a los datos del cuadro 1 se considerarán para la variable Edad $[u_i] = [1, 2, 3, 4, 5]$ y para Desnutrición $[v_j] = [1, 2, 3]$.

Considérese primero el ajuste del modelo de completa independencia, es decir

$$logm_{ijk} = \mu + \lambda_i^E + \lambda_j^D + \lambda_k^S \tag{3}$$

Para las pruebas de bondad de ajuste es útil la estadística de razón de verosimilitud

$$G^2 = 2 \sum \quad \text{(observados)} \quad log\left[\frac{\text{observados}}{\text{esperados}}\right]$$

la cual tiene una distribución asintótica ji-cuadrada con grados de libertad igual a la diferencia entre el número de celdas y el número de parámetros linealmente independientes estimados. Esta estadística tiene la propiedad de que se puede particionar de manera que se pueden comparar dos modelos cuando uno es un caso particular del otro.

Al ajustar el modelo (3) a los datos del cuadro 1 se tiene que $G^2 = 33.62$ que comparado con el valor de ji-cuadrada con 22 grados de libertad se ubica en un nivel de significancia descriptivo de $p = 0.054$ por lo que se tiene alguna evidencia de correlación entre las variables sobre todo si se toma en cuenta que el tamaño de muestra utilizado es chico.

Para analizar con mayor detalle dónde falla este ajuste conviene analizar los residuales "ajustados" propuestos por Haberman (1972) que consisten en dividir los residuales observados entre los estimadores de sus desviaciones estándar asintóticas correspondientes, los cuales tienen una distribución asintótica $N(0,1)$.

En el cuadro 2 se presentan estos residuales; cuatro de ellos están marcados con un asterisco para señalar que son demasiado grandes como para provenir de una normal estándar. Como se puede ver estos residuales se ubican cerca de las esquinas de la tabla.

CUADRO 2

			Edad		
Desnutri.	0	1	2	3	4+
No	*2.15	-0.81	-0.75	-0.31	1.63
Leve	-0.1	-1.15	1.35	-1.59	-0.59
Moderada	-1.21	0.35	-1.22	0.88	-0.94
			Edad		
Desnutri.	0	1	2	3	4+
No	-0.76	0.75	0.45	-0.44	*-2.64
Leve	-0.29	1.02	1.54	-0.75	0.89
Moderada	-0.52	-0.08	-1.01	*2.45	*2.23

El único negativo de los cuatro es -2.64 y corresponde al valor mas bajo de desnutrición y al mas alto de edad en el sexo femenino, los otros tres residuales marcados

son positivos, dos de ellos corresponden a valores altos de las dos variables ordinales y al sexo femenino y el otro corresponde a valores bajos de las dos variables en el otro sexo. Esto parece indicar una importante relación entre las dos variables ordinales y una posible relación entre sexo y una de las variables ordinales.

para seleccionar el "mejor" modelo se utilizará el método de la G^2 particionada que consiste en proponer un conjunto de modelos anidados, tomar los que tienen un ajuste adecuado y compararlos por parejas comenzando por los más complejos; en la primera pareja donde la G^2 particionada resulta significativa se selecciona el modelo mas complejo de la pareja.

Se consideró el siguiente conjunto de los modelos anidados incluyendo al de completa independencia.

a. $logm_{ijk} = \mu + \lambda_i^E + \lambda_j^D + \lambda_k^S$

b. $logm_{ijk} = \mu + \lambda_i^E + \lambda_j^D + \lambda_k^S + \beta^{ED}[u_i - \overline{u}][v_j - \overline{v}]$

c. $logm_{ijk} = \mu + \lambda_i^E + \lambda_j^D + \lambda_k^S + \beta^{ED}[u_i - \overline{u}][v_j - \overline{v}] + \tau_k^{DS}[v_j - \overline{u}]$

d. $logm_{ijk} = \mu + \lambda_i^E + \lambda_j^D + \lambda_k^S + \beta^{ED}[u_i - \overline{u}][v_j - \overline{v}] + \tau_k^{DS}[v_j = \overline{v}] + \tau_k^{ES}[u_i - \overline{u}]$

En el cuadro 3 se presentan las estadísticas G^2 despueés de ajustar cada uno de estos modelos y sus diferencias para la comparación entre pares de modelos.

En el cuadro 3 se observa que al revisar las estadísticas particionadas la única que resulta significativa al 5% es la diferencia entre las de los modelos (b) y (c) por lo que se escoge el modelo (c) como el mejor modelo.

CUADRO 3

	G^2	$g.l.$	p
modelo (a)	33.62	22	.054
modelo (b)	30.06	21	.095
diferencia entre (a) y (b)	3.56	1	.059
modelo (c)	25.71	20	.176
diferencia entre (b) y (c)	4.35	1	.037
modelo (d)	25.70	19	.18
diferencia entre (d) y (c)	0.01	1	.90

En el cuadro 4 se presentan los valores esperados (estimados) al considerar el modelo (c).

Se puede verificar que

$$\hat{\beta}^{ED} = log\hat{\theta}_{ij(k)} = log\frac{(34.69)(2.49)}{(12.50)(6.22)} = \ldots = \frac{(2.31)(5.96)}{(3.86)(3.21)} = 0.1055$$

$$\hat{\tau}_1^{DS} - \hat{\tau}_2^{DS} = log\hat{\theta}_{(i)jk} = log\frac{(34.69)(6.70)}{(25.97)(6.22)} = \ldots + \frac{(2.98)(5.96)}{(3.84)(3.21)} = -0.3648$$

CUADRO 4

			Edad			
Desnutri.	0	1	2	3	4+	total
No	34.69	12.50	9.29	8.70	10.88	77
Leve	6.22	2.49	2.06	2.14	2.98	14
Moderada	5.26	2.34	2.15	2.48	3.84	17
Total	50	14	12	12	20	108
			Edad			
Desnutri.	0	1	2	3	4+	total
No	25.97	9.36	6.95	6.51	8.15	56
Leve	6.70	2.68	2.22	2.31	3.21	19
Moderada	8.16	3.63	3.33	3.86	5.96	24
Total	37	19	14	14	15	99

El modelo escogido finalmente se ajusta sastisfactoriamente a los datos y considera la relación entre edad y sexo con un solo parámetro en lugar de los 8 que se hubieran tenido que estimar si no se hubiera considerado la información ordinal. Además tanto el parámetro β como los τ's tienen una interpretación más simple.

BIBLIOGRAFIA

Agresti, A. 1984 "Analysis of Ordinal Categorical Data", John Wiley & Sons, New York.

Baker, R.J. and Nelder, J.A. 1978 "The GLIM System. Release 3. Generalized Linear Interactive Modelling Manual". Numerical Algorithms Group, Oxford.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W., 1975 "Discrete Multivariate

Analysis: Theory and Practice", Mit Press, Cambridge.

Fienberg, S.E., 1979 "The Analysis of Cross-Classified Categorical Data", Mit Press, Cambridge.

Secretaría de Salud, Subsecretaría de Planeación, "Sistema Estatal de Información Básica, Subsistema de Servicios, 1er nivel de atención" 1987. México, D. F.

DISEÑO DE UNA BANDA AUTOEXTINGUIBLE POR EXPERIMENTACION CON METODO TAGUCHI

Arroyo López, Pilar E. y Hernández Alvarado Ramón
Facultad de Química de la U.A.E.M.
Gates Rubber de México, S.A. de C.V.

RESUMEN

Se desea definir una formulación para emplearse en la construcción de una banda transportadora autoextinguible. Los tiempos de flama y de emisión de gases fueron las variables de respuesta analizaedas. El objetivo principal es tener para los tiempos de flama valores por debajo de los 20 segs. y como objetivo secundario, reducir al mínimo la emisión de gases. Cinco factores fueron investigados asó como dos interacciones de primer orden, el diseño experimental fué asignado sobre un arreglo ortogonal $L_{16}(2^{15})$. La formulación óptima es determinada mediante el ANOVA de los cocientes señal a ruido.

INTRODUCCION

Calidad se ha definido tradicionalmente como adecuación para el uso, y la industria ha tratado de alcanzarla con el propósito de competencia en ventas. Con la revolución de calidad generada en Japón en los años 60's y con la visita de asesores norteamericanos en calidad a este país, Japón se ha convertido en el país líder en calidad. La excelente calidad del producto japonés ha fomentado en occidente el interés por mejorar la producción. Algunas de las ideas innovadoras en Japón fueron propuestas en los años 70's oor el Dr. Genichi Taguchi, cuando trabajaba en Electrical Communication Laboratories. Su trabajo fué conocido en Estados Unidos en 1984, en una conferencia organizada por A.T.& T. Bell Laboratories en Nueva York, donde se discutió sobre el impacto que el diseño experimental tiene sobre productividad y calidad.

La revista Journal of Quality Technology, dedicó su volumen 17, de octubre de 1985, a la Metodología Taguchi, y desde entonces numerosas publicaciones técnicas y prácticas han aparecido, dándose un interés creciente por el empleo de la metodología. La demanda de aplicación en las técnicas llevó a la creación de The American Supplier Institute, en Dearborn, Michigan, el centro en Estados Unidos para la difusión y enseñanza del Método Taguchi. La industria

automotriz es una de las que más ha fomentado el uso de la metodología dada su orientación a costos y mejoras en la variabilidad, contribuyendo a que otro tipo de industrias, proveedoras de materiales automotrices se inclinen también por su uso.

En México, la metodología Taguchi ha sido promovida inicialmente por el ITESM, campus Monterrey, en donde se ha tenido la visita del Dr. Taguchi en dos ocasiones. En enero de 1986 se promovió el primer curso de Método Taguchi en México y en 1987 el Congreso de Calidad, organizado por Mitutoyo (1987) sirvió como foro de presentación de los primeros experimentos Taguchi en nuestro país. En la zona industrial de Toluca, sólo una empresa ha iniciado el empleo de la metodología, el presente trabajo es uno de los experimentos Taguchi efectuados por esta industria en el presente año.

EL PROBLEMA

La empresa Gates Rubber de México fabrica banda transportadora, el hule con que se elabora el producto debe tener propiedades de resistencia al fuego. La tolerancia actual para el material es de 60 seg. y el polímero base en la fórmula es neopreno, la empresa desea sustituir el neopreno por estireno butadieno, producto de fabricación nacional y menor costo. Considerando que el producto actual a base de neopreno tiene tiempos de flama de 5 seg. en promedio, se desea que el nuevo producto no exceda el límite de 20 seg. Además de los tiempos de flama, se tiene como segunda variable importante, el tiempo de emisión de gases también flamables, el cual desea reducirse al mínimo, siendo la tolerancia actual 3 min.

En base al reporte de Culverhouse (1983), y a consultas directas de los proveedores de materiales de la empresa, se seleccionaron cinco factores a investigar, los niveles fueron definidos también en base a la bibliografía y la experiencia del personal en el área de formulaciones. Los factores investigados son:

A: ALUMINA HIDRATADA	A1	A2		
B:TRIOXIDO DE ANTIMONIO	B1=0 PCH*	B2		
C:PARAFINA CLORADA	C1=0 PCH	C2	C3	
D:FOSFATO DE TRICRESILO	D1=0 PCH	D2	D3	
E:GRAFITO	E1=0 PCH	E2		

* PCH = PARTES POR CIEN DE HULE.

Fórmula base: Cantidades fijas de aceite, polímero, cargas reforzantes, antioxidante y aceleradores, más el agente vulcanizante.

Para cuantificar la resistencia al fuego del material, se emplearon las siguientes variables de respuesta:

Tiempo de flama: El tiempo (seg) de duración de la flma en la muestra de hule después que se expone a la flama de un mechero.

Tiempo de emisión de gases: El tiempo (seg) que dura el desprendimiento de humo en la muestra luego de la exposición al fuego.

En adición a los cinco efectos principales, las interacciones AxC y AxD también fueron estimadas. El experimento se asignó sobre un arreglo ortogonal $L_{16}(2^{15})$ de acuerdo a las tablas proporciondas por Taguchi (1986 a). Los factores a tres niveles se asignaron empleando el método de tratamiento "mudo" según lo propuesto en Wu y Moore (1986) y Taguchi (1986 b). El nivel repetido para el factor C fue el tercero y para el D el segundo. La disposición de efectos en las columnas del arreglo se muestra en seguida (e = error experimental):

(1)	(2,4)	(3,5)	(6)	(7)	(8,15)	(9,14)	(10)	(11)	(12)	(13)
A	C	AxC	e	e	D	AxD	E	e	B	e

En el cuadro I se muestran las combinaciones corridas, los datos obtenidos para tres repeticiones a cada combinación y los valores del cociente señal a ruido.

ANALISIS DE RESULTADOS

El concepto en audio de señal respecto a ruido ha sido empleado por el Dr. G. Taguchi para generar expresiones que permitan determinar las condiciones óptimas al experimento al seleccionar el valor máximo del cociente calculado. Para este experimento, donde el objetivo es minimizar la respuesta, el cociente señal a ruido (η), expresado en decibeles (db), se calcula según Wu y Moore (1986) y Taguchi (1986 b) como sigue:

$$\eta = \text{cociente s/r} = -10 \log \Sigma y_i^2/n$$

observemos que al minimizar la respuesta, también se estará minimizando la variabilidad de la misma.

El análisis de varianza (ANOVA) para el cociente s/r de las dos variables de respuesta se muestra a continuación. Para detalles respecto a la tabla, consultar Wu y Moore (1986), Taguchi (1986 b) y Box, Hunter y Hunter (1978). Las gráficas de los efectos significantes se muestran en las Fig. 1 y 2 del artículo.

VARIABLE DE RESPUESTA: TIEMPO DE FLAMA

ANOVA

FUENTE	gl	S	V	F	F(com)	S'	(%)
A	1	19.3421+	19.3421	<1	---	-----	---
B **	1	1226.9745	1226.9745	49.03	61.29	1206.9562	56.5
C **	2	584.587	292.2935	11.68	14.6	544.5504	25.5
D	2	83.363	41.6815	1.67	2.09	43.3264	2.04
E	1	0.6744+	0.6744	<1	---	-----	---
AxC	2	54.224	27.112	1.08	1.35	14.1874	<1
AxD	2	67.0229	33.5114	1.34	1.67	26.9863	1.3
e	4	100.0932	25.0233				
e(com)	6	120.1097	20.0183			300.2744	14.1
TOTAL	15	2136.281					

** = factores altamente significantes

\+ = efectos combinados para estimar error experimental

(com) = abreviatura para indicar que las cantidades fueron evaluadas después de combinar efectos con el error.

De acuerdo a las gráficas, las condiciones óptimas serían la combinación B2C3, el cociente señal a ruido estimado en el óptimo:

$$\hat{\eta} = B2 + C3 - T = -10.46$$

entonces $1/n \Sigma y_i^2 = 11.1173$

lo que indica una variación mínima de la respuesta además de valores bajos para ésta.

VARIABLE DE RESPUESTA: TIEMPO DE EMISION DE GASES

ANOVA

FUENTE	gl	S	V	F	F(com)	S'	(%)
A**	1	48.4138	48.4138	20.5	42.2	47.2667	27.3
B**	1	82.9466	48.4138	35.4	72.3	81.7997	47.3
C	2	8.5243	4.2621	4.16	3.72	6.2299	3.6
D	2	10.4399	5.22	2.2	4.33	8.1455	4.7
E	1	0.3667+	0.3667	<1	----	----	---
AxC*	2	14.67	7.335	7.16	6.39	12.3756	7.2
AxD	2	3.57 +	1.785	1.74	----	----	---
e	4	4.0944	1.0236				
e(com)	7	8.0304	1.1472			17.208	10.0
TOTAL	15	173.025					

Para esta segunda variable de respuesta. resulta significante la interacción de los factores A y C, por tanto, una combinación de ellos será la elegida como óptima, las gráficas de la Fig. 2 permiten determinar estas condiciones.

La combinación que maximiza el cociente s/r es A1B1C2, el cociente s/r será estimado entonces por:

$$\hat{\eta} = T + (B1-T) + (A1C2-A1-C2+T) + T = -28.98$$

entonces $1/n \Sigma y_i^2 = 790.68$

la media estimada en seg. utilizando expresión similar será:

$$\hat{\mu} = B1 + A1C2 - A1 - C2 + T = 27.6251$$

EXPERIMENTOS PARA DISMINUIR EL COSTO

El producto trióxido de antimonio (factor B) es bastante caro, por lo cual la formulación en las condiciones óptimas para tiempo de flama (B2C3) sería de alto precio, además según el análisis de emisión de gases, cantidades altas de este producto incrementan esta segunda variable de respuesta. Dado que los tiempos de flama están muy por debajo del objetivo 20 seg. y la tolerancia 3 min. respectivamente, se decidió realizar otra serie de experimentos, involucrando únicamente los factores A, B y C. Para el factor B se probaron cantidades menores a los niveles en el primer experimento, sin eliminarlo totalmente; para el facotr C, que es un reactivo económico, cantidades superiores a las anteriores, esperando encontrar una relación de B y C que satisfaga los objetivos del problema y sea de bajo costo. Respecto a la alúmina hidratada (factor A), se investigaron niveles inveriores a los previos, dado que la emisión de gases se espera menor cuando el reactivo esté en baja proporción. Los nuevos niveles se dan en seguida, el símbolo (') indica que nos referimos al nivel cuantitativo utilizado en la anterior serie de experimentos.

A: ALUMINA HIDRATADA	A1=0 PCH	A2= A1'/2	A3=A1'	
B:TRIOXIDO DE ANTIMONIO	B1=B2'-15	B2=B2'-5	B3=B2'	
C:PARAFINA CLORADA	C1=C2'	C2=C3'	C3=C3'+10	

Un tercio del factorial 3^3 fué corrido, con tres repeticiones a cada combinación, los resultados se muestran en el cuadro II.

El análisis de varianza para los valores del cociente s/r de las dos variables de respuesta se da a continuación.

VARIABLE DE RESPUESTA: TIEMPO DE FLAMA

ANOVA

FUENTE	gl	S	V	F	F(com)	S'	(%)
A	2	66.9009	33.4505	<1	----	----	---
B	2	250.1187	125.0594	1.58	2.22	137.2907	8.81
C*	2	1082.7305	541.3653	6.82	9.6	969.9025	62.23
e	2	158.755	79.3775				
TOTAL	8	1558.5051					

El factor C es el único declarado significante en este ANOVA y el tercer nivel produce el máximo valor del cociente η. Notando que los tiempos de flama para la primera combinación de los factores son notablemente altos, se efectuó también el ANOVA sobre los valores originales, estimándose parcialmente la interacción AxC, ver John (1981), la cual se esperaba significante desde el inicio de la experimentación. La parte AB de la interacción AxB es alias dela parte AC^2 de AxC y en el arreglo ortogonal, los efectos se estimarían de la 4a. columna del arreglo ortogonal. El ANOVA se muestra en seguida.

VARIABLE DE RESPUESTA: TIEMPO DE FLAMA (SEGUNDOS)

ANOVA

FUENTE	gl	SC	CM	F	
A	2	5872.063	2936.0315		
B=AC	2	6216.0741	3108.0371		
C	2	11596.0741	5798.0371		
e_1=AB=AC^2	2	5766.3073	2883.1537	23.7	**
e_2	18	2190	121.6667		
TOTAL	26	31640.5185			

e_1= error experimental e_2= error de muestreo

Este ANOVA permite observar que los efectos de A, B y la parte AC^2 de la interacción AxC son comporables, las gráficas de la Fig. 3 permiten observar que cuando la cantidad de parafina clorada es baja, se reportan los mayores tiempos de flama, pero estos no son tan considerables como en ausencia de alúmina (A1), cuando el factor C está a niveles más altos, los tiempos mejoran notablemente aún cuando no se tenga alúmina hidratada en la fórmula, por tanto la interacción AxC se considera significante y la combinación A1C1 ha de evitarse. El factor dominante es C, y los tiempos de flama son buenos cuando su cantidad es elevada, independientemente de A y B.

VARIABLE DE RESPUESTA: TIEMPO DE EMISION DE GASES

ANOVA

FUENTE	gl	S	V	F
A	2	4.3996	2.1998	<1
B	2	2.6779	1.3389	<1
C	2	2.7037	1.3519	<1
e	2	9.7812	4.8906	
TOTAL	8	19.1675		

Ningún factor es declarado significante, al reducir la cantidad de alúmina hidratada y de trióxido de antimonio respecto a los niveles de la primera serie de experimentos, se tienen tiempos de emisión de gases por debajo de tres minutos, los que no disminuyen considerablemente a las combinaciones investigadas de los factores en estudio.

Con esta segunda serie de experimentos, se concluye que es posible mantener al mínimo la cantidad de trióxido de antimonio sin aumentar considerablemente los tiempos de flama, la cantidad de alúmina tampoco afecta considerablmente estos valores siempre que se use con proporciones elevadas de parafina clorada, por tanto, la formulación elegida como óptima para las dos variables de respuesta y respecto a costos es A1B1C3.

El valor del cociente s/r estimado en las condiciones óptimas está dado para tiempos de flama por:

$$\hat{\eta} = A1+C3+B1-2T = -11.32 \text{ db}$$

La media estimada en segundos, considerando la interacción AxC, es estimada en menos de cero en las condiciones óptimas.

Para tiempo de emisión de gases, la respuesta en el óptimo para cociente s/r y segundos sería igual a:

$$\hat{\mu} = 60.7037 \qquad Y \qquad \hat{\eta} = -35.56 \text{ db}$$

EXPERIMENTOS CONFIRMATORIOS

Una serie de cinco experimentos confirmatorios fué corrida bajo las condiciones óptimas, obteniéndose los siguientes resultados:

TIEMPO DE FLAMA					TIEMPO DE EMISION DE GASES				
MIN	MAX	PROMEDIO	η	$\hat{\sigma}^2$	MIN	MAX	PROMEDIO	η	$\hat{\sigma}^2$
1	2	1.4	-3.42	0.3	47	68	59.2	-35.51	67.7

Experimentos adicionales con el trióxido de antimonio mostraron que una reducción de más del 1% respecto a la usada en las

condiciones anteriores incrementa los tiempos de flama hasta 60 segs. permitiendo considerar la formulación A1B1C3 de la segunda serie de experimentos como la óptima.

CONCLUSIONES

- El fosfato de tricresilo y el grafito, agregados a la fórmula base no mejoran significativamente la resistencia al fuego del hule a emplear en la banda.
- Es necesario agregar trióxido de antimonio a la fórmula base, pero su proporción puede ser mínima y obtener tiempos de flama bajos, ventajas adicionales a esta baja cantidad son mínimas emisiones de gases y bajo costo.
- La alúmina hidratada en altas proporciones incrementa el tiempo de emisión de gases y su proporción en la fórmula afecta los tiempos de flama sólo si se usa en combinación con baja cantidad de parafina clorada.
- La formulación resultante al vulcanizar proporciona hule de buena resistencia al fuego, los tiempos de flama no exceden los 2 segs. y los de emisión de gases son menores de 2 minutos.
- Experimentación usando el método Taguchi permite determinar reactivos y proporciones a usar en la formulación de hule, tal que se tengan propiedades óptimas para el material. La decisión se basa en métodos estadísticos y no en la tradicional experiencia de años del formulista.

BIBLIOGRAFIA

1.- III Congreso de Metrologia y Control de Calidad. La Tecnologia mas avanzada del mundo a su alcance. Organizado por Mitutoyo. Noviembre 1987.

2.- Culverhouse, D. "Compunding rubber for fire resistance". Rubber World. Vol. 188, No. 1. April, 1983.

3.- Orthogonal Arrays and Linear Graphs. American Supplier Institute, incorporated. Center for Taguchi Methods.

4.- Wu, Yuin & Moore, W. H. Quality Engineering: Product & Process Design Optimization. American Supplier Institute, Inc. Center for Taguchi Methods. Dearborn, MI, 1986.

5.- Taguchi, G. Introduction to Quality Engineering. Asian Productivity Organization. New York, 1986.

6.- Box, G.E.P., Hunter, W.G. & Hunter, J. S. Statistics for Experimenters. John Wiley & Sons. New York, 1978.

7.- John, P. W. M. Statistical Design and Analysis of Experiments. Macmillan Co. New York, 1981.

CUADRO I

DATOS DE LA PRIMERA SERIE DE EXPERIMENTOS

A	B	C	D	E	TIEMPO DE FLAMA			η	TIEMPO DE EMISION			η
11	1	1	1		70	55	110	-38.24	13	26	28	-27.35
1	2	1	2	2	13	10	10	-20.90	36	35	27	-30.35
1	2	3	2	1	3	1	2	- 6.69	34	44	53	-32.94
1	1	3	3	2	19	15	22	-25.52	43	39	40	-32.19
1	1	2	2	2	51	27	47	-32.66	13	21	15	-24.45
1	2	2	3	1	7	12	18	-22.36	35	43	35	-31.56
1	2	3	1	2	2	3	2	- 7.53	43	47	46	-33.13
1	1	3	2	1	110	50	42	-37.37	15	29	28	-27.90
2	1	1	2	1	140	54	110	-40.62	20	56	33	-31.88
2	2	1	3	2	68	66	46	-35.68	92	59	60	-37.14
2	2	3	1	1	2	2	2	- 6.02	68	69	58	-36.28
2	1	3	2	2	30	38	41	-31.28	25	32	39	-30.24
2	1	2	1	2	45	72	47	-34.97	20	28	43	-30.05
2	2	2	2	1	4	8	15	-20.07	67	64	61	-36.13
2	2	3	2	2	5	2	1	-10.00	53	62	54	-35.04
2	1	3	3	1	43	24	24	-30.00	37	34	41	-31.47

CUADRO II

RESULTADOS DE EXPERIMENTOS PARA MINIMIZAR COSTOS

A	B	C	TIEMPO DE FLAMA			η	TIEMPO DE EMISION			η
1	1	1	85	145	95	-40.94	65	60	100	-37.74
1	2	2	2	2	2	- 6.02	43	40	53	-33.19
1	3	3	1	2	2	- 4.77	54	48	48	-33.99
2	1	3	1	1	1	0.00	55	54	49	-34.41
2	2	1	24	26	15	-26.92	41	54	75	-35.32
2	3	2	2	2	1	- 4.77	58	58	54	-35.07
3	1	2	4	7	2	-13.62	48	43	48	-33.33
3	2	3	1	1	1	0.00	54	54	54	-34.65
3	3	1	8	3	12	-18.59	47	42	36	-32.45

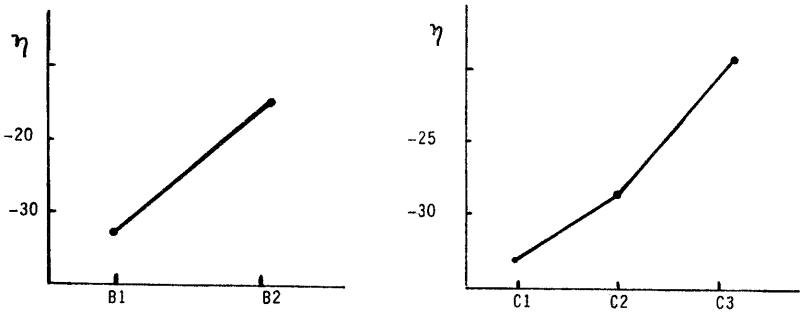


FIGURA 1. GRAFICA DE LOS EFECTOS SIGNIFICANTES
VARIABLE DE RESPUESTA: TIEMPO DE FLAMA

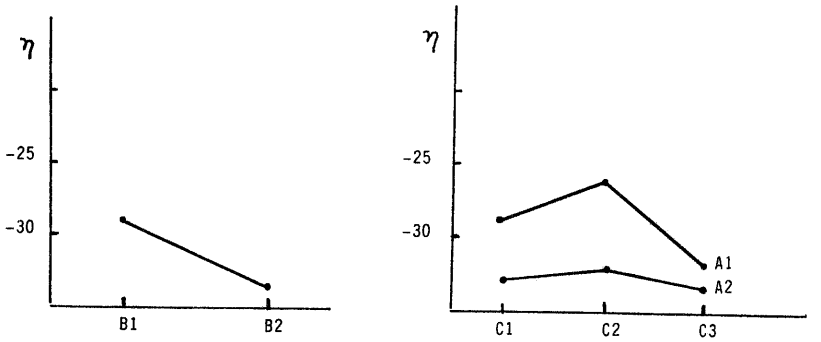


FIGURA 2. GRAFICA DE LOS EFECTOS SIGNIFICANTES
VARIABLE DE RESPUESTA: TIEMPO DE EMISION DE GASES

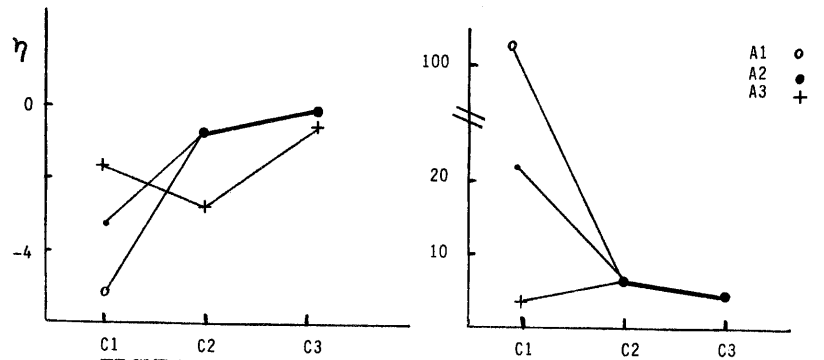


FIGURA 3. GRAFICA DE LOS EFECTOS SIGNIFICANTES
VARIABLE DE RESPUESTA: TIEMPO DE FLAMA

PROCEDIMIENTOS GRAFICOS DE CONTROL DE CALIDAD PARA PROCESOS DE POISSON MULTIVARIADOS

Por Anell V. Daniel G.
Facultad de Estadística (LINAE).
Universidad Veracruzana
Av. Xalapa esq. Avila Camacho s/n.
Xalapa, Ver. Mexico.

RESUMEN

En la práctica del control de calidad industrial se cuanta con algunas técnicas específicas que pueden reflejar el comportamiento de un proceso de producción a partir de una característica. De igual manera, cuando la calidad de un producto se da en p variables, y estas poseen la distribución Normal p variada, se cuenta con técnicas que nos auxilian en la verificación de la misma. En este trabajo se aborda el caso en el que la calidad se da en dos variables teniendo estas distribución Poisson bivariada.

INTRODUCCION.

Cuando se trata de vigilar la calidad de un producto, elaborado en algun proceso de producción, a partir de una sola característica, existe una gran variedad de procedimientos [Grant y Leavenworth (1984)], que pueden reflejar el comportamiento de dicho proceso. La principal limitante de estas herramientas es que consideran la verificación de la calidad sobre una sola característica.

En la práctica del control de la calidad, existen diversas situaciones en donde esta se verifica en mas de una variable sobre el mismo producto. A raíz de esto se hace necesario observar en un mismo grafico el comportamiento de las p características que determinan la calidad; dado que si observamos cada una de ellas por separado, podrían cometerse errores en el momento de decidir, por no considerar la posible correlación que existe entre las variables, lo cual desde luego no se veria reflejado en graficos por separado.

La idea de construir gráficos para controlar procesos multivariados tiene varias orientaciones. Una de estas se basa en la técnica multivariada de componentes principales. Otro enfoque bastante reciente, es debido a Kulkarni y Paranjape [(1984) y (1986)], el cual se basa en las graficos

propuestos por Andrews (1972). En otros artículos [Anell (1987) y (1988)] se ha presentado la metodología y los aspectos computacionales para realizar aplicaciones de gráficos de control en procesos Gaussianos bivariados y trivariados. Aquí siguiendo la línea metodológica, presentamos una extensión de estos procedimientos a procesos de Poisson bivariados basados en los resultados de Hudson, Tucker y Veeh (1986) y algunos resultados de simulación para fundamentar su uso.

DESCRIPCION DEL METODO DE KULKARNI Y PARANJAPE.

Sea $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ la observación multivariada que caracteriza al i-ésimo producto en una muestra de tamaño n, de un proceso que se desea vigilar. Sea $f_i(t) = x_i't$ donde $t' = t'(\theta) = (t_1(\theta), t_2(\theta), \ldots, t_p(\theta))$ un conjunto de funciones trigonométricas ortogonales sobre $[-\pi, \pi]$. Entonces $f_i(t)$ representará una curva para cada x_i; i=1,2,. . . n sobre $[-\pi, \pi]$. Kulkarni y Paranjape (1984) explotan la propiedad de f(t) de preservar una estructura de distancia dependiendo de x. Siguiendo una idea de Gnanadesikan (1977), Kulkarni y Paranjape (1986), sugieren un procedimiento gráfico de control, para el caso trivariado, bajo la siguiente regla: El procedimiento será declarado fuera de control si, y solo si, f(t) no esta completamente contenida en la region

$$R = \{ x \mid li < x_i't < ls, -\pi < \theta < \pi \}$$

donde li y ls son el límite inferior y superior de control respectivamente. Sin embargo, para el caso de mas de dos variables no se tiene una función t que cubra completamente la esfera p-dimensional, creandose así una región de "hoyos" por la que el vector t no pasa [Gnanadesikan (1977)], pudiendo esto hacernos declarar a un proceso dentro de control cuando en realidad no lo esta. A partir de esto, entonces, los mismos autores proponen determinar dos regiones, una basada en t y la otra en t^*, donde t^* es un vector de funciones trigonometricas, ortogonal a t, de tal manera que $t'(\theta)t^*(\theta) = 0$ para todo $\theta \in [-\pi, \pi]$. Así, esta nueva consideracion nos lleva a declarar fuera de control al proceso si alguna $f_i(t)$ no esta completamente contenida en ambas gráficas. Esta es básicamente la idea de Kulkarni y Paranjape (1986). Las pruebas matemáticas para evaluar la eficiencia de este procedimiento no se han reportado, debido a la dificultad que involucran, pero para el caso Gaussiano trivariado,

Kulkarni y Paranjape (1986) proveen estudios de simulación que avalan la eficiencia del metodo.

Es bien sabido que existen situaciones en la práctica donde se cuenta el número de defectos de varios tipos. Si las variables descriptoras de la calidad en el producto, salido del proceso, fuesen estadísticamente independientes, entonces sería posible utilizar algún procedimiento grafico univariado para el control de la calidad por atributos, pero debido a que la estructura de correlación de las variables indica no independencia, dicho conjuto de herramientas resulta inapropiado.

Hudson, Tuker y Veeh (1986), sugieren una aproximación a la Normal Multivariada de la Poisson Multivariada; asi, se antoja posible usar el procedimiento grafico de control de Kulkarni y Paranjape diseñado para procesos Gaussianos con procesos de Poisson, bajo la fundamentación de la aproximación propuesta.

Sin embargo, es menester determinar las condiciones bajo las cuales la aproximación sugerida se logra de buena manera.

En un primer intento se han diseñado estudios de simulacion para el caso bivariado, con el fin de evaluar la potencia y significancia de la aproximación . Asi que el procedimiento grafico de control con el cual se investigara la bondad de tal sugerencia se define de la siguiente manera:

Para el caso bivariado existe una función t que cubre completamente el espacio bidimensional. En este articulo se considerara el primer método sugerido por Kulkarni y Paranjape, dada la propiedad de la funcion $t = (\text{sen } \theta, \cos \theta)'$. Siguiendo los lineamientos para el control grafico multidimensional expuestos en parrafos anteriores se dira que si $f_i(t) = x_i't$ no se encuentra contenida en la region R, el proceso sera declarado fuera de control. Aqui $x_i = (x_{i1}, x_{i2})'$ es el vector de observaciones, con distribución Poisson bivariada [Holgate (1964)], que caracteriza al i-esimo producto en una muestra de tamaño n; en otras palabras, el proceso sera declarado bajo control si cada $f_i(t)$ aparece contenida en

$$\mu' t - C_\alpha \sqrt{t' \Sigma t} \leq f_i(t) \leq \mu' t + C_\alpha \sqrt{t' \Sigma t}$$

donde $\underset{\sim}{\mu}$ es el vector de medias; $\underset{\sim}{\mu} \equiv \begin{bmatrix}0\\0\end{bmatrix}$; $t \equiv (\text{sen } \theta, \cos \theta)'$, con C_α siendo el cuantil (1-α) x 100% de la distribucion ji-cuadrada con dos grados de libertad y Σ la matriz de varianzas y covarianzas .

Los estudios de simulación fueron implementados en el lenguaje Basic. En ellos se consideraron diferentes situaciones por las que se supuso un proceso de produccion podría pasar. El conjunto de valores que tomo el vector que define completamente a la distribucion Poisson bivariada se encontro en el intervalo de uno a cinco, para ambos. Una limitante de la aplicabilidad de la distribucion Poisson, en el caso bivariado, es que la correlacion de las variables no puede exceder a la raiz cuadrada del cociente entre la mas pequeña y la mas grande de las medias [Holgate (1964)]. Esta condicion hizo imposible que se evaluaran algunas situaciones por las que el proceso podria pasar en la practica.

Tres correlaciones fueron utilizadas en los estudios de simulacion, estas fueron ρ=0.25, ρ=0.50 y ρ=0.75. Se emplearon tamaños de muestra desde 5 y hasta 500, con alrededor de 10000 simulaciones.

RESULTADOS

A continuación se presentan algunos de los resultados más relevantes.

Tabla que muestra la significancia empírica con ρ = 0.25 ,(λ_1=2, λ_2=2) y 50 muestras.

Tamaño de muestra	5	6	7	8	9	10
No. de muestras fuera .	1	3	4	7	12	13
No. de muestras fuera esperadas.	13	15	18	20	23	25

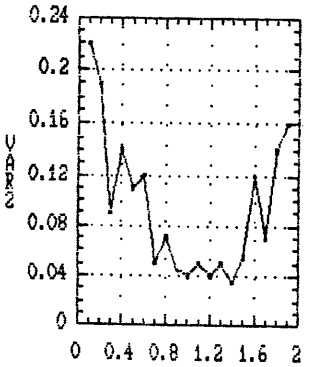
Tabla que muestra la significancia empírica con ρ = 0.50 (λ_1=2, λ_2=4) y 100 muestras.

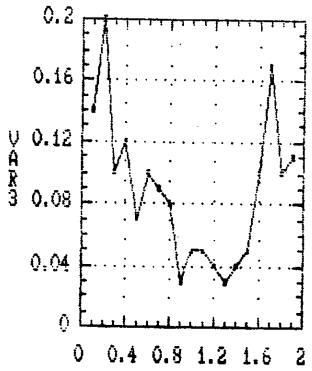
Tamaño de muestra	5	6	7	8	9	10
No. de muestras fuera .	5	4	9	11	23	28
No. de muestras fuera esperadas.	25	30	35	40	45	50

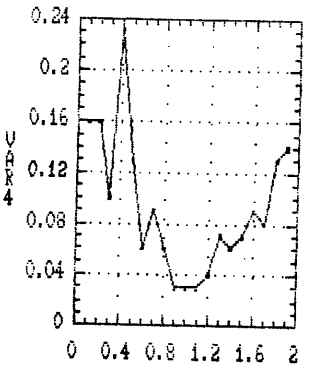
Se observa que el nivel de significancia en general se sobreprotege.

En cuanto a potencia, se presentan algunos gráficos que nos muestran el poder de la aproximación. Para esta parte, se simularon muestras de tamaño 100 y hasta 500, considerando ρ = 0.75. Para evaluar la potencia de la aproximación en algunos casos se varió un parametro, fijando el otro y en otras situaciones, se variaron ambos.

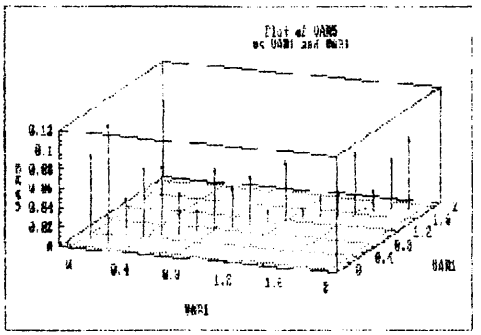
Los siguientes gráficos muestran el caso en que se fija λ_1=1 y se varía λ_2=1 en el intervalo [0, 2], para tres tamaños de muestra, n = 100, n = 200 y n = 500.

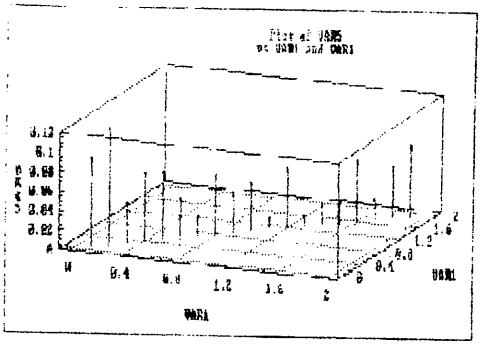


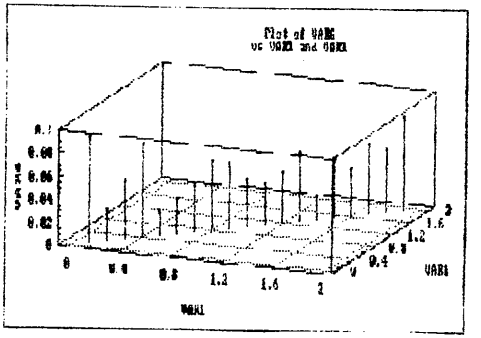




Ahora se presenta uno de los casos en los que se variaron ambos parámetros y donde $\lambda_1 = 1$ y $\lambda_2 = 1$, tambien para los diferentes tamanos de muestra.







CONCLUSIONES

El presente trabajo ha tenido como propósito fundamental difundir la metodología estadística necesaria para la construcción de gráficos de control en el caso multivariado considerando que las características que determinan la calidad tengan distribución Normal y Poisson Multivariadas.

Recientemente en el Congreso de la Sociedad Americana de Control de Calidad (ASQC), celebrado en Dallas Texas, se sostuvo en una ponencia [Constable K.G.] la importancia de realizar investigación en los aspectos prácticos y metodológicos del control de la calidad para el caso multivariado. También se hizo presente la necesidad de implementar esta metodología en procesos de producción cuyos datos no tuvieran distribución Normal p variada. El presente trabajo considera este caso, las características que se han supuesto aquí, poseen una distribución Poisson Multivariada. Con esto se intenta contribuir, en alguna medida, al desarrollo de los procedimientos teórico-prácticos que aporten mejoras al control de calidad industrial. Consideramos que de aquí se pueden seguir proyectos que cubran los siguientes aspectos:

1 Realizar estudios de simulacion mas amplios.
2 Difundir los resultados obtenidos.
3 Promover aplicaciones e implementar paquetes computacionales que faciliten su operabilidad en la practica.
4 Ampliar el rango de accion de estos procesos a otras tecnicas de analisis de datos.

BIBLIOGRAFIA

1.- Andrews D.F. "PLOTS OF HIGH-DIMENSIONAL DATA". Biometrics 28, 125-36. 1972

2.- Anell D.V.G. "PROCEDIMIENTOS PARA LA ELABORACION DE GRAFICOS DE CONTROL PARA PROCESOS GAUSSIANOS Y DE POISSON" . Facultad de Estadística e Informatica, U.V. Trabajo presentado en el XX Congreso Nacional de Matematicas.

3.- Holgate P. "ESTIMATION FOR THE BIVARIATE POISSON DISTRIBUTION". Biometrika 51,1 y 2 p. 241. 1964.

4.- Hudson N.W., Tucker G.H. y Veeh A.J. "LIMIT THEOREMS FOR THE MULTIVARIATE BINOMIAL DISTRIBUTION". Annals of Multivariate Analysis. 18, 32-45. 1986.

5.- Johnson N. L. and Kotz S. "DISTRIBUTIOS IN STATISTICS". Volumen II, Discrete Distributions. Wiley 1970.

6.- Kulkarni S. R. and Paranjape S.R. "AN IMPROVED GRAPHICAL PROCEDURE FOR MULTIVARIATE QUALITY CONTROL". Comm. Statist.- Simula, 15(1), 135-146. 1986.

7.- Marshall W.A. and Olkin I. "A FAMILY OF BIVARIATE DISTRIBUTIONS GENERATED BY THE BIVARIATE BERNOULLI DISTRIBUTION". Journal of the American Statistical Association. 80, 390. 1985.

8.- Ojeda M.M., Ortíz C.M., Armas G. "GRAFICOS DE CONTROL PARA PROCESOS GAUSSIANOS MULTIVARIADOS". Facultad de Estadística e Informatica, U.V. Artículo Inedito. 1986.

9.- Constable K. G. "MULTIVARIATE CHARTS- WHEN AND WHEN NOT USE". ASQC Quality Congress Transaction- Dallas 1988.

UNA SOLUCION BASADA EN MODELOS ARIMA PARA EL PROBLEMA DE LA DESAGREGACION TEMPORAL DE SERIES

Guerrero, V.M.
Banco de México, Av. Juárez 90, México 06040, D.F.

RESUMEN

Muchas series de tiempo económicas se encuentran disponibles sólo en forma agregada temporalmente (por ejemplo, anualmente). Cuando el análisis indica que se debe trabajar con datos desagregados (por ejemplo, mensual o trimestralmente) el analista se topa con el problema de tener que derivar estos datos en forma razonable. Para utilizar los métodos propuestos con anterioridad, se debe elegir con sensatez entre la suavidad esperada de la serie desagregada, la estructura estocástica del término de error implícito o algún otro aspecto que sustente al método respectivo.

En este trabajo se desarrolla un método que no depende de la opinión subjetiva del analista, sino de los datos mismos y que produce un estimador óptimo de la serie desagregada. Para usarlo se requiere de una estimación preliminar de la serie, la cual después se ajusta para cumplir con las restricciones impuestas por los datos agregados.

1. INTRODUCCION

Se han publicado muchos trabajos que versan sobre el problema de la desagregación temporal de series, visto desde diferentes perspectivas y llamado de distintas formas (veáse al respecto los artículos de la bibliografía). Este artículo presenta un nuevo método para desagregar una serie de tiempo, con el cual se combina óptimamente la información provista por los datos agregados con estimaciones preliminares de los valores desagregados. Este planteamiento considera a la serie desagregada como valores estimados de una serie de tiempo no observada. El criterio de optimización empleado consiste en minimizar la varianza condicional generalizada del error de estimación.

El método basado en modelos ARIMA que aquí se propone, conduce a usar una estimación preliminar (la cual se puede obtener, por ejemplo, a partir de un modelo de regresión, que use variables relacionadas y observadas en forma desagregada) combinada linealmente con un elemento que sirve para cancelar las discrepancias entre los datos agregados y los desagregados. De hecho, la estimación resultante tiene, en esencia, la misma forma que la de Chow y Lin (1971) o la de Denton (1971). Además, este método concuerda con las propuestas de Harvey y Pierse (1984) y Nijman y Palm (1986) en el sentido de que utiliza una representación de modelo ARIMA. No obstante, la característica más importante del

método sugerido es su objetividad basada en los datos, la cual evita el insumo subjetivo del analista con respecto a la uniformidad esperada o la estructura de autocorrelación. Aun cuando se apoya en el supuesto de que se cuenta con una estimación preliminar de la serie.

2. UN METODO BASADO EN MODELOS ARIMA

Sea $\{Z_t\}$ una serie de tiempo no observada que se va a estimar (desagregar) para los periodos $t=1, \ldots, mn$, en donde $n \geq 1$ es el número del total de años y $m \geq 2$ denota la frecuencia dentro del año (es decir, $m=4$ para una serie trimestral y $m=12$ para una serie mensual).

La fuente de información principal sobre $\{Z_t\}$ la constituye una serie anual $\{Y_i\}$, observada en los momentos $i=1\ldots, n$, la cual es una combinación lineal de las Z's dentro de los años. O sea, si $\mathbf{Z}_{m(i-1,i)}=(Z_{m(i-1)+1},\ldots,Z_{mi})'$ y $\mathbf{c}=(c_1,\ldots,c_m)'$ son vectores columna, con $c_1,\ldots,c_m$ algunas constantes conocidas (no todas iguales a cero) entonces

$$Y_i=\mathbf{c}'\mathbf{Z}_{m(i-1,i)} \text{ para } i=1,\ldots, n \qquad (2.1)$$

Algunas de las formas de $\mathbf{c}$ que se encuentran con mayor frecuencia son

$$\mathbf{c} = \begin{cases} (1,0, \ldots, 0,0)' \\ (0,0, \ldots, 0,1)' \\ (1,1, \ldots, 1,1)' \\ \frac{1}{m}\ (1,1, \ldots, 1,1)' \end{cases}$$

donde las dos primeras están asociadas con el llamado problema de interpolación y las otras dos con el de distribución de una serie de tiempo.

2.1 Una solución teórica óptima

Ahora supóngase que $\{Z_t\}$ admite una representación ARIMA del tipo

$$\phi(B)d(B)Z_t = \tau(B)a_{Z,t} \qquad (2.2)$$

en donde $\{a_{Z,t}\}$ es un proceso de ruido blanco Gaussiano con media cero y varianza σ_Z^2. B es el operador de retraso tal que $BZ_t=Z_{t-1}$, mientras que $\phi(B)$, $d(B)$ y $\tau(B)$ representan a los operadores: autorregresivo, de diferencias y de promedios móviles, respectivamente. Es sabido (véase Guerrero, 1983, cap. 6) que el estimador lineal con error cuadrático medio mínimo (ECMM) de Z_t, basado en información hasta el período $0<t$, lo proporciona la esperanza condicional

$$E(Z_t|Z_0,Z_{-1},\ldots) = \underset{0}{E}(Z_t) \qquad (2.3)$$

El error de pronóstico, en términos de los coeficientes de la representación de promedios móviles pura, viene a ser entonces

$$Z_t - \underset{0}{E}(Z_t) = \sum_{j=0}^{t-1} \theta_j a_{Z,t-j} \quad \text{con } \theta_0 \equiv 1 \qquad (2.4)$$

en donde los valores $\theta_1,\theta_2,\ldots$ se obtienen al igualar los coeficientes de las potencias de B en la ecuación $\theta(B)\phi(B)d(B)\tau^{-1}(B)=1$. La expresión (2.4) es válida para $t=1,\ldots,mn$ y se obtiene a partir del Teorema de la Descomposición de Wold para el caso en que $\{Z_t\}$ sea estacionaria y a partir del Teorema 1 de Bell (1984) cuando la serie es no-estacionaria.

Para derivar el método, es conveniente suponer (pretender) que $\theta_1,\theta_2,\ldots$ y σ_{Z^2} son conocidos. Así, al escribir $\mathbf{Z}=(Z_1,\ldots,Z_{mn})^1$ resulta que

$$\mathbf{Z} - \underset{0}{E}(\mathbf{Z}) = \theta\mathbf{a}_Z \qquad (2.5)$$

donde θ es la matriz triangular inferior, de dimensión mn x mn, formada por la sucesión de ponderaciones 1, $\theta_1,\ldots,$ θ_{mn-1} en la primera columna, las ponderaciones 0,1, $\theta_1,\ldots,$ θ_{mn-2} en la segunda columna, etc. Mientras que el vector aleatorio $\mathbf{a}_Z=(a_{Z,1},\ldots, a_{Z,mn})'$ es tal que $E(\mathbf{a}_Z)=\mathbf{0}$ y $E(\mathbf{a}_Z\mathbf{a}_Z')=\sigma_{Z^2}I$.

Para tomar en consideración las restricciones impuestas por $\{Y_i\}$, defínase $C=I\otimes\mathbf{c}'$ en donde $\otimes$ denota el producto Kronecker y $\mathbf{Y} = (Y_1,\ldots, Y_n)'$, entonces el grupo completo de restricciones anuales (2.1) se puede escribir

$$\mathbf{Y}=C\mathbf{Z}=C\underset{0}{E}(\mathbf{Z})+C\theta\mathbf{a}_Z \qquad (2.6)$$

Por lo tanto, la solución al problema de la desagregación de $\{Z_t\}$ se obtiene con el siguiente:

Teorema 1.- El estimador lineal con ECMM de $\mathbf{Z}$ que implica a $\mathbf{Y}$ como en (2.6) es

$$\hat{\mathbf{Z}} = \underset{0}{E}(\mathbf{Z})+\hat{A}[\mathbf{Y}-C\underset{0}{E}(\mathbf{Z})] \qquad (2.7)$$

donde

$$\hat{A} = \theta\theta'C'(C\theta\theta'C')^{-1} \qquad (2.8)$$

y

$$E_0[(\hat{\mathbf{Z}}-\mathbf{Z})(\hat{\mathbf{Z}}-\mathbf{Z})'] = \sigma_Z^2(I-\hat{A}C)\theta\theta'. \qquad (2.9)$$

Demostración: Todo estimador lineal de $\mathbf{Z}$ es de la forma

$$\tilde{\mathbf{Z}} = A\mathbf{Y} = AC[E_0(\mathbf{Z})+\theta a_Z] \qquad (2.10)$$

donde A es una matriz mn x n de constantes. Para que resulte insesgado, es necesario que

$$\mathbf{0} = E_0(\tilde{\mathbf{Z}}-\mathbf{Z}) = (AC-I)E_0(\mathbf{Z}) \qquad (2.11)$$

así que se debe tener $\tilde{\mathbf{Z}}-\mathbf{Z} = (AC-I)\theta a_Z$ y

$$E_0[(\tilde{\mathbf{Z}}-\mathbf{Z})(\tilde{\mathbf{Z}}-\mathbf{Z})'] = \sigma_Z^2(I-AC)\theta\theta'(I-AC)' \qquad (2.12)$$

Por consiguiente, para que $\tilde{\mathbf{Z}}$ tenga ECMM se debe escoger la matriz A que minimice la varianza generalizada, $Var(\tilde{\mathbf{Z}}-\mathbf{Z})$, definida como la traza de (2.12). Por lo tanto, se debe resolver la ecuación (condición de primer orden) $\mathbf{0} = dVar(\tilde{\mathbf{Z}}-\mathbf{Z})/dA|_{A=\hat{A}}$ y suponer que se cumplen las condiciones de segundo orden para un mínimo. Esto conduce a obtener (2.8). Luego a partir de (2.10) y (2.11) se deduce que

$$\hat{\mathbf{Z}} = E_0(\mathbf{Z})+\hat{A}C\theta a_Z \qquad (2.13)$$

que por (2.6) produce (2.7). Finalmente, la expresión (2.9) se obtiene insertando $\hat{A}$ en (2.12).

El resultado que da el teorema 1 es útil una vez que se tienen suficientes observaciones desagregadas, ya sea al principio o al final de la serie, pues con ello se podrían estimar los valores de θ_1, θ_2, ... y σ_Z^2. En este sentido, el teorema 1 corresponde a la solución propuesta por Harvey y Pierse (1984). Sin embargo, por lo general, en la práctica no se cuenta con esas observaciones y esto limita la aplicabilidad del resultado, motivo por el cual se le considera una solución puramente teórica.

2.2 Una solución factible de aplicarse en la práctica

En aplicaciones prácticas, es razonable suponer que existe una estimación preliminar $\mathbf{W} = (W_1,\ldots,W_{mn})'$, de modo que se puedan emplear estos valores suponiendo que $\{Z_t\}$ y $\{W_t\}$ tienen esencialmente la misma estructura de autocorrelación (y que por ello admiten el mismo modelo ARIMA), excepto por el hecho de que los valores agregados de $\{Z_t\}$ satisfacen (2.1), en tanto que los de $\{W_t\}$ no. En estas condiciones supóngase que

$$\text{i) } E(Z_t|\mathbf{W})=W_t \text{ y ii) } \underset{0}{E(\mathbf{Z})} \text{ es independiente de } \mathbf{W} \qquad (2.14)$$

así que, al tomar esperanza condicional de (2.5), se tiene

$$\mathbf{W}-\underset{0}{E(\mathbf{Z})}=E(\mathbf{Z}|\mathbf{W})-\underset{0}{E[E(\mathbf{Z})|\mathbf{W}]}=\theta E(\mathbf{a}_Z|\mathbf{W}) \qquad (2.15)$$

expresión que al sustraerse de (2.5) conduce a

$$\mathbf{Z} - \mathbf{W} = \theta\mathbf{a} \qquad (2.16)$$

donde $\mathbf{a}=\mathbf{a}_Z-E(\mathbf{a}_Z|\mathbf{W})$ es un vector aleatorio tal que

$$E(\mathbf{a}|\mathbf{W})=\mathbf{0} \quad \text{y} \quad E(\mathbf{aa}'|\mathbf{W})=\sigma^2 P \qquad (2.17)$$

Además como se supone que

$$\phi(B)d(B)W_t = \tau(B)a_{W,t} \qquad (2.18)$$

donde $\{a_{W,t}\}$ es un proceso de ruido blanco con media cero y varianza σ_W^2, resulta que θ_1, θ_2,... y σ_W^2 se pueden estimar al construir un modelo ARIMA para la serie observada $\{W_t\}$ (en Guerrero, 1983, se muestra detalladamente la construcción de este tipo de modelos). Entonces, siguiendo el mismo razonamiento básico de la prueba del teorema 1 y advirtiendo que (2.16) asume ahora el papel de (2.5), se obtiene el siguiente resultado:

Teorema 2.- El MELI de $\mathbf{Z}$, dado que se conoce $\mathbf{W}$ y que $\mathbf{Y} = C\mathbf{Z}$, es

$$\hat{\mathbf{Z}} = \mathbf{W} + \hat{A}(\mathbf{Y}-C\mathbf{W}) \qquad (2.19)$$

con

$$\hat{A}=\theta P\theta'C'(C\theta P\theta'C')^{-1} \qquad (2.20)$$

y donde

$$\mathrm{Cov}[(\hat{\mathbf{Z}}-\mathbf{Z})|\mathbf{W}] = \sigma^2(\mathbf{I}-\hat{\mathbf{A}}\mathbf{C})\Theta\mathbf{P}\Theta'. \qquad (2.21)$$

En la expresión (2.19) se aprecia que $\hat{\mathbf{Z}}$ toma en consideración la estimación preliminar y la corrige de conformidad con las discrepancias que existan entre las dos series anuales observadas: $\mathbf{Y}$ y $\mathbf{CW}$. Por otra parte, se requiere conocer $\Theta\mathbf{P}\Theta'$ para calcular $\hat{\mathbf{Z}}$, y σ^2 para obtener su matriz de varianza-covarianza. En particular, un caso especialmente sencillo ocurre cuando P=I, ya que entonces las expresiones (2.20) y (2.21) se convierten en

$$\hat{\mathbf{A}}=\Theta\Theta'\mathbf{C}'(\mathbf{C}\Theta\Theta'\mathbf{C}')^{-1} \qquad (2.22)$$

y

$$\mathrm{Cov}[(\hat{\mathbf{Z}}-\mathbf{Z})|\mathbf{W}]=\sigma^2(\mathbf{I}-\hat{\mathbf{A}}\mathbf{C})\Theta\Theta' \qquad (2.23)$$

además, en este caso una estimación de σ^2 se obtiene de

$$(mn-r)\hat{\sigma}^2=\hat{\mathbf{a}}'\hat{\mathbf{a}}=(\hat{\mathbf{Z}}-\mathbf{W})'(\Theta\Theta')^{-1}(\hat{\mathbf{Z}}-\mathbf{W}) \qquad (2.24)$$

donde r es el número de parámetros estimados del modelo ARIMA implícito en (2.18).

Ahora bien, por lo que toca al caso general P≠I, con P desconocida, se sugiere a continuación un procedimiento de estimación en dos etapas, similar al que se emplea con el método de Mínimos Cuadrados Generalizados Estimados (véase Judge, Griffiths, Hill y Lee, 1980). En principio obsérvese que (2.16) conduce a

$$\mathbf{e}=\mathbf{Q}\mathbf{a}=\mathbf{Q}\Theta^{-1}(\mathbf{Z}-\mathbf{W})=\Omega^{-1}(\mathbf{Z}-\mathbf{W}) \qquad (2.25)$$

con la matriz no-singular Q definida de tal manera que QPQ'=I. De (2.17) se sigue entonces que

$$E(\mathbf{e}|\mathbf{W})=\mathbf{0} \quad \text{y} \quad E(\mathbf{e}\mathbf{e}'|\mathbf{W})=\sigma^2\mathbf{I} \qquad (2.26)$$

Estas últimas relaciones son útiles para estimar $\Theta\mathbf{P}\Theta'$ y σ^2, y sugieren el siguiente procedimiento de dos etapas:

1) Supóngase tentativamente que P=I y calcúlense entonces las expresiones (2.22) y (2.19). Constrúyase la serie $(\hat{a}_t)$ a partir de $\hat{\mathbf{a}}=\Theta^{-1}(\hat{\mathbf{Z}}-\mathbf{W})$. i) Si esta serie resulta ser ruido blanco, entonces se está de hecho en el caso P=I, para el cual son aplicables las expresiones (2.23) y (2.24). ii) Si dicha serie no es ruido blanco entonces constrúyase un modelo ARIMA para $(\hat{Z}_t-W_t)$ y obténgase su representación MA pura,

con lo cual se tendrá una estimación de la matriz Ω que aparece en (2.25).

2) Hágase uso de la relación

$$\Omega\Omega' = \theta(Q'Q)^{-1}\theta' = \theta P \theta' \qquad (2.27)$$

para calcular la expresión (2.20) y a partir de ella la expresión (2.19). Estímese σ^2 mediante

$$(mn-r)\hat{\sigma}^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}} = (\mathbf{Z}-\mathbf{W})'(\theta P \theta')^{-1}(\mathbf{Z}-\mathbf{W}) \qquad (2.28)$$

con r el número de parámetros estimados del modelo ARIMA implícito en la estimación de Ω y calcúlese la expresión (2.21).

3. RELACIONES CON OTROS METODOS

Dentro de los procedimientos que se utilizan con más frecuencia para desagregar una serie de tiempo en la actualidad, se encuentran el de Denton (1971) y el de Chow y Lin (1971). De modo que resulta lógico comparar el método propuesto con estos dos.

El procedimiento de Denton resulta como solución al problema de optimización:

$$\min_{\tilde{\mathbf{Z}}_D} \{(\tilde{\mathbf{Z}}_D-\mathbf{W})'G(\tilde{\mathbf{Z}}_D-\mathbf{W})\} \text{ sujeto a } \mathbf{Y}=C\tilde{\mathbf{Z}}_D \qquad (3.1)$$

en donde $\mathbf{W}$ se toma como un vector de valores originales (preliminares) de la serie y G es una matriz mn x mn de constantes, definida por $G = D'D$, donde D es una matriz cuadrada y especificada de conformidad con la función de penalización que seleccione el analista. Por ejemplo, cuando la función de penalización está basada en las diferencias que existen entre las estimaciones preliminar y final, se tendrá

$$p(\tilde{\mathbf{Z}}_D,\mathbf{W}) = \sum_{t=1}^{mn} [(1-B)(\tilde{Z}_t - W_t)]^2 \qquad (3.2)$$

con B el operador de retraso; entonces D se convierte en una matriz triangular inferior, con 1's en la diagonal principal, -1's en la subdiagonal inmediata inferior a la principal y ceros en todo lo demás.

La solución al problema (3.1) que se obtiene con la minimización Lagrangiana da como resultado

$$\hat{\mathbf{Z}}_D = \mathbf{W}+G^{-1}C'(CG^{-1}C')^{-1}(\mathbf{Y}-C\mathbf{W}) \qquad (3.3)$$

Entonces, es relativamente fácil ver la forma en que el procedimiento de Denton encaja dentro del método basado en el modelo ARIMA que aquí se propone: simplemente tómese $G^{-1}=\theta P\theta'$ y se tiene (3.3) como caso especial de (2.19). Sin embargo, obsérvese que G, como lo sugiere Denton, está especificada en forma subjetiva y no depende necesariamente de la estructura estocástica de alguna serie observada, contrario al método propuesto en este artículo. Además, ahora se puede medir la precisión de $\mathbf{Z}$, dada la estimación preliminar, mediante la expresión (2.21).

También cabe hacer notar que la solución basada en modelos ARIMA planteada en este trabajo, se pudo haber obtenido como solución de un problema de optimización semejante al (3.1). Esto se debe a que (2.25) implica

$$\mathbf{e}'\mathbf{e}= (\mathbf{Z}-\mathbf{W})'(\Omega\Omega')^{-1}(\mathbf{Z}-\mathbf{W}) \qquad (3.4)$$

de modo que $G=(\Omega\Omega')^{-1}$. Así pues, la función de penalización implícita en el método basado en modelos ARIMA, se convierte en

$$p(\hat{\mathbf{Z}},\mathbf{W})= \sum_{t=1}^{mn} [\Omega^{-1}(B)(\hat{Z}_t-W_t)]^2 \qquad (3.5)$$

donde $\Omega^{-1}(B) = (1 + \Omega_1 B + \Omega_2 B^2 + \ldots)^{-1}$.

Por otro lado, el procedimiento de Chow y Lin (1971) se deriva de la suposición de que $\mathbf{Z}$ se puede representar mediante un modelo de regresión lineal de la forma

$$\mathbf{Z} = X\beta+\mathbf{u} \qquad (3.6)$$

donde X es la matriz de observaciones de las k variables relacionadas con $\mathbf{Z}$ y $\mathbf{u}$ es un vector aleatorio que satisface $E(\mathbf{u}) = 0$ y $Cov(\mathbf{u}) = V$. A partir de aquí, Chow y Lin obtuvieron un resultado, semejante al del teorema 2, para obtener el MELI de $\mathbf{Z}$ y β:

$$\hat{\mathbf{Z}}_{Ch-L} = X\hat{\beta}+VC'(CVC')^{-1}(\mathbf{Y}-C\hat{X\beta}) \qquad (3.7)$$

y

$$\hat{\beta} = [X'C'(CVC')^{-1}CX]^{-1}X'C'(CVC')^{-1}\mathbf{Y} \qquad (3.8)$$

con

$$\mathrm{Cov}(\hat{\mathbf{z}}_{Ch-L}-\mathbf{z}) = [X-VC'(CVC')^{-1}CX][X'C'(CVC')^{-1}CX]^{-1}$$

$$[X'-X'C'(CVC')^{-1}CV] + [V-VC'(CVC')^{-1}CV] \qquad (3.9)$$

Al comparar (3.7) con (2.19), se observa que $\hat{\mathbf{z}}=\hat{\mathbf{z}}_{Ch-L}$ cuando $\mathbf{W}=X\hat{\boldsymbol{\beta}}$ y $\sigma^2\Theta P\Theta'=V$, lo que significa que la estimación preliminar se obtiene a partir del modelo de regresión y que se acertó en la especificación de V. Por ello se puede considerar que el procedimiento de Chow y Lin está estrechamente relacionado con el método que aquí se presenta. No obstante, cabe señalar que la matriz V, que da la estructura estocástica en el método de Chow-Lin, está relacionada con el vector de errores $\mathbf{u}$ (no observado). Así pues, el analista debe postularla subjetivamente, mientras que en el método sugerido, la matriz se deriva de las estimaciones preliminares (observadas) de la serie. Otro punto a notar es que la matriz de covarianza (3.9) no es comparable con (2.21), puesto que en la presente propuesta la matriz de covarianza se obtiene al condicionar en los valores observados de la estimación preliminar $\{W_t\}$. En realidad, cuando $V=\sigma^2\Theta P\Theta'$ y $\hat{A}=\Theta P\Theta'C'(C\Theta P\Theta'C')^{-1}$, (3.9) se convierte en

$$\mathrm{Cov}(\hat{\mathbf{z}}_{Ch-L}-\mathbf{z}) = \sigma^2(I-\hat{A}C)X[X'C'(C\Theta P\Theta'C')^{-1}CX]^{-1}X'(I-\hat{A}C)'$$

$$+ \sigma^2(I-\hat{A}C)\Theta P\Theta' \qquad (3.10)$$

aquí se pueden apreciar claramente dos componentes aditivos de varianza, el segundo de los cuales corresponde a la covarianza de la esperanza condicional de las desviaciones de $\hat{\mathbf{z}}$ respecto a $\mathbf{z}$ (ver 2.21).

4. CONCLUSIONES

En este trabajo se presentaron dos teoremas que proporcionan soluciones teóricamente aceptables al problema de la desagregación temporal de series. Sin embargo, el método basado en modelos ARIMA que se propone en este trabajo, para usarse en la práctica se centra en la existencia de una estimación preliminar de la serie que debe desagregarse. A partir de dicha estimación preliminar, se puede construir un modelo ARIMA cuya estructura, así como la del modelo para la serie de diferencias entre series desagregada y serie preliminar se explota de varias maneras: (i) para considerar las restricciones impuestas por los datos agregados; (ii) para obtener una expresión de la matriz de varianza-covarianza de los errores de desagregación y (iii) para derivar una estadística de prueba que permita juzgar la adecuación de la estimación preliminar. Existe otra manera de explotar el modelo ARIMA de la serie preliminar y ésta es el uso más común de tales modelos: para obtener pronósticos irrestrictos de la serie

desagregada. Así, aunque sólo se mencionaron explícitamente los problemas de interpolación y de distribución, el problema de extrapolación también puede resolverse con mucha naturalidad con este enfoque.

La razón principal para desarrollar este procedimiento fue la de contar con un método objetivo (basado en los datos) que redujera las elecciones discrecionales del analista para su aplicación. Esto se logró aportando un marco unificado dentro del contexto de los modelos ARIMA. Se apreció que el método así obtenido está relacionado estrechamente con dos de los métodos hasta ahora conocidos y a los que se recurre con mucha frecuencia en la práctica

BIBLIOGRAFIA

Bell, W. (1984) "Signal Extraction for Nonstationary Time Series", The Annals of Statistics 12, 646-664.

Chow, G.C. y Lin, A. (1971) "Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series". Review of Economics and Statistics 53, 372-375.

Denton, F.T. (1971) "Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization". Journal of the American Statistical Association 66, 99-102.

Guerrero, V.M. (1983) Análisis Estadístico de Series de Tiempo Económicas. Libro no publicado, disponible en fotocopia.

Judge, G G., Griffiths, W.E., Hill, R.C. y Lee, T.C. (1980). The Theory and Practice of Econometrics. John Wiley and Sons.

Harvey, A.C. y Pierse, R.G. (1984) "Estimating Missing Observations in Economic Time Series". Journal of the American Statistical Association 79, 125-131.

Nijman, T.E. y Palm, F.C. (1986) "The Construction and Use of Approximations for Missing Quarterly Observations: A Model-Based Approach". Journal of Business and Economic Statistics 4, 47-58.

*mzoo

BASES TEORICAS PARA LA INFERENCIA EN MODELOS SEMI-PARAMETRICOS A PARTIR DE N MOMENTOS CONDICIONALES

Sabau, H. C.
CIDE* / ITAM

RESUMEN

En este trabajo desarrollamos bases para una teoría de la inferencia con series de tiempo en una clase muy general de modelos condicionales semi-paramétricos. Mediante un proceso de condicionamiento sobre la información extraída de los momentos de orden menor que r, la información se extrae del momento r utilizando condiciones de ortogonalidad que incorporan criterios de eficiencia asintótica, definiendo así estimadores del método generalizado de momentos. La información adicional obtenida sobre parámetros comunes se incorpora mediante medias ponderadas matriciales. Lo anterior permite definir una estrategia secuencial de modelística en la que cada momento se puede tratar en forma virtualmente independiente de los demás. La necesidad de utilizar residuos de la estimación de la media condicional para la estimación de los parámetros de los momentos de orden superior genera condiciones cualitativamente diferentes para los casos en que la distribución es simétrica y aquéllas en que no lo es. Esta diferencia cualitativa es analizada y se derivan sus implicaciones sobre la teoría asintótica presentada.

§1.- INTRODUCCION

Una muy alta proporción del trabajo econométrico aplicado se basa de una forma más o menos explícita en un supuesto de normalidad. Así, el esfuerzo del investigador se concentra en modelar los dos primeros momentos de la distribución ya que los de orden superior quedan automáticamente determinados. Más aún, la atención se ha concentrado fuertemente en el análisis del primer momento bajo un supuesto artificial de constancia del segundo (homoscedasticidad), y cuya violación se presenta como un "problema", soluble mediante el uso de algunas técnicas convencionales. Aún los más recientes textos de econometría son ejemplos claros de este enfoque (Amemiya [1985], Spanos [1986]).

En los últimos años se ha cuestionado este acercamiento a la inferencia por su extrema simplicidad, si bien se han reconocido sus múltiples logros empíricos. En relación al segundo momento, el trabajo iniciado por Engle [1982] con el modelo ARCH ha abierto una nueva perspectiva para la interpretación de la varianza condicional y

* Domicilio: CIDE, A.C., División de Economía, Carretera México-Toluca Km. 16.5, Apartado Postal 10-883, Col. Lomas de Santa Fe, Delegación Alvaro Obregón, 01210 México D. F.

ha mostrado que se pueden obtener ganancias importantes en eficiencia en la estimación de los parámetros del primer momento si se utiliza la información que acerca de los mismos puede estar contenida en la varianza condicional del proceso. Por lo que respecta al tercero y cuarto momentos, el trabajo de White y MacDonald [1980], Jarque y Bera [1980] y Bera y Jarque [1982] ha producido nuevas bases para su diagnóstico, pero hay pocas propuestas constructivas para su tratamiento. Gallant y Nychka [1987] y Gallant y Tauchen [1986] han propuesto el uso de expansiones hermíticas de la función de densidad, que permiten disponer efectivamente del supuesto de normalidad a cambio de una teoría de la inferencia sustancialmente más compleja. En el presente trabajo buscamos sentar las bases para una teoría de la inferencia que, siendo en cobertura de la misma generalidad que la propuesta por Gallant y sus co-autores, mantenga la simplicidad de un conjunto de ecuaciones en el contexto del modelo clásico. Para lograr esto extendemos el enfoque en Sabau [1987], presentamos a cada momento como una ecuación, y extraemos la información de cada momento utilizando condiciones de ortogonalidad siguiendo el método generalizado de momentos (MGM, Hansen [1982]).

En la sección §2 se presentan los criterios para la parametrización del modelo y la diagonalización de la matriz de covarianzas, condicionando así las ecuaciones. En la sección §3 se presenta la teoría de la estimación, ignorando la no observabilidad de las innovaciones en la media al estimar los momentos de orden dos y superiores. En la sección §4 se utilizan residuos de la estimación de la media para la estimación factible de los demás momentos, generándose diferencias cualitativas entre problemas simétricos y asimétricos. Finalmente, se presentan conclusiones en la sección §5. Por razones de espacio, las demostraciones de teoremas se presentan en un anexo.

§2.- PARAMETRIZACION Y DIAGONALIZACION DEL SISTEMA

La clase de modelos condicionales semi-paramétricos que consideramos está definida por

$$y_t \mid \mathcal{F}_t \sim \mathcal{D}[\mu_t, h_t^{(2)}, h_t^{(3)}, \ldots, h_t^{(N)}], \qquad (1)$$

donde y_t es la variable estudiada; $\mathfrak{F}_t$ es el conjunto de información condicionante (σ-álgebra) que incluye el pasado de y_t y el presente y pasado de otras variables relevantes para la explicación de y_t; $\mathfrak{D}$ es una distribución de probabilidad caracterizada por sus primeros N momentos; $\mu_t = E[y_t \mid \mathfrak{F}_t]$, y $h_t^{(r)} = E[(y_t - \mu_t)^r \mid \mathfrak{F}_t]$, $r = 1,\ldots,N$. Nótese que $h_t^{(1)} = 0$. La dimensión no paramétrica del modelo está dada por el desconocimiento de la distribución $\mathfrak{D}$, y la dimensión paramétrica está dada por la especificación de los momentos condicionales

$$\mu_t = \mu_t(\theta_1; \mathfrak{F}_t),$$

$$h_t^{(r)} = E[u_t^r \mid \mathfrak{F}_t] = h_t^{(r)}(\theta_r; \mathfrak{F}_t), \quad r \geq 2,$$

donde $u_t = y_t - \mu_t$ son las innovaciones en la media de y_t, y θ_r es un vector de parámetros en el espacio paramétral $\Theta_r \subset \mathbb{R}^{Pr}$, $r = 1,\ldots, N$. Sería demasiado pretencioso el pensar que con el estado actual del arte en economía podemos derivar de la teoría directamente parametrizaciones para momentos de orden superior, por lo que una exploración empírica podría comenzar a partir de tres consideraciones:

- Interpretemos la proposición teórica $Y = f(Z)$ como que "toda la distribución condicional de Y dada Z es una función de Z, y no solamente su media". Por lo tanto, usemos funciones de las variables explicativas Z como argumentos de los momentos altos.
- Aunque nuestro interés se centra en momentos libres (i.e. no determinados por momentos de orden inferior), parece sensato vincularlos a momentos de menor orden. Esto permite, entre otras cosas, probar si la información está siendo generada por un proceso más simple que puede representarse por menos momentos.
- Según la información se va haciendo disponible, las expectativas deben irse revisando en todos los momentos y no solamente en los primeros. Esto introduce dinámica al relacionar $h_t^{(r)}$ con su pasado y con valores rezagados de u_t^r, para toda r. Lo anterior extiende el argumento de Engle [1982] a momentos superiores al segundo.

Estas consideraciones pueden resumirse en la propuesta

$$h_t^{(r)} = E[u_t^r \mid \mathfrak{F}_t] = h_t^{(r)}[z_t, \mu_t, h_t^{(2)}, \ldots, h_t^{(r-1)}; u_{t-j}^r, h_{t-j}^{(r)}, j > 0], \qquad -(2)$$

con la cual los momentos se parametrizan secuencialmente, y un nuevo conjunto de parámetros, digamos α_r, es incorporado con cada momento adicional. Así

$$h_t^{(r)} = h_t^{(r)}(\theta_r ; \mathfrak{F}_t) = h_t^{(r)}(\theta_{r-1}, \alpha_r ; \mathfrak{F}_t) = h_t^{(r)}(\beta , \alpha_2 , \ldots , \alpha_r ; \mathfrak{F}_t) , \qquad -(3)$$

donde $\theta_r = (\beta', \alpha_2', \ldots, \alpha_r')' = (\theta_{r-1}', \alpha_r')'$, y hemos definido $\beta \equiv \alpha_1$ por coherencia con la notación usual en que β es el vector de parámetros de la media, de modo que $\Theta_r \subseteq \Theta_{r+1}$. Sea $k_r = \dim (\alpha_r)$ y $p_r = \dim (\theta_r) = p_{r-1} + k_r$. El vector total de parámetros es $\theta = \theta_N$, de dimensión $p = p_N$. Definimos las innovaciones en el r-ésimo momento como

$$\varepsilon_t^{(r)} = u_t^r - h_t^{(r)} ,$$

de forma que $E[\varepsilon_t^{(r)} | \mathfrak{F}_t] = 0$, y las $\varepsilon_t^{(r)}$ son una sucesión de martingalas. Notemos que $\varepsilon_t^{(1)} = u_t$. La ecuación de regresión en el r-ésimo momento se define naturalmente como

$$u_t^r = h_t^{(r)}(\theta_r) + \varepsilon_t^{(r)} ,$$

y la covarianza entre innovaciones de distintos momentos es

$$E[\varepsilon_t^{(r)} \varepsilon_t^{(s)} | \mathfrak{F}_t] = h_t^{(r+s)} - h_t^{(r)} h_t^{(s)} .$$

Encimando las N ecuaciones obtenemos el sistema de ecuaciones

$$\eta_t = g_t(\theta) + v_t , \qquad -(4)$$

con innovaciones

$$v_t = \eta_t - E[\eta_t | \mathfrak{F}_t] = \eta_t - g_t(\theta) , \qquad -(5)$$

donde $\eta_t = (y_t , u_t^2 , \ldots , u_t^N)'$, $g_t = (\mu_t , h_t^{(2)} , \ldots , h_t^{(N)})'$, y $v_t = (u_t , \varepsilon_t^{(2)} , \ldots , \varepsilon_t^{(N)})'$. Suponiendo que la distribución condicional posee 2 N momentos, la matriz de covarianzas condicionales es

$$\Sigma_t = E[v_t v_t' | \mathfrak{F}_t] = \| \sigma_{trs} \| = \| h_t^{(r+s)} - h_t^{(r)} h_t^{(s)} \| . \qquad -(6)$$

Esta matriz de covarianzas no es diagonal. Por lo tanto las ecuaciones están correlacionadas y existe información compartida entre ellas. En coherencia con la parametrización secuencial de los momentos, existe una forma simple de diagonalizar la matriz de covarianzas de forma tal que se produzca una versión de la ecuación del r-ésimo momento condicionada a la información contenida en los momentos inferiores. Para este efecto, sea $\bar{\varepsilon}_t^{(r)}$ la innovación en la r-ésima ecuación condicionada, y que se obtiene secuencialmente definiendo $\bar{\varepsilon}_t^{(1)} = \varepsilon_t^{(1)} = u_t$ y, para $r = 2, \ldots, N$,

$$c_t(r,s) = \mathrm{cov}[\varepsilon_t^{(r)} , \bar{\varepsilon}_t^{(s)}] \quad \text{para } s \leq r ,$$

$$v_t(r) = \text{var}[\,\bar{\varepsilon}_t^{(r)}\,],$$

y

$$\bar{\varepsilon}_t^{(r)} = \varepsilon_t^{(r)} - \sum_{j=1}^{r-1} \frac{c_t(r,j)}{v_t(j)}\ \bar{\varepsilon}_t^{(j)}. \qquad -(7)$$

Entonces tenemos el siguiente

Lema 1.- $\text{cov}[\,\bar{\varepsilon}_t^{(r)}, \bar{\varepsilon}_t^{(s)}] = 0$, para $r, s = 1,\ldots, N$, $r \neq s$. □

Las funciones $c_t(r,s)$ y $v_t(r)$ se pueden calcular recursivamente para $r = 2,\ldots,N$ usando (6) y la expresión

$$c_t(r,s) = \text{cov}[\,\varepsilon_t^{(r)}, \varepsilon_t^{(s)} - \sum_{j=1}^{s-1} \frac{c_t(s,j)}{v_t(j)}\ \bar{\varepsilon}_t^{(j)}] = \text{cov}[\varepsilon_t^{(r)}, \varepsilon_t^{(s)}] - \sum_{j=1}^{s-1} \frac{c_t(s,j)\,c_t(r,j)}{v_t(j)},$$

para $s \leq r$, notando que $v_t(r) = c_t(r,r)$ en vista de la naturaleza dondicionada de $\bar{\varepsilon}_t^{(j)}$.

La ecuación condicionada para el r-ésimo momento es entonces

$$u_t^r - \sum_{j=1}^{r-1} \frac{c_t(r,j)}{v_t(j)}\ u_t^j = h_t^{(r)}(\theta_r) - \sum_{j=1}^{r-1} \frac{c_t(r,j)}{v_t(j)}\ h_t^{(j)}(\theta_j) + \bar{\varepsilon}_t^{(r)}, \qquad -(8)$$

que definiendo

$$u_t^{(r/r-1)} = u_t^r - \sum_{j=1}^{r-1} \frac{c_t(r,j)}{v_t(j)}\ u_t^j,$$

y

$$h_t^{(r/r-1)}(\theta_r) = h_t^{(r)}(\theta_r) - \sum_{j=1}^{r-1} \frac{c_t(r,j)}{v_t(j)}\ h_t^{(j)}(\theta_j),$$

se puede expresar como

$$u_t^{(r/r-1)} = h_t^{(r/r-1)}(\theta_r) + \bar{\varepsilon}_t^{(r)}, \qquad -(9)$$

o bien como sistema de ecuaciones

$$\eta_t^* = g_t^*(\theta) + v_t^*, \qquad -(10)$$

donde $\eta_t^* = (y_t, u_t^{(2/1)},\ldots, u_t^{(N/N-1)})'$, $g_t^* = (\mu_t, h_t^{(2/1)},\ldots, h_t^{(N/N-1)})'$, y $v_t^* = (u_t, \bar{\varepsilon}_t^{(2)},\ldots, \bar{\varepsilon}_t^{(N)})'$. Con base en el Lema 1 se puede construir una descomposición tipo Cholesky de Σ_t, $\Sigma_t = R_t D_t R_t'$, donde R_t es triangular y D_t es diagonal, de forma tal que $\eta_t^* = R_t^{-1}\eta_t$, $g_t^*(\theta) = R_t^{-1} g_t(\theta)$, y $v_t^* = R_t^{-1} v_t$, y por tanto θ se puede estimar equivalentemente a partir de (4) ó

(10). La diferencia cualitativa entre los dos sistemas radica en que la matriz de covarianzas condicional de v_t^* es diagonal (D_t) , y por tanto la información en cada ecuación en (10) puede extraerse en forma independiente de las demás.

§3.- BASES GENERALES PARA LA ESTIMACION

Desde un punto de vista teórico, la estimación de (4) ó (10) no presenta aspectos cualitativamente diferentes a los de la estimación de modelos heteroscedásticos, y la discusión de problemas como el de identificabilidad se ignora en aras de simplificar la exposición (ver Sabau [1987, 1988]). Las condiciones de regularidad que garantizar la consistencia y normalidad asintótica de estimadores incluyen identificabilidad, de restricciones sobre la heterogeneidad y/o memoria del proceso (Hansen [1982], White y Domowitz [1984]), existencia de momentos hasta de orden 2 N, suavidad, acotamiento y dominancia en las funciones $h_t^{(r)}$. Estas condiciones se detallan en el anexo. La no observabilidad de u_t^r introduce aspectos cualitativamente distintos para distribuciones simétricas y asimétricas, y en aras de la generalidad su trato se pospone a la siguiente sección. Por el momento, los estimadores son caracterizados con una '*' para notar que dependen de inobservables (y por tanto no son propiamente estimadores).

Podemos obtener estimadores iniciales $\hat{\theta}_\ell$ de θ utilizando mínimos cuadrados (no lineales, MCNL) en la ecuación de la media para obtener $\hat{\beta}_\ell$ de β , y generando secuencialmente estimadores $\hat{\alpha}_r^\ell$ de α_r mediante MCNL para $r = 2,\ldots,N$ en

$$u_t^r = h_t^{(r)}(\alpha_r\,;\, \hat{\theta}_{r-1}^\ell) + \bar{e}_t^{(r)},$$

donde $\bar{e}_t^{(r)} = e_t^{(r)} + \{ h_t^{(r)}(\theta_r) - h_t^{(r)}(\alpha_r\,;\, \hat{\theta}_{r-1}^\ell) \}$, y $\hat{\theta}_{r-1}^\ell = (\hat{\beta}_\ell', \hat{\alpha}_2^{\ell\prime}, \ldots, \hat{\alpha}_{r-1}^{\ell\prime})'$ (Ver Jobson y Fuller [1980] para el caso N=2). Se tiene entonces el siguiente

Lema 2- Bajo las condiciones de regularidad, $\hat{\theta}_\ell$ es un estimador fuertemente consistente del valor real θ_0 de θ . □

Dada la existencia de estos estimadores consistentes, $c_t(r,s)$ y $v_t(r)$ pueden tomarse como dadas para la construcción y estimación del sistema en (10). Esto puede hacerse porque la estimación de θ_r no se ve afectada al orden $T^{1/2}$ si se utilizan estimadores consistentes en la construcción de $c_t(r,s)$ y $v_t(s)$ (ver Sabau [1987]).

Ahora bien, para estimar θ_r eficientemente utilizando información solamente del r-ésimo momento, consideremos las condiciones de ortogonalidad

$$\psi_r^*(\theta_r) = T^{-1} \sum_{t=1}^{T-1} v_t(r)^{-1} g_{tr}\, \bar{e}_t^{(r)} = T^{-1} S_r' \Omega_r^{-1} \bar{e}^{(r)},$$

donde $g_{tr} = \partial h_t^{(r/r-1)} / \partial \theta_r$, $S_r = (g_{1r}, \ldots, g_{Tr})'$, $\Omega_r = \text{diag}\{v_t(r)\}$, y $\bar{e}^{(r)} = (\bar{e}_1^{(r)}, \ldots, \bar{e}_T^{(r)})'$. Mostramos el siguiente

Teorema 3. - Bajo las condiciones de regularidad, el estimador $\hat{\theta}_r^*$ obtenido de $\psi_r^*(\theta_r)$ es fuertemente consistente del valor real θ_r^0 de θ_r, y su distribución asintótica es

$$T^{1/2}(\hat{\theta}_r^* - \theta_r^0) \xrightarrow{d} N[0, \mathcal{C}\{T^{-1} S_r' \Omega_r^{-1} S_r\}^{-1}],$$

donde $\mathcal{C}\{\cdot\} = \lim E[\cdot]$, y la esperanza para la matriz de covarianzas se evalúa en θ_0. □

Dado que existe información en distintos momentos acerca de parámetros comunes, es de interés el considerar la estimación conjunta a partir de las condiciones de ortogonalidad

$$\psi^*(\theta) = (\psi_m(\beta)', \psi_2^*(\theta_2)', \ldots, \psi_N^*(\theta)')',$$

con matriz de ponderación

$$A_T = \text{diag}\{T(X'\Omega^{-1}X)^{-1}, T(S_2'\Omega_2^{-1}S_2)^{-1}, \ldots, T(S_N'\Omega_N^{-1}S_N)^{-1}\}.$$

Entonces tenemos el siguiente

Teorema 4. - Bajo las condiciones de regularidad, el estimador $\hat{\theta}_N^*$ obtenido de las condiciones de ortogonalidad $\psi^*(\theta)$ y matriz de ponderaciones A_T, es fuertemente consistente de θ_0 y tiene distribución asintótica dada por

$$T^{1/2}(\hat{\theta}_N^* - \theta_0) \xrightarrow{d} N[0, \mathcal{C}\{T^{-1} \sum_{r=1}^{N} \bar{S}_r' \Omega_r^{-1} \bar{S}_r\}^{-1}],$$

donde $S_j = \partial h^{(j/j-1)} / \partial \theta_j$, y $\bar{S}_j = \partial h^{(j/j-1)} / \partial \theta_j = (S_j, 0)$, $j = 1, \ldots, N$ y la esperanza para la matriz de covarianzas es evaluada en θ_0. □

Si definimos las matrices $B_r = (I_{p_r}, 0)'$, de dimensión $p \times p_r$, entonces

$$V(\hat{\theta}_N^*) = (\sum_{r=1}^{N} B_r V(\hat{\theta}_r^*)^{-1} B_r')^{-1},$$

y como $T^{-1/2}\overline{S}_r'\Omega_r^{-1}\overline{\varepsilon}(r) = (T^{-1/2}\overline{\varepsilon}(r)'\Omega_r^{-1}S_r, 0)' = B_r V(\hat{\theta}_r^*)^{-1} T^{1/2}(\hat{\theta}_r^* - \theta_r^0) + o_p(1)$, se sigue que

$$T^{1/2}(\hat{\theta}_N^* - \theta_0) = V(\hat{\theta}_N^*) \sum_{r=1}^{N} B_r V(\hat{\theta}_r^*)^{-1} T^{1/2}(\hat{\theta}_r^* - \theta_r^0) + o_p(1), \qquad -(11)$$

que muestra que el estimador conjunto tiene una estructura de media ponderada matricial (MPM). Una estrategia interesante consiste en proceder secuencialmente, explorando la posibilidad de extraer información de un momento adicional, hasta que no es posible obtener mayor eficiencia. Cada vez que una ecuación condicionada posee información acerca de parámetros de momentos inferiores, esta puede incorporarse mediante una MPM. Para ver ésto, partamos $\hat{\theta}_r^* = (\hat{\theta}_{r-1}^{(r)*\prime}, \hat{\alpha}_r^{*\prime})'$ (el estimador de θ_r con la información en el r-ésimo momento exclusivamente), y $\hat{\theta}_r^* = (\hat{\theta}_{r-1}^{(r)*\prime}, \hat{\alpha}_r^{*\prime})'$ (el estimador de θ_r con toda la información existente en los primeros r momentos), y demostremos el siguiente

Teorema 5. - Bajo las condiciones de regularidad,

$$\hat{\theta}_{r-1}^{(r)*} = V(\hat{\theta}_{r-1}^{(r)*})[V(\hat{\theta}_{r-1}^*)^{-1}\hat{\theta}_{r-1}^* + V(\hat{\theta}_{r-1}^{(r)*})^{-1}\hat{\theta}_{r-1}^{(r)*}] + o_p(T^{-1/2}),$$

$$V(\hat{\theta}_{r-1}^{(r)*})^{-1} = V(\hat{\theta}_{r-1}^*)^{-1} + V(\hat{\theta}_{r-1}^{(r)*})^{-1},$$

$$\hat{\alpha}_r^* = \tilde{\alpha}_r^* + \mathrm{cov}[\hat{\alpha}_r^*, \hat{\theta}_{r-1}^{(r)*}] V(\hat{\theta}_{r-1}^{(r)*})^{-1}[\hat{\theta}_{r-1}^{(r)*} - \hat{\theta}_{r-1}^{(r)*}] + o_p(T^{-1/2}),$$

$$V(\hat{\alpha}_r^*) = V_r + \mathrm{cov}[\hat{\alpha}_r^*, \hat{\theta}_{r-1}^{(r)*}] V(\hat{\theta}_{r-1}^{(r)*})^{-1}\mathrm{cov}[\hat{\theta}_{r-1}^{(r)*}, \hat{\alpha}_r^*],$$

y

$$\mathrm{cov}[\hat{\alpha}_r^*, \hat{\theta}_{r-1}^{(r)*}] = -V_r\, \mathcal{E}\{T^{-1}S_{rr}'\Omega_r^{-1}S_{r\,r-1}^*\} V(\hat{\theta}_{r-1}^{(r)*}),$$

donde $V_r = \mathcal{E}\{T^{-1}S_{rr}'\Omega_r^{-1}S_{rr}\}$, y $S_r = (S_{r\,r-1}^*, S_{rr})$. □

Si ya no existe en el r-ésimo momento información acerca de θ_{r-1}, esto puede detectarse con el siguiente

Teorema 6. - Bajo las condiciones de regularidad (e identificabilidad), $\hat{\theta}_{r-1}^*$ es eficiente con respecto a la información contenida en los primeros r momentos si, y solo si,

$$\mathrm{rango}[\mathcal{E}\{T^{-1}(S_r - \sum_{j=1}^{r-1} C_{rj}\,\Omega_j^{-1}S_{jr})'(\Omega_r - \sum_{j=1}^{r-1} C_{rj}\,\Omega_j^{-1}C_{rj})^{-1}(S_r - \sum_{j=1}^{r-1} C_{rj}\,\Omega_j^{-1}S_{jr})\}] = k_r,$$

donde $C_{rj} = \mathrm{diag}\{c_t(r,j)\}$, and $S_{jr} = (S_j, 0)$ de forma que sea $T \times p_r$, $j = 1, \ldots, r-1$. □

Una condición más simple, suficiente pero no necesaria es dada en el

Corolario 7.- Bajo la condiciones del Teorema 6, $\hat{\theta}^{*}_{r-1}$ es eficiente con respecto a la información contenida en los primeros r momentos si

$$\partial h_t^{(r)}/\partial\theta_{r-1} - \sum_{j=1}^{r-1} v_t(j)^{-1} c_t(r,j)\, \partial h_t^{(j)}/\partial\theta_{r-1}$$

se anula en θ_0, para todo t. □

Ejemplos simples de la aplicación de este corolario son la eficiencia del estimador de mínimos cuadrados generalizados (MCG) bajo distribuciones gamma y Poisson, así como la inexistencia de información en la kurtosis de la normal (Sabau [1988]).

§4.- ESTIMACION FACTIBLE BAJO SIMETRIA Y ASIMETRIA

Hasta ahora hemos ignorado la inobservabilidad de u_t, por lo que los estimadores de la sección previa no son factibles. Para obtener resultados aplicables, debemos utilizar residuos de la estimación de la media para construir las variables dependientes para las ecuaciones para los momentos de orden superior. La versión operativa de la ecuación para el r-ésimo momento es

$$\tilde{u}_t^r = h_t^{(r)}(\theta_r) + e_t^{(r)},$$

donde $e_t^{(r)} = \varepsilon_t^{(r)} + \{\tilde{u}_t^r - u_t^r\}$, y la versión operativa de la ecuación condicionada para el r-ésimo momento es

$$\tilde{u}_t^{(r/r-1)} = h_t^{(r/r-1)}(\theta_r) + \bar{e}_t^{(r)},$$

donde $\bar{e}_t^{(r)} = \bar{\varepsilon}_t^{(r)} + \{\tilde{u}_t^{(r/r-1)} - u_t^{(r/r-1)}\}$, y $\tilde{u}_t^r$ y $\tilde{u}_t^{(r/r-1)}$ se obtienen sustituyendo estimadores consistentes $\tilde{\beta}$ de β. $\hat{\beta}_m$ (estimador MCG de β) se puede obtener sin información de los momentos de orden superior al segundo, y por tanto en lo que sigue suponemos que $\tilde{\beta} = \hat{\beta}_m$. Esto constituye la mejor herramienta para efectos de eficiencia, ya que las matrices de covarianzas de los estimadores factibles serán funciones directas de $V(\tilde{\beta})$. Las condiciones de ortogonalidad $\psi_r^{*}(\theta_r)$ deben ser reemplazadas para obtener estimadores factibles $\tilde{\theta}_r$ por

$$\psi_r(\theta_r) = T^{-1} \sum_{t=1}^{T-1} v_t(r)^{-1} g_{tr}\, \bar{e}_t^{(r)} = T^{-1} S_r' \Omega_r^{-1} \bar{e}^{(r)},$$

con lo que tenemos

Teorema 8. - Bajo las condiciones de regularidad,

$$T^{1/2}(\tilde{\theta}_r - \theta_r^0) = T^{1/2}(\tilde{\theta}_r^* - \theta_r^0) + V(\tilde{\theta}_r^*) A_r T^{1/2}(\hat{\beta}_m - \beta_0) + o_p(1), \qquad -(12)$$

donde $A_r = \mathcal{E}\{T^{-1} \sum_{t=1}^{T} v_t(r)^{-1}[h_t^{(r-1)} - \sum_{j=3}^{r-1} \frac{c_t(r,j)}{v_t(j)} h_t^{(j-1)}] s_{rt} x_t'\}$, $r > 2$, y $A_r = 0$, $r \leq 2$. □

Este Teorema establece que los estimadores factibles se mantienen consistentes, lo que se debe a que $T^{-1} \sum \tilde{f}_t [\tilde{u}_t^r - u_t^r] \xrightarrow{as} 0$ siempre que $E[u_t^r]$ exista. Pero la distribución de los estimadores factibles y no-factibles será distinta a menos que el segundo término en (12) se anule lo suficientemente rápido, lo que requiere la condición más fuerte de que $T^{-1/2} \sum \tilde{f}_t [\tilde{u}_t^r - u_t^r] \xrightarrow{as} 0$. El Lema 2 de Sabau [1987] establece que esta condición es satisfecha por los dos primeros momentos, y sugiere que para momentos de orden superior las propiedades de los estimadores factibles diferirán según la distribución de y_t sea simétrica o asimétrica. De hecho, tenemos el siguiente

Corolario 9. - Bajo los supuestos del Teorema 8, y si la distribución condicional de y_t es simétrica, entonces para r par,

$$T^{1/2}(\tilde{\theta}_r - \theta_r^0) = T^{1/2}(\tilde{\theta}_r^* - \theta_r^0) + o_p(1). \qquad \square$$

Por lo tanto, los estimadores factibles tienen la misma distribución asintótica que los no factibles bajo simetría, y las ecuaciones nones se eliminan del sistema. Newey [1986] ha utilizado estas ecuaciones nones como restricciones de momentos condicionales para mejorar la eficiencia en la estimación. Para distribucviones asimétricas tenemos el siguiente

Corolario 10. - Bajo los supuestos del Teorema 8, y si la distribución condicional de y_t es asimétrica, entonces para r > 2,

$$T^{1/2}(\tilde{\theta}_r - \theta_r^0) \xrightarrow{d} N[0, V(\tilde{\theta}_r^*) + V(\tilde{\theta}_r^*) A_r V(\hat{\beta}_m) A_r' V(\tilde{\theta}_r^*)]. \square \qquad -(13)$$

El primer componente de $V(\tilde{\theta}_r)$ se obtiene de cualquier paquete de regresión luego de aplicar MCG a la ecuación condicionada operativa del r-ésimo momento, pero el segundo término no puede ser calculado simplemente utilizando la matriz de White [1980] y por tanto tiene que calcularse por separado.

Consideremos ahora la estimación conjunta del sistema. Sustituyendo en (11) los estimadores factibles obtenemos el estimador factible conjunto $\hat{\theta}_N$ de θ_0,

$$T^{1/2}(\hat{\theta}_N - \theta_0) = V(\hat{\theta}^*_N) \sum_{r=1}^{N} B_r V(\hat{\theta}^*_r)^{-1} T^{1/2}(\tilde{\theta}_r - \theta_r^0) + o_p(1), \qquad -(14)$$

para el que establecemos el siguiente

Teorema 11.- Bajo las condiciones de regularidad, si la distribución condicional de y_t es simétrica,

$$T^{1/2}(\hat{\theta}_N - \theta_0) = T^{1/2}(\hat{\theta}^*_N - \theta_0) + o_p(1),$$

y si la distribución condicional de y_t es asimétrica,

$$T^{1/2}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N[0, V(\hat{\theta}^*_N) + V(\hat{\theta}^*_N) B V(\hat{\theta}^*_N)], \qquad -(15)$$

donde

$$B = (\sum_{r=3}^{N} B_r A_r) V(\hat{\beta}_m) (\sum_{r=3}^{N} A_r' B_r') + B_1 (\sum_{r=3}^{N} A_r' B_r') + (\sum_{r=3}^{N} B_r A_r) B_1'. \square \qquad -(16)$$

En problemas asimétricos, el primer término de la matriz de covarianzas sería producido directamente por un paquete de regresión tras la estimación conjunta de los N momentos a partir de las ecuaciones condicionadas. Obsérvese que las sumas para la construcción de la matriz B empiezan a partir de 3 , ya que los dos primeros momentos no tienen problema derivado de la observabilidad de su variable dependiente.

Lo atractivo de construir el estimador conjunto de esta manera es que podemos seguir una búsqueda secuencial de información en momentos cada vez más altos hasta que no encontremos ganancias en eficiencia. Un procedimiento alternativo consistiría en estimar directamente el sistema de ecuaciones *sin* condicionar los momentos superiores. Dado que la transformación entre (4) y (10) es no singular, si la especificación es correcta *a priori* el estimador no factible $\hat{\theta}^*_N$ se puede obtener equivalentemente de las condiciones de ortogonalidad

$$\psi(\theta) = T^{-1} \sum_{t=1}^{T} \frac{\partial g_t'}{\partial \theta} \Sigma_t^{-1} v_t = T^{-1} G' \Sigma^{-1} v,$$

donde ahora definimos $G = (\frac{\partial g_1'}{\partial \theta}, \ldots, \frac{\partial g_T'}{\partial \theta})'$, $\Sigma = \text{diag}\{\Sigma_t\}$, y $v = (v_1', \ldots, v_T')'$. Por tanto una expresión equivalente para $V(\hat{\theta}^*_N)$ es

$$V(\hat{\theta}^*_N) = \mathcal{E}\{T^{-1} \sum_{t=1}^{T} \frac{\partial g_t'}{\partial \theta} \Sigma_t^{-1} \frac{\partial g_t}{\partial \theta'}\} = \mathcal{E}\{T^{-1} G' \Sigma^{-1} G\},$$

evaluando la esperanza en θ_0. Además,

$$T^{1/2}(\hat{\theta}_N^* - \theta_0) = V(\hat{\theta}_N^*)\, T^{-1/2} G' \Sigma^{-1} v + o_p(1). \qquad -(17)$$

Sea $\tilde{\eta}_t = (y_t, \tilde{u}_t^2, \ldots, \tilde{u}_t^N)'$ y $\tilde{v}_t = \tilde{\eta}_t - g_t = v_t + \{\tilde{\eta}_t - \eta_t\}$, y encimando en forma obvia $\tilde{v} = \tilde{\eta} - g = v + (\tilde{\eta} - \eta)$. Sustituyendo $\tilde{v}$ en $\psi(\theta)$ produce el estimador $\hat{\theta}_N^+$, ésto es,

$$T^{1/2}(\hat{\theta}_N^+ - \theta_0) = T^{1/2}(\hat{\theta}_N^* - \theta_0) + V(\hat{\theta}_N^*)\, T^{-1/2} G' \Sigma^{-1}(\tilde{\eta} - \eta) + o_p(1), \qquad -(18)$$

donde hemos utilizado (17). Definiendo $\bar{g}_t = (0, 0, h_t^{(2)}, \ldots, h_t^{(N-1)})'$ y utilizando el Lema 2 de Sabau [1987] se obtiene que

$$T^{-1/2} G' \Sigma^{-1}(\tilde{\eta} - \eta) = T^{-1/2} \sum_{t=1}^{T} \frac{\partial \bar{g}_t'}{\partial \theta} \Sigma_t^{-1}(\tilde{\eta}_t - \eta_t) = C\, T^{1/2}(\hat{\beta}_m - \beta_0) + o_p(1),$$

donde $C = \mathcal{E}\{T^{-1} \sum_{t=1}^{T} \frac{\partial \bar{g}_t'}{\partial \theta} \Sigma_t^{-1} \bar{g}_t x_t'\}$. Substituyendo en (18) y usando $\mathrm{cov}(\hat{\theta}_N^*, \hat{\beta}_m) = V(\hat{\theta}_N^*) B_1$ (ver demostración del Teorema 11), obtenemos

$$T^{1/2}(\hat{\theta}_N^+ - \theta_0) \xrightarrow{d} N[0, V(\hat{\theta}_N^*) + V(\hat{\theta}_N^*) D V(\hat{\theta}_N^*)], \qquad -(19)$$

donde

$$D = C V(\hat{\beta}_m) C' + B_1 C' + C B_1'. \qquad -(20)$$

La inspección de (15) y (19) muestra que los dos procedimientos con y sin condicionamiento son asintóticamente equivalentes si $B = D$, donde B se define en (16) y D en (20). Ya que $V(\hat{\beta}_m) = B_1' V(\hat{\theta}_N^*) B_1$ y $B_1 = (I_k, 0)'$, una condición necesaria y suficiente es que $C = \sum_{r=2} B_r A_r$. Evidentemente, este es el caso cuando la distribución es simétrica y en cuyo caso $B = D = 0$. A pesar de que no intentamos prueba formal, es razonablemente claro que aún en el caso asimétrico se da que $B = D$, en vista de la transformación no singular entre los sistemas con y sin condicionamiento. Lo importante de este resultado es que permite el uso de la estrategia secuencial que permite tratar la búsqueda de especificación en cada momento en forma independiente, y al final existe un procedimiento simple para obtener los estimadores conjuntos dada una especificación seleccionada.

§5.- CONCLUSIONES

En el presente artículo hemos propuesto una teoría de la estimación para una clase muy general de modelos semi-paramétricos en los que la dimensión no paramétrica está dada por la generalidad de la distribución, y la dimensión paramétrica por la especificación de los momentos condicionales. El aspecto central es

el de generar una ecuación para cada momento a partir de la definición de las correspondientes innovaciones. Así, la aplicación del método generalizado de momentos como criterio de estimación resulta natural. El problema de especificación de un modelo con tal generalidad parece colosal. Sin embargo, se puede simplificar utilizando una diagonalización del sistema que condiciona la inferencia en el momento r a la información que ya ha sido extraida de los primeros r-1 momentos. Con ésto, la inferencia en cada momento puede hacerse en forma prácticamente independiente de los restantes, si bien hay un condicionamiento sobre la existencia de errores en momentos de orden inferior.

La no observabilidad de las innovaciones de la media, que aparecen como variables dependientes en las ecuaciones de regresión de los restantes momentos, hace que los estimadores factibles de los parámetros sean cualitativamente distintos según la distribución con la que se esté trabajando sea simétrica a no. En el trabajo se ha obtenido la distribución asintótica de los estimadores para ambos casos.

Para ayudar en la especificación, se puede mostrar que los resultados sobre error de especificación de modelos heteroscedásticos desarrollados en Pagan y Sabau [1987] y Sabau [1988] generalizan en forma natural a los modelos considerados en este artículo, y que lo mismo aplica a las pruebas de diagnóstico a través del análisis de la coherencia de la información en distintos momentos sobre parámetros comunes (Sabau [1987]), de la consistencia de estimadores (Pagan y Sabau [1988]), y la eficiencia de los mismos (Sabau [1988]). Estos temas los dejamos para un trabajo futuro por motivos de espacio.

REFERENCIAS BIBLIOGRAFICAS

Amemiya, T. [1985], *Advanced Econometrics*, Basil Blackwell.

Bera, A.K. y Jarque, C.M. [1982], Model specification tests: a simultaneous approach, *Journal of Econometrics* 20, 59-82.

Engle, R.F. [1982], Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica* 50, 987-1007.

Gallant, A.R. y Nychka, D.W. [1987], Semi-nonparametric maximum likelihood estimation, *Econometrica* 55, 363-390.

Gallant, A.R., y Tauchen, G. [1986], Seminonparametric estimation of conditionally constrained heterogeneous processes: asset pricing applications, *North Carolina State University y Duke University*, manuscrito.

Hansen, L.P. [1982], Large sample properties of generalized method of moments estimators, *Econometrica* 50, 1029-1054.

Jarque, C.M. y Bera, A.K. [1980], Efficient tests for normality, homoscedasticity and serial independence of regression residuals, *Economics Letters* 6, 255-259.

Jobson, J.D. y Fuller, W.A. [1980], Least squares estimation when the covariance matrix and parameter vector are functionally related, *Journal of the American Statistical Association* 75, 176-181.

Newey, W.K. [1986], Adaptive estimation of regression models via moment restrictions, *Princeton University*, Research Memorandum 330.

Pagan, A.R. [1986], Two stage and related estimators and their applications, *Review of Economic Studies* 53, 517-538.

Pagan, A.R. y Sabau, H. [1987], On the inconsistency of the MLE in certain heteroskedastic regression models, *University of Rochester y ANU*, manuscrito.

Pagan, A.R. y Sabau, H. [1988], Consistency tests for heteroskedastic and risk models, *University of Rochester y ANU*, manuscrito.

Sabau, H.C. [1987], The structure of GMM and ML estimators in conditionally heteroskedastic models, *ANU Working Papers in Economics and Econometrics* 153.

Sabau, H.C. [1988], *Some Theoretical Aspects of Econometric Inference with Heteroskedastic Models*, Tesis de Doctorado, ANU.

Spanos, A. [1986], *Statistical Foundations of Econometric Modelling*, Cambridge University Press.

White, H. [1980], A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* 48, 817-838.

White, H. y Domowitz, I. [1984], Nonlinear regression with dependent observations, *Econometrica* 52, 143-161.

White, H. y McDonald, G.M. [1980], Some large-sample tests for nonnormality in the linear regression model, *Journal of the American Statistical Association* 75, 16-28.

ANEXO

I.- CONDICIONES DE REGULARIDAD

(C0) *La media condicional de* y_t *es* $\mu_t(\beta_0) = E[y_t | \mathcal{F}_t]$ *, y los momentos de orden dos y superiores están dados por* $h_t^{(r)}(\theta_0) = E[\{y_t - \mu_t(\beta_0)\}^r | \mathcal{F}_t]$ *,* $r \geq 2$ *. Los momentos existen y son finitos condicionalmente en* $\mathcal{F}_t$ *para* $r \leq 2N$.

(C1) (y_t, z_t') *es estacionario y ergódico, ó alternativamente* (y_t, z_t') *mezcla con* $\phi(m)$ *de tamaño* $s/(2s-1)$ *,* $s \geq 1$ *, o con* $\alpha(m)$ *de tamaño* $s/(s-1)$ *,* $s > 1$ *.*

(C2) Θ *es un subespacio compacto de* $\mathbb{R}^p$ *y* θ_0 *es un punto interior de* Θ *.*

(C3) μ_t *y* $h_t^{(r)}$, $2 \leq r \leq N$, *son funciones medibles de* $\mathcal{F}_t$ *para todo* $\theta \in \Theta$ *, y son continuamente diferenciables al orden 2 en* Θ *, uniformemente en* t *. Más aún,* μ_t *y las* $h_t^{(r)}$ *y sus primeras dos derivadas con respecto a* θ *son acotadas por arriba en valor absoluto, y los momentos pares están acotados por abajo en valor absoluto por una cantidad positiva casi en cualquier lado en* Θ *, uniformemente en* t *.*

(C4) $\varepsilon_t^{(r)} = \varepsilon_t(\theta_0)^r$ *es uniformemente* $(s+\delta)$ *-integrable para* $2 \leq r \leq N$ *,* $s \geq 1, \delta > 0$, *mientras que* $[\mu_t(\beta_0) - \mu_t(\beta)]^2$ *y* $[h_t^{(r)}(\theta_0) - h_t^{(r)}(\theta)]^2$ *están dominados por funciones uniformemente* $(s+\delta)$ *-integrables.*

(C5) $\dfrac{\partial g_t'}{\partial \theta} \Sigma_t^{-1} \dfrac{\partial g_t}{\partial \theta'}$ *, que es una función matricial de* θ *, está dominada por funciones* s *-integrables,* $s > 1$ *, y tiene esperanza finita en* θ_0 *.*

(C6) $h_t^{-1}\left\{\dfrac{\partial \mu_t}{\partial \beta}\dfrac{\partial \mu_t}{\partial \beta'} - u_t(\beta)\dfrac{\partial^2 \mu_t}{\partial \beta\, \partial \beta'}\right\}$ *y* $[h_t^{(2r)} - h_t^{(r)2}]^{-1}\left\{\dfrac{\partial h_t^{(r)}}{\partial \theta_r}\dfrac{\partial h_t^{(r)}}{\partial \theta_r'} - \varepsilon_t^{(r)}(\theta)\dfrac{\partial^2 h_t^{(r)}}{\partial \theta_r\, \partial \theta_r'}\right\}$ *,*

$2 \leq r \leq N$, *están dominadas por funciones uniformemente* s *-integrables,* $s > 1$.

(C7) $\mathrm{Var}[T^{-1/2} G' \Sigma^{-1} v] = E[T^{-1} G' \Sigma^{-1} G]$ *es uniformemente positiva definida en una vecindad abierta de* θ_0.

II.- DEMOSTRACIONES

Demostración del Lema 1 : Usamos inducción sobre r . Como $\mathrm{cov}[\bar{\varepsilon}_t^{(r)}, \bar{\varepsilon}_t^{(s)}] = \mathrm{cov}[\bar{\varepsilon}_t^{(s)}, \bar{\varepsilon}_t^{(r)}]$ es suficiente considerar $s < r$. Para $r = 2$, $c_t(2,1) = \mathrm{cov}[\varepsilon_t^{(2)}, u_t] = h_t^{(3)} = \delta_t$, $v_t(1) = \mathrm{var}[u_t] = h_t^{(2)} = h_t$. Entonces $\bar{\varepsilon}_t^{(2)} = \varepsilon_t^{(2)} - h_t^{-1}\delta_t u_t$, y por tanto $\mathrm{cov}[\bar{\varepsilon}_t^{(2)}, \bar{\varepsilon}_t^{(1)}] = E[(\varepsilon_t^{(2)} - h_t^{-1}\delta_t u_t)u_t] = 0$, puesto que $E[\varepsilon_t^{(2)} u_t] = h_t$ y $E[h_t^{-1}\delta_t u_t^2] = h_t$.

Supongamos ahora que $\text{cov}[\,\bar{\varepsilon}_t^{(r-1)}, \bar{\varepsilon}_t^{(s)}] = 0$, $s < r-1$. Entonces, usando (7) , tenemos para $s < r$ que

$$\text{cov}[\,\bar{\varepsilon}_t^{(r)}, \bar{\varepsilon}_t^{(s)}] = \text{cov}[\,\{\varepsilon_t^{(r)} - \sum_{j=1}^{r-1} \frac{c_t(r,j)}{v_t(j)}\ \bar{\varepsilon}_t^{(j)}\}, \bar{\varepsilon}_t^{(s)}] = c_t(r,s) - \frac{c_t(r,s)}{v_t(s)}\, v_t(s) = 0 .$$ □

Demostración del Lema 2 : ver Pagan [1986] . □

Los siguientes resultados del texto se demuestran esencialmente como referencia a los resultados de Hansen [1982] o bien de White y Domowitz [1984], para condiciones de alternativas de heterogeneidad y memoria sobre los procesos estocásticos involucrados. Por ello es conveniente hacer notar que las condiciones de regularidad mencionadas son satisfechas en este contexto e incluir el siguiente

Teorema A.- Supongamos que un vector de parámetros $\lambda \in \Lambda \subset \mathbb{R}^n$ va a ser estimado mediante el criterio

$$\hat{\lambda} = \min_{\lambda \in \Lambda}\ \psi(\lambda)' A_T \psi(\lambda)$$

donde $\psi(\lambda) = T^{-1}\sum \psi_t(\lambda)$, $E[\,\psi_t(\lambda_0)\,] = 0$, y $A_T \xrightarrow{as} \mathcal{C}\{T\,\psi(\lambda_0)\,\psi(\lambda_0)'\}$, donde λ_0 es el valor real de λ . Supongamos que las condiciones de regularidad de Hansen [1982] y/o White y Dopmowitz [1984] se satisfacen. Entonces

$$T^{1/2}(\hat{\lambda} - \lambda_0) \xrightarrow{d} N[0, V(\hat{\lambda})],$$

donde

$$V(\hat{\lambda}) = [\,\mathcal{C}\{\frac{\partial \psi(\lambda_0)'}{\partial \lambda}\}\,\mathcal{C}\{T\,\psi(\lambda_0)\,\psi(\lambda_0)'\}^{-1}\,\mathcal{C}\{\frac{\partial \psi(\lambda_0)}{\partial \lambda'}\}\,]^{-1}.$$

Demostración: Ver Teoremas 3.1 y 3.2 de White y Domowitz [1984], y Teoremas 2.1 y 3.1 de Hansen [1982]. □

Demostración del Teorema 3 : se satisfacen las condiciones del Teorema A. □

Demostración del Teorema 4 : Los supuestos conforman al Teorema A. Como las ecuaciones no están correlacionadas,

$$E[\,T\,\psi^*(\theta)\,\psi^*(\theta)'] = E[\,A_T(\theta_J)\,] = \text{diag}\,\{E[\,T^{-1}S_r'\Omega_r^{-1}S_r]^{-1}\},$$

y usando expectativas iteradas se ve que $E[\,\partial\,\psi_r^*(\theta_r^0)/\partial\theta'] = -E[\,T^{-1}S_r'\Omega_r^{-1}\bar{S}_r]$, y así

$$E[\frac{\partial \psi^*(\theta_0)'}{\partial \theta}] = -(E[T^{-1}\bar{X}'\Omega^{-1}X],\ldots,E[T^{-1}\bar{S}_N'\Omega_N^{-1}S_N]),$$

y la matriz de covarianzas de $\hat{\theta}$ es

$$V(\hat{\theta}_N^*) = \mathcal{E}\{T^{-1}\sum_{r=1}^{N}\bar{S}_r'\Omega_r^{-1}S_r(S_r'\Omega_r^{-1}S_r)^{-1}S_r'\Omega_r^{-1}\bar{S}_r\}^{-1}.$$

Pero $S_r = \bar{S}_r(I, 0)$, y por tanto $\bar{S}_r'\Omega_r^{-1}S_r(S_r'\Omega_r^{-1}S_r)^{-1}S_r'\Omega_r^{-1}\bar{S}_r = \bar{S}_r'\Omega_r^{-1}\bar{S}_r$, lo que completa la prueba. □

Demostración del Teorema 5 : En vista de la parametrización secuencial de $h_t^{(r)}$ y la naturaleza condicionada de las ecuaciones, el problema de incorporar la información del r-ésimo momento dado que los primeros r-1 han sido tomados en cuenta es un problema con la misma estructura que la incorporación de información de la varianza en un modelos heteroscedástico bajo simetría. Aplicamos por tanto a $\hat{\theta}_r = (\hat{\theta}_{r-1}^{(r)*\prime}, \hat{\alpha}_r')'$ los Corolarios 1.1 y 1.2 de Sabau [1987] para establecer su relación con $\hat{\theta}_{r-1}^*$ y $\tilde{\theta}_r = (\tilde{\theta}_{r-1}^{(r)*\prime}, \tilde{\alpha}_r')'$. □

Demostración del Teorema 6 : Sea V la matriz en el enunciado, y nótese que

$$S_r - \sum_{j=1}^{r-1}C_{rj}\,\Omega_j^{-1}S_{jr} = (S_{r\,r-1}^* - \sum_{j=1}^{r-1}C_{rj}\,\Omega_j^{-1}S_{jr}^*, S_{rr}),$$

donde $S_{jr} = (S_{jr}^*, 0)$. Por tanto,

$$\text{rango}[V] \geq \text{rango}[\mathcal{E}\{T^{-1}S_{rr}'(\Omega_r - \sum_{j=1}^{r-1}C_{rj}\,\Omega_j^{-1}C_{rj})^{-1}S_{rr}\}] = k_r,$$

donde la última igualdad se sigue de la identificabilidad de α_r en la r-ésima ecuación condicionada. Si el rango de V es k_r existen solamente k_r funciones identificables de θ_r en la r-ésima ecuación condicionada, pero como la parametrización es secuencial y θ_r es identificable a partir de los primeros r momentos, la información en la r-ésima ecuación sirve para identificar α_r solamente, y por tanto $V[\hat{\theta}_{r-1}^*] = V[\hat{\theta}_{r-1}^{(r)*}]$. De igual forma, si $V[\hat{\theta}_{r-1}^*] = V[\hat{\theta}_{r-1}^{(r)*}]$ la r-ésima ecuación no proporciona información acerca de θ_{r-1}, y se sigue que rango $[V] \leq k_r$. Por lo tanto, rango $[V] = k_r$. □

Demostración del Corolario 7: Cuando $\partial h_t^{(r)}/\partial\theta_{r-1} - \sum_{j=1}^{r-1} v_t(j)^{-1}c_t(r,j)\,\partial h_t^{(j)}/\partial\theta_{r-1}$ es nulo en θ_0, para todo t, la matriz V del Teorem 6 tiene ceros en todos lados excepto por la

submatriz

$\mathcal{E}\{ T^{-1}S_{rr}'(\Omega_r - \sum_{j=1}^{r-1} C_{rj}\ \Omega_j^{-1} C_{rj})^{-1} S_{rr} \}$, y por tanto, rango $[V] = k_r$. □

Demostración del Teorema 8 : Usando $\tilde{e}_t^{(r)}$ en lugar de $\tilde{\varepsilon}_t^{(r)}$ en las condiciones de ortogonalidad $\psi_r(\theta_r)$ resulta en

$$T^{1/2}(\tilde{\theta}_r - \theta_r^0) = T^{1/2}(\tilde{\theta}_r^* - \theta_r^0) + V(\tilde{\theta}_r^*) T^{-1/2} \sum_{t=1}^{T} v_t(r)^{-1} s_{tr} \{ \tilde{u}_t^{(r/r-1)} - u_t^{(r/r-1)} \} + o_p(1),$$

donde el segundo término del lado derecho es la contribución debida a la inobservabilidad de u_t^r. Substituyendo $u_t^{(r/r-1)}$ de (8) y (9), con el Lema 2 de Sabau [1987],

$$T^{1/2}(\tilde{\theta}_r - \theta_r^0) = T^{1/2}(\tilde{\theta}_r^* - \theta_r^0) + V(\tilde{\theta}_r^*) A_r T^{1/2}(\hat{\beta}_m - \beta_0) + o_p(1),$$

donde $A_r = \mathcal{E}\{ T^{-1} \sum_{t=1}^{T} v_t(r)^{-1} [h_t^{(r-1)} - \sum_{j=3}^{r-1} \frac{c_t(r,j)}{v_t(j)} h_t^{(j-1)}] s_{rt} x_t' \}$ para $r > 1$. $A_1 = 0$ pues la ecuación de la media no tiene problema de observabilidad, y $A_2 = 0$ pues $h_t^{(1)} = 0$. □

Demostración del Corolario 9 : Cuando la distribución es simétrica $h_t^{(r)} \equiv 0$ para r non y las ecuaciones de orden non para $r > 1$ son eliminadas de los sistemas (4) - (6) y (10). Por tanto las ecuaciones condicionadas solamente remueven la correlación con momentos pares de orden inferior, pero para estos la matriz A_r es función de $h_t^{(j)}$ solo para j non, siguiéndose que $A_r = 0$. □

Demostración del Corolario 10 : Por construcción, cov $(\tilde{\theta}_r^*, \tilde{\theta}_s^*) = 0$ para $r \neq s$, y en particular si $r \geq 2$ cov $(\tilde{\theta}_r^*, \hat{\beta}_m) = 0$. Por tanto los dos términos en (12) son asintóticamente independientes y la matriz de covarianzas en (13) es inmediata. □

Demostración del Teorema 11 : Para distribuciones simétricas el resultado sigue del Corolario 9. Para distribuciones asimétricas sustituimos (12) en (14) para obtener

$$T^{1/2}(\hat{\theta}_N - \theta_0) = T^{1/2}(\tilde{\theta}_N^* - \theta_0) + V(\tilde{\theta}_N^*) (\sum_{r=1}^{N} B_r A_r) T^{1/2}(\hat{\beta}_m - \beta_0) + o_p(1).$$

Ahora bien,

$$\text{cov}(\tilde{\theta}_N^*, \hat{\beta}_m) = V(\tilde{\theta}_N^*) \sum_{r=1}^{N} B_r V(\tilde{\theta}_r^*)^{-1} \text{cov}(\tilde{\theta}_r^*, \hat{\beta}_m) = V(\tilde{\theta}_N^*) B_1,$$

puesto que $T^{1/2}(\hat{\beta}_m - \tilde{\theta}_1^*) = o_p(1)$ y $cov(\tilde{\theta}_r^*, \hat{\beta}_m) = 0$ para $r > 1$ por construcción. Por lo tanto,

$$T^{1/2}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N[0, V(\tilde{\theta}_N^*) + V(\tilde{\theta}_N^*) B V(\tilde{\theta}_N^*)],$$

donde $B = (\sum_{r=0}^{N} B_r A_r) V(\hat{\beta}_m) (\sum_{r=0}^{N} A_r' B_r') + B_1 (\sum_{r=0}^{N} A_r' B_r') + (\sum_{r=0}^{N} B_r A_r) B_1'$. □

La Combinación de Pronósticos desde un punto de vista de la programación matemática.

Gaytán Iniestra J.
Faculta de de Ingeniería de la UAEM,
Cerro de Coatepec, Ciudad Universitaria
Toluca México.

RESUMEN: La combinación de pronósticos ha probado ser una estrategia que mejora la calidad de los pronósticos. Entre otras estrategias, la programación matemática ha sido sugerida como una estrategia de combinación (Reeves et. al (1982), Moussourakis (1987)). En este artículo se proponen extensiones al modelo propuesto por Reeves y Lawrence que mejoran la calidad de los pronósticos. La estrategia fundamental se basa en considerar a los pesos asignados a los pronósticos dependientes de la edad de las observaciones. Finalmente, el desempeño de experimentos iniciales de esas revisiones son presentados.

1. INTRODUCCION

En un artículo fundamental, Bates et. al (1969) demostraron que una combinación lineal de pronósticos opaca a los pronósticos individuales. Estudios posteriores se han concentrado en el cálculo de promedios simples de pronósticos individuales (por ejemplo Makridakis et. al, (1983)-a), o de promedios pesados donde los pesos están restringidos a sumar uno (por ejemplo Newbold et. al (1974)). posteriormente Granger et. al (1984)-a proponen expresar el pronóstico utilizando regresión lineal, donde los pronósticos individuales son las variables independientes. La idea es minimizar los errores del pronóstico asignando pesos que tomen en cuenta las dependencias entre los pronósticos individuales, y sus precisiones relativas, imponiendo la usual suposición de normalidad. Lawrence et. al (1985) han investigado la combinación de pronósticos basados en juicios subjetivos y en métodos cuantitativos usuales. Moussourakis (1987) suguiere algunas revisiones al trabajo de Granger et. al (1984)-a, e incrementa la eficiencia de los pronósticos. Moussourakis construye modelos de programación matemática involucrando variables desviacionales, las cuales hacen el papel de los errores del pronóstico.

Este trabajo se organiza como sigue. En la sección 2 se revisan brévemente las estrategias de combinación relacionadas con este trabajo. En la sección 3 se presentan las extensiones al modelo básico de Reeves y Lawrence. En la sección 4 se consideran los resultados experimentales obtenidos al hacer uso de las extensiones propuestas. Conclusiones y direcciones para trabajos futuros son señaladas en la sección 5.

2. MOTIVACION

Supongamos que se tienen dos métodos de pronóstico, A y B. Un pronostico combinado es obtenido al utilizar un modelo compuesto, en el cual los métodos A y B son primero combinados.

Supongamos que se tienen disponibles dos pronósticos $\hat{Y}_A$ y $\hat{Y}_B$, hechos al tiempo t, utilizando los métodos A y B respectivamentedel valor futuro Y_{t+1} de la serie de tiempo $\{Y_t\}$. Sea $\hat{Y}_{C,t}$ un pronóstico de Y_{t+1} combinado mediante el promedio ponderado siguiente:

$$\hat{Y}_{C,t} = w_A Y_{A,t} + w_B \hat{Y}_{B,t} \qquad (1)$$

donde w_A y w_B son las ponderaciones dadas a los pronósticos A y B, satisfaciendo las siguientes condiciones

$$w_A + w_B = 1 \qquad (2)$$

y

$$w_A, w_B \geq 0 \qquad (3)$$

El pronóstico $\hat{Y}_{C,n}$ es insesgado si $\hat{Y}_{A,n}$ y $\hat{Y}_{B,n}$ lo son. Si denotamos por $e_{A,t}$, $e_{B,t}$ y $e_{C,t}$ a los errores de los pronósticos A, B, C al final del periodo t, tenemos por (1) y (2):

$$Y_{t+1} - \hat{Y}_{C,t} = w_A(Y_{t+1} - \hat{Y}_{A,t}) + w_B(Y_{t+1} - \hat{Y}_{B,t})$$

o bién

$$e_{C,t} = w_A e_{A,t} + w_B e_{B,t} \qquad (4)$$

Por lo que la varianza de $e_{C,t}$ está dada por

$$\mathrm{Var}(e_{c,t}) = w_A^2 \mathrm{Var}(e_{A,t}) + w_B^2 \mathrm{Var}(e_{B,t}) + 2 w_A w_B \mathrm{Cov}(e_{A,t}, e_{B,t}) \qquad (5)$$

La determinación de w_A Y w_B en (1) se basa en la minimización de la varianza de $e_{c,t}$ dada por (5) sujeta a la condición (2). Para simplificar la notación, escribamos $\sigma_A^2 = \mathrm{Var}(e_{A,t})$, $\sigma_B^2 = \mathrm{Var}(e_{B,t})$, y $\sigma_{AB}^2 = \mathrm{Cov}(e_{A,t}, e_{B,t})$. Los pesos que minimizan el problema anterior son:

$$w_A = \frac{\sigma_B^2 - \sigma_{AB}^2}{\sigma_A^2 + \sigma_B^2 - 2\sigma_{AB}^2} \qquad (6)$$

$$w_B = \frac{\sigma_A^2 - \sigma_{AB}^2}{\sigma_A^2 + \sigma_B^2 - 2\sigma_{AB}^2} \qquad (7)$$

El método anterior es llamado el de varianza-covarianza. Los pesos (6) y (7) tienen la gran desventaja de involurar

la covarianza de los errores $e_{A,t}$ y $e_{B,t}$, la cual en general no es conocida. Sin embargo, si los errores del pronóstico $e_{A,t}$ y $e_{B,t}$ no están correlacionados, las fórmulas (6) y (7) se reducen a

$$w_A = \frac{\sigma_B^2}{\sigma_A^2 + \sigma_B^2} \qquad (8)$$

$$w_B = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2} \qquad (9)$$

las cuales son más populares.

La estrategia anterior puede generalizarse a más de dos métodos como sigue. supongase que se tienen disponibles r métodos ($r \geq 2$), Y sea Σ la matriz de covarianza de los errores del pronóstico hechos para el periodo t+1. Entonces el vector de pesos $\mathbf{W} = (w_1, w_2, \ldots, w_r)^T$ en el pronóstico:

$$\hat{Y}_{c,t} = \mathbf{W}^T\hat{\mathbf{Y}} = w_1\hat{Y}_1 + w_2\hat{Y}_2 + \ldots + w_r\hat{Y}_r \qquad (10)$$

se obtiene minimizando la varianza del pronóstico dada por:

$$\mathrm{Var}(e_{c,t}) = 1/2\ \mathbf{W}^T\Sigma\ \mathbf{W} \qquad (11)$$

sujeto a

$$\mathbf{W}^T\mathbf{1} = 1 \qquad (12)$$

donde 1 es un r-vector columna con unos como elementos.

La solución de ese problema es:

$$\mathbf{W} = \Sigma^{-1}\mathbf{1} \ / \ \mathbf{1}^T\Sigma^{-1}\mathbf{1}$$

La expresión anterior es la generalización de (6) y (7). Si las covarianzas en la matriz Σ son iguales a cero, $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2)$, donde σ_i^2 es la varianza del error del método de pronóstico i. En ese caso, W queda

$$\mathbf{W} = \frac{1}{\sum_{j=1}^{r} (\sigma_j^2)^{-1}} \left[1/\sigma_1^2,\ 1/\sigma_2^2, \ldots, 1/\sigma_r^2\right]^T$$

Por lo que el peso w_i asociado al i-ésimo método de pronósticos es:

$$(\sigma^2)^{-1}$$

$$W_i = \frac{(\sigma_i^2)^{-1}}{\sum_{j=1}^{r} (\sigma_j^2)^{-1}} \qquad i=1,2,\ldots,r$$

La relación anterior la generalización de (8) Y (9). En la práctica, **W** se estima reemplazando a Σ con una estimación Σ, donde

$$\hat{\Sigma}_{ij} = \sum_{t=1}^{T} e_{it} e_{jt}$$

Varios autores han reconocido que la matriz Σ no es fija en el tiempo (y por consiguiente **W**). En ese caso el uso de W puede ser severamente sub-óptimo. La mayoría de los procedimientos propuestos calculan los pesos **W** mediante una estrategia adaptiva en el tiempo, por ejemplo Granger y Newbold (1977), sin embargo, esto requiere de gran cantidad de trabajo pues se requiere estimar la matriz de varianza-covarianza en cada punto del tiempo.

Numerosos estudios han demostrado que una combinación más eficiente puede ser encontrada utilizando esquemas mucho más simples. Por ejemplo, la experiencia obtenida por los estudios de Makridakis y otros indica que utilizar los promedios pesados basados en la matríz de covarianza de los errores del pronóstico no se desempeña tan bien como el simple promedio de los pronósticos de los métodos usados. Pocas evidencias existían hasta antes de 1982, sin embargo, Makridakis et. al (1983)-b, y (1982)-c, y Makridakis y Winkler (1983)-a reportan resultados basados en estrategias empíricas en los que sobresale la evidencia de que un simple promedio o un promedio ponderado de pronósticos obtenidos por varios métodos, mejora virtualmente todos los pronósticos individuales. Específicamente, Makridakis y Winkler reportan una reducción del 7.2% en el error cuando se combinan dos métodos, hasta un máximo de 16.3% de reducción en el error al combinar 5 métodos.

La principal ventaja del promedio simple o ponderado, es que no requiere información de la precisión de los métodos involucrados, o de la correlación existente entre sus errores, aunque tienen la desventaja de tratar imparcialmente a los métodos, sin distinguir si un método puede desempeñarse mejor que otro. Clemen et. al (1986)-a han hipotetizado que si los métodos siendo combinados exhiben diferencias sustanciales en las varianzas de los errores, el asignar pesos iguales a los pronósticos es probable que la combinación tenga un pobre desempeño.

Otra estrategia de combinación que es más manipulable es la que proporciona la regresión, la cual consiste en expresar la observación Y_t mediante una combinación lineal de los pronósticos individuales F_{1t}, F_{2t},... F_{Kt} como sigue:

$$Y_t = \beta_1 F_{1t} + \beta_2 F_{2t} + \ldots \beta_k F_{Kt} + \varepsilon_t \qquad (13)$$

Donde los parámetros β's se estiman utilizando regresión lineal múltiple, y la variable ε es una variable aleatoria.

Para que los pronósticos sean insesgados (suponiendo que

individualmente los son), es necesario que

$$\beta_1 + \beta_2 + \ldots + \beta_k = 1 \qquad (14)$$

Sin embargo, Granger et. al (1984)-a demuestran que una mejor estrategia consiste en agregar un término constante, digamos β_o, a la combinación (13) y no imponer que los pesos sumen 1 como lo señala la condición (14).

Diebold et. al (1987) demuestran que la estrategia de minimizar la varianza de los errores del pronóstico (11), sujeta a la restricción (12), es un caso particular de la estrategia de estimar los parámetros β's en (13), sujetos a la restricción (14).

La estrategia de regresión ha sido generalizada por Diebold et. al (1987) utilizando parámetros dependientes del tiempo, y reportan experimentos donde el pronóstico combinado tiene un EMAP de sólo el 10% del EMAP del pronóstico individual más malo, y un 40% de la SCE del modelo de regresión ordinario irrestricto.

La estrategia de regresión con parámetros dependientes del tiempo evita la necesidad de estimar explícitamente las matrices de varianza-covarianza utilizando datos móviles, y manipula su evolución implícitamente. La estrategia de combinación de pronósticos basada en la programación matemática es debida a Reeves et. al (1982) y está basada en el siguiente modelo

$$\text{Min } f(d^+, d^-) \qquad (15)$$

Sujeto a

$$\sum_{j=i}^{n} w_j F_{tj} + d_t^- - d_t^+ = Y_t \qquad t = 1,2,\ldots,m \qquad (16)$$

$$\sum_{j=1}^{n} w_j = 1 \qquad (17)$$

$$w_j \geq 0 \qquad j = 1,2,\ldots,n \qquad (18)$$

$$d_t^-,\ d_t^+ \geq 0 \qquad t = 1,2,\ldots,m \qquad (19)$$

donde

n = número de métodos de pronóstico utilizados

m = número de periodos observados de la serie $\{Y_t\}$

w_j = es el peso asignado al pronóstico j, $j = 1,2,\ldots,n$

Y_t = es el valor de la serie observado en el periodo t, $t = 1,2,\ldots,m$

F_{tj} = el pronóstico hecho por el método j para el periodo t, $t = 1,2,\ldots,m$, $j = 1,2,\ldots,n$

$d^+ = (d_1^+, d_2^+, \ldots, d_m^+)^T$ es el vector $_{mx1}$ de variables desviacionales que sobre-estiman el pronóstico combinado respecto al valor observado.

$d^- = (d_1^-, d_2^-, \ldots, d_m^-)^T$ es el vector $_{mx1}$ de variables desviacionales que sub-estiman el pronóstico combinado respecto al valor observado.

$f(d^+, d^-)$ = es la función de los errores del pronóstico. Observe que de esta manera se tienen $_{2m}$ variables desviacionales, y son las que asumen el papel de los errores del pronóstico.

El modelo de programación por metas múltiples anterior tiene grandes ventajas, pues proporciona una manera versatil de obtener los pesos w_j modelando diferentes situaciones reales. Por ejemplo, al permitir uno o varios objetivos se pueden expresar como primer objetivo alguna de las funciones de error más comunes (EMAP, DMA, ECM), y como los restantes, alguna función de los errores para los más recientes periodos. Más aún, se pueden formular situaciones como minimizar el riesgo de sobre-estimar la demanda de algún artículo, lo cual tendría desastrozas consecuencias, sobre todo en la producción o almacenamiento de productos perecederos. Esta última situación requiere de más análisis. Observe que cuando el único objetivo es la Desviación Media Absoluta (DMA), o el Error Medio Absoluto en Porcentaje (EMAP), la función objetivo f es lineal, y el problema (15)-(19) es uno de programación lineal, y cuando la función objetivo es el Error Cuadrático Medio (ECM), el problema es uno de programación cuadrática.

Supongamos que dejamos el número de periodos m fijo, nos preguntamos El pronóstico combinado mejorará agregando más métodos de pronóstico al modelo?, y si mejora, Cuantos y cuales métodos serán adecuados incluir?. Para responder a la primera pregunta supongamos que la medida de error es la DMA o el EMAP, y observemos que en ese caso el modelo anterior es uno de programación lineal. Dado que se tienen $2m+n$ variables y $m+1$ restricciones, en el óptimo se tendrán $m+1$ variables no cero (suponiendo una solución no degenerada), lo cual significa que necesariamente al menos un peso w_j es distinto de cero (debido a (17), y las demás variables básicas son algunas de las variables desviacionales d_t^+ ó d_t^-. Pero si más de un peso w_j es variable básica en el óptimo, se tendrán pronósticos perfectos en algunos de los periodos. Por lo tanto, a mayor número de métodos de pronóstico individuales efectivos incorporados en el modelo, indicado por sus pesos no cero. mayor es el número de pronósticos perfectos, hasta un total de n-1 pronósticos perfectos cuando el número de métodos efectivos es n. Ahora bién, una manera de lograr que el número de pesos distintos de cero en el óptimo sea mayor de uno, es permitiendo que los pesos w_j sean variables irrestrictas en signo, lo cual a su vez sugiere que la restricción (17) sea eliminada. Esta idea fué utilizado por Moussourakis (1987), quién aplicando el modelo (15)-(19) anterior con las tres medidas de desempeño básicas (DMA, EMFA, y ECM), combinó 5 métodos de pronóstico distintos, encontrando que el no restringir los pesos a sumar uno y

dejarlos irrestrictos en signo (restricciones (18) y (19)), produce menores valores de esas medidas de desempeño que restringiendolos. Este resultado está de acuerdo con lo encontrado por Granger et. al (1984)-a.

3. Extensiones al modelo de Reeves y Lawrence.

El modelo básico definido por las expresiones (15)-(19) tiene el inconveniente de no tomar en cuenta la edad de las observaciones. Por lo que la propuesta básica de este trabajo es *pesar a las observaciones de acuerdo a su edad.*

Concretamente, las modificaciones que se proponen son:

i) Determinar los valores de las variables w_j, d_t^-, y d_t^+ cuando la función objetivo que se minimiza es:

$$\text{ia)} \quad f(d^-, d^+) = \sum_{t=1}^{m} C_t (d_t^- + d_t^+)/m \qquad \text{criterio DMA} \qquad (20)$$

$$\text{ib)} \quad f(d^-, d^+) = \sum_{t=1}^{m} C_t (d_t^- + d_t^+)/Y_t\, m \qquad \text{criterio EMAP} \qquad (21)$$

$$\text{ic)} \quad f(d^-, d^+) = \sum_{t=1}^{m} C_t (d_t^- + d_t^+)^2/m \qquad \text{criterio ECM} \qquad (22)$$

donde las constantes C_t asignan un peso distinto al las observaciones dependiendo de su edad. Los valores asignados a las constantes C_t son:

$$\text{id)} \quad K\alpha(1-\alpha)^{m-t} \quad \text{con} \quad 0<\alpha<1,\ t=1,\ldots,m$$

$$\text{ie)} \quad K\, t^{\lambda} \quad \text{con} \quad \lambda \geq 1,\ t=1,\ldots,m$$

La constante K se elige de tal manera que la suma de los pesos sea uno.

ii) Considerar a los pesos asignados a cada método de pronóstico como una función lineal del tiempo. De esta manera los pesos son:

$$\text{iia)} \quad w_j(t) = w_j + v_j\, t, \text{ con } t=1,2,\ldots,m$$

donde w_j y v_j son parámetros a determinar.

La primera estrategia, intenta pesar la edad de las observaciones según su edad. Esto tiene su análogo en regresión lineal ponderada. El esquema id) asigna un peso exponencial a las observaciones, mientras que el esquema ie) incluye los pesos constantes con $\lambda=0$ y el lineal con $\lambda=1$, además de producir pesos con tasa creciente si $\lambda \geq 1$.

La estrategia ii), reemplaza a las restricciones (16) por las siguientes:

$$\sum_{j=1}^{n} (w_j + v_j t) F_{tj} + d_t^- - d_t^+ = Y_t \qquad t = 1,2,\ldots,m \qquad (16')$$

las cuales siguen siendo lineales. El nuevo problema definido por las relaciones (15),(16')-(19) está constituido de m+1 restricciones y 2m+2n variables en el caso restringido, y m restricciones con 2m+2n variables en el caso irrestricto.

4. Resultados experimentales

Para mostrar los resultados obtenidos al usar las dos extensiones propuestas al modelo, utilizaremos como referencia de comparación la misma serie de tiempo y los mismos métodos de pronóstico reportados por Moussourakis (1987), y que mostramos en el cuadro 1.

Los resultados encontrados utilizando los objetivos ia) y ib) con pesos id) son mostrados en el cuadro 2, mientras que utilizando los pesos ie) se presentan en el cuadro 3.

Los valores de los pesos w_j encontrados utilizando los pesos exponenciales id) son reportados en el cuadro 4.

La estrategia i) propuesta reduce notáblemente la variablilidad respecto a la asignación de pesos iguales, siendo consistente en las dos medidas de error estudiadas.

Los resultados obtenidos al utilizar la estrategia ii) también reducen sensíblemente la variabilidad de las 2 medidad de desempeño estudiadas. En particular, las reducciones que se encontraron al utilizar el EMAP y la DMA se indican en el cuadro 5. En dicho cuadro, se incluyen también, para fines de comparación, los valores de las medidas de desempeño cuando en el objetivo es el ie).

Los valores de los pesos w_j y v_j obtenidos para estos experimentos indican que se obtienen 10 pronósticos exactos. Este resultado es interesante, y sus implicaciones en cuanto a la calidad de los pronósticos requieren ser estudiadas con más detalle. La reducción en el EMAP para pesos iguales es del 28% respecto al valor obtenido por Moussourakis, y un 57% en el DMA también para pesos iguales. Este resultado es favorable, y aunque es obtenido para una sola serie, sugiere que se pueden obtener reducciones en las medidas de desempeño considerando las edad de las observaciones. La verificación de este resu ado con series que han sufrido cambios estructurales en su comportamiento es parte de una investigación por parte del autor.

5. Conclusiones

En este reporte se han mencionado brévemente algunas de las estrategias de combinación de pronósticos más populares. La aceptación de estas ideas por parte la comunidad científica ha sido notable, siendo una de las áreas de la estadística más activas en cuanto a investigación. Uno de los aspectos que resaltan por el número de investigaciones dedicadas, es el referente al utilizar regresión restringiendo los pesos w_j a

sumar 1 (Granger et. a_ (1984)-a; Clemen (1986); Trenkler et. al (1986)). Los estudios de Moussourakis (1987), y los presentados en este reporte, indican que se obtienen reducciones notables de las medidas de desempeño más comunes, y más pronósticos perfectos al no restringir los pesos a sumar uno y permitiendolos ser negativos. El que las medidas de desempeño para el caso irrestricto sean de valor a lo más igual a las medidas del caso restringido es una consecuencia de que el problema irrestricto es una relajación del problema restringido. Sin embargo, comparando las medidas de error para un mismo modelo, encontramos que existe una reducción muy notable utilizando los esquemas propuestos.

Las reducciones obtenidas al considerar la dependencia del tiempo de las observaciones en un esquema de programación matemática son notables y van de acuerdo con los resultados de Diebold et. al (1987), aunque ellos utilizan regresión lineal para sus estudios.

Una ventaja de la programación matemática sobre el esquema de regresión, es la posibilidad de considerar más de un objetivo, permitiéndose así involucrar otro tipo de aspectos relacionados con el pronóstico.

Un aspecto que requiere ser estudiado y que puede ser manipulado con la programación lineal con metas múltiples, es el de introducir en el modelo un objetivo que permita disminuir los riesgos generados al proponer un pronóstico muy por encima (o por debajo) del valor verdadero.

BIBLIOGRAFIA

Bates, J.M. and Granger, C.W.J, (1969) "The Combination of Forecasts," Operations Research Quarterly, 20.

Clemen, R.T. and Winkler, R.L, (1986)-a "Combining Economic Forecasts " Journal of Business and Economic Statistics, 4.

Clemen, R.T. (1986)-b "Linear Constraints and the Efficiency of Combined Forecasts," Journal of Forecasting 5, 21-38.

Diebold, E, and Pauly, P. (1987) "Structural change and the Combination of Forecasts", Journal of Forecasting, Vol. 6. 21-40.

Granger, C.W.J., and Ramanathan, R. (1984)-a "Improved Methods of Combining Forecasts", Journal of Forecasting 3, 197-204.

Granger, C. W. J. and Newbold, P., (1977)-b Forecasting Economic Time Series, New York: Academic Press.

Lawrence, M.J, R.H. Edmundson and M.J. O' Connor, (1985) "An Examination of the Accuracy of Judmental Extrapolation of Time Series", International Journal of Forecasting 1 25-35.

Makridakis, S. and Winkler, R.L, (1983)-a "Averages of Forecasts: Some Empirical Results," Management Science , 29.

Makridakis, S. et. al, (1983)-b "The Accuracy of Major Extrapolation (Time Series) Methods. London: Wiley.

Makridakis, S. et. al, (1982)-c The Accuray of Extrapolation (Time Series) Methods: Results of a Foracasting Competition," J. Forecasting, 1 pp 111-153.

Newbold, P. and Granger, C.W.J, (1974) "Experience With Forecasting Univariate Time Series and the Combination of Forecasts," Journal of the Royal Statistical Society, A, 137, 131-165.

Moussourakis, J., (1987) "An Efficient Approach to Combining Forecasts," Manuscrito, Rider College.
Reeves, G.R. and Lawrence, K.D., (1982) "Combining Multiple Forecasts Given Multiple objetives," Journal of Forecasting 1.
Winkler, R.L. and Makridakis, S., (1983) "The Combination of Forecasts," J.R. Statist. Soc. A, 146, Part 2, pp.150-157.
Trenkler, G., and Liski, E., (1986) "Linear Constraints and the Efficiency of Combined Forecasts," Journal of Forecasting 5(3), 197-202.

Cuadro 1.

Periodo t	Valor Y_t	PRONOSTICOS F_{t1}	F_{t2}	F_{t3}	F_{t4}	F_{t5}
1	161	147	161	147	149	151
2	139	153	170	154	153	157
3	137	148	171	149	157	156
4	174	144	169	144	162	155
5	142	156	176	157	166	164
6	141	152	172	152	170	161
7	162	148	167	148	174	159
8	180	154	168	155	178	164
9	164	165	172	167	183	172
10	171	166	172	169	187	173
11	206	170	173	172	191	176
12	193	186	183	191	196	189
13	207	192	189	197	200	194
14	218	202	198	207	204	203
15	229	213	208	219	208	212
16	225	224	220	231	213	222
17	204	230	229	237	217	228
18	227	225	230	231	221	227
19	223	230	236	236	225	231
20	242	232	239	237	230	234
21	239	240	245	245	234	241
22	266	245	249	249	238	245

Cuadro 2.

Criterio	Modelo	Pesos iguales	Pesos exponenciales α=0.1	α=0.2	α=0.3
DMA	Restringido	12.00	0.4943	0.4821	0.5029
	Irrestricto	8.45	0.3136	0.2613	0.2091
EMAP	Restringido	6.646	0.24	0.2101	0.2058
(%)	Irrestricto	4.826	0.249	0.1232	0.0958

Cuadro 3.

Criterio	Modelo	Pesos iguales	Pesos potencia λ=1.	λ=2.	λ=3.	λ=4.
DMA	Restringido	12.00	0.49	0.47	0.46	0.47
	Irrestricto	8.45	0.32	0.28	0.26	0.27
EMAP	Restringido	6.64	2.40	2.15	2.04	2.00
(%)	Irrestricto	4.82	1.82	1.38	1.31	1.12

Cuadro 4.

Crite- rio	modelo	Pesos	w_1	w_2	w_3	w_4	w_5	Numero de pron. exactos
DMA	restr.	iguales	0.0	0.0	0.6	0.4	0.0	2
	irrest.	iguales	11.8	-2.43	-9.84	-0.05	1.73	5
	restr.	α=0.1	0.0	0.0	0.6	0.4	0.0	2
	irrest.	α=0.1	12.31	-2.4	-10.14	-.05	1.59	5
	restr.	α=0.2	0.0	0.0	0.461	0.53	0.0	2
	irrest.	α=0.2	11.98	-2.5	-10.21	-0.19	2.26	5
	restr.	α=0.3	0.0	0.0	0.6	0.4	0.0	2
	irrest.	α=0.3	21.41	1.9	-10.13	6.1	-17.97	5
EMAP	restr.	iguales	0.0	0.0	0.6	0.4	0.0	2
	irres.	iguales	11.89	-2.43	-9.84	-.05	1.71	5
	restr.	α=0.1	0.0	0.0	0.6	0.4	0.0	2
	irres.	α=0.1	13.19	0.86	-3.54	3.41	-12.75	5
	restr.	α=0.2	0.0	0.0	0.6	0.4	0.0	2
	irres.	α=0.2	14.67	-1.38	-9.89	1.10	-3.26	5
	restr.	α=0.3	0.0	0.0	0.6	0.4	0.0	2
	irres.	α=0.3	21.50	1.52	-9.40	5.62	-17.87	5

Cuadro 5.

Criterio	Modelo	Pesos iguales	Pesos potencia (ie) λ=1	λ=2	λ=3	λ=4
DMA	Irrestricto	4.81	0.22	0.21	0.14	0.07
EMAP (%)	Irrestricto	1.37	1.06	0.94	0.71	0.45

UN METODO ECONOMETRICO PARA LA ESTIMA Y JERARQUIZACION DE LA DEMANDA DE DERIVADOS DEL PETROLEO SOBRE LA COSTA MEXICANA DEL PACIFICO A PARTIR DEL ANALISIS FACTORIAL

FRANCISCO CASANOVA DEL ANGEL
INSTITUTO POLITECNICO NACIONAL

Acueducto de Guadalupe No. 125
Col. Acueducto de Guadalupe
07270 México 14 D. F.

RESUMEN.

Se presenta un metodo de análisis para calcular coeficientes de elasticidad y jerarquizar la demanda de cuatro productos derivados del petróleo sobre la costa mexicana del pacífico a partir del tratamiento de la demanda, haciendo uso de la construcción de tablas triangulares de datos. Los coeficientes de elasticidad calculados expresan la relación entre el consumo de derivados de zonas de influencia predeterminadas y el PIB de la región, a partir de lo cual se obtiene el incremento relativo de la demanda entre el año base y su horizonte.

Las funciones que expresan la relación existente entre el consumo de la zona de influencia, a través del puerto de destino, y el PIB de la región, han sido deducidas en función de los factores obtenidos del análisis factorial aplicado a las tablas triangulares de datos.

La metodología de jerarquización se realiza, de la misma manera que para el cálculo de los coeficientes de elasticidad, a partir de las formas triangulares de datos.

0. Introducción.

En esta comunicación se muestra un método de análisis y jerarquización de la demanda de productos siendo estos, en forma particular, cuatro derivados del petróleo sobre la costa mexicana de Océano Pacífico, (en la figura 1 se encuentran identificadas las zonas de influencia, a partir del tratamiento factorialista de la demanda), haciendo uso de la construcción de nuevas formas triangulares de datos.

Para efectuar las proyecciones de la demanda dentro de los planes de desarrollo económico, es necesario apreciar cuantitativamente la influencia del aumento de los ingresos en el consumo del producto. Para esto, es necesario tratar las relaciones entre el ingreso y el consumo a partir de series de tiempo de los recursos nacionales.

1. Estima de las funciones de consumo.

Las funciones de consumo establecen la relación entre el consumo por entidad o zona de influencia y sus ingresos caracterizados por el PIB de la entidad a partir de lo cual se obtiene el incremento relativo de la demanda entre el año base y su horizonte.

En base a las funciones de consumo descritas y desarrolladas en varias publicaciones se han adaptado y experimentado 5 funciones-tipo de consumo, tomando como variables independientes a los factores obtenidos del análisis factorial de correspondencias, las cuales se indican en el cuadro 1. Las funciones consideradas permiten calcular las elasticidades de la demanda. Ellas se prestan al ajuste de la regresión lineal a partir de una sencilla transformación. Claro está que la selección final de la función de consumo depende de parámetros tales como los de ajuste simple y cuadrático, desviación estándar, tests de significancia y verosimilitud de la función dentro de la teoría del consumo; es decir, correspondencia entre los coeficientes de elasticidad de la función en estudio y los estimados con funciones específicas.

De estas funciones, se puede comentar brevemente que la más común es la lineal y que tiene una mala adaptación al ajuste de una función de demanda ya que el coeficiente de elasticidad obtenido tiende a la unidad cuando los ingresos aumentan indefinidamente. En nuestro estudio, los ingresos estan caracterizados por el PIB de la zona de influencia.

La función semilogaritmica, cuyo coeficiente de elasticidad es inversamente proporcional a la cantidad consumida del producto, es muy utilizable cuando el consumo del producto se expresa en cantidades. La función log-inversa da buenos resultados (coeficientes de

elasticidad) cuando se cuenta con una buena estimación del PIB de la zona de influencia lo que permite estudiar la evolución de la cantidad de consumo. Las otras 2 funciones se construyeron e implementaron pero sus primeros resultados fueron un poco anormales por lo que se decidió no tomarlas en cuenta.

	Lineal	
	Original	Transformada
Función tipo	$f(x) = a_0 + \sum_j a_j F_j(x)$	
Coeficiente de elast	$a_j \frac{\sum_j a_j F_j(x)}{f(x)} = \frac{\sum_j a_j F_j(x)}{\sum_j a_j F_j(x) + a_1}$	
Indice de crecimiento		$1 + z \frac{\sum_j F_j f_j(x)}{\sum_j F_j^j(x)}$

	Doble logaritmica	
	Original	Transformada
Función tipo	$f(x) = e^{a_j}(\sum_i F_i(x)\, a_i)$	$Ln(f(x)) = a_i + \sum_i a_i Ln(F_i(x))$
Coeficiente de elast	a_0	
Indice de crecimiento	$(\frac{\sum_i F_i f(x)\, z}{\sum_i F_i^j(x)})$	

	Semi logaritmica	
	Original	Transformada
Función tipo	$e^{f(x)} = e^{a_j}(\sum_i F_i(x)\, a_i)$	$f(x) = a_i + \sum_i a_i\, Ln(F_i(x))$
Coeficiente de elast	$\frac{a_0}{f(x)}$ $\frac{a_0}{a_i + a_0\, Ln(a_i)}$	
Indice de crecimiento	$1 + z\, Ln(\frac{\sum_i F_i f_i(x)}{\sum_i F_i^j(x)})$	

	Logaritmica inversa	
	Original	Transformada
Función tipo	$f(x) = e^{(a_i + a_i / \sum_i F_i(x))}$	$Ln(f(x)) = a_i - (a_0 / \sum_i F_i(x))$
Coeficiente de elast	$-\frac{a_0}{\sum_i F_i(x)}$	
Indice de crecimiento	$e^z(1 - \frac{\sum_i F_i^j(x)}{\sum_i F_i f_i(x)})$	

Continuación .....

	Log-log inversa	
	Original	Transformada
Función tipo	$f(x) = x\, a_j\, e^{(a_1 + a_0 / \Sigma_j F_j(x))}$	$Ln(f(x)) = a_1 + (a_0 / \Sigma_j F_j(x) + \Sigma_j a_j Ln F_j(x))$
Coeficiente de elast:	$- \frac{a_0}{\Sigma_j F_j(x)} + a_0$	
Indice de crecimiento		

Cuadro 1: Funciones que expresan la relación existente entre el consumo de la zona de influencia a través del puerto de destino del producto y el PIB de la región.

2. Los coeficientes de elasticidad como insumo factorialista.

En base a las funciones del cuadro 1, se calcularon las elasticidades y se construyó un arreglo matricial con sus valores-proyección de demanda entre 1984 y el año 2000 para 4 de los principales derivados del crudo en 17 zonas de influencia o puertos de la costa mexicana del Océano Pacífico. Vease la figura 2, donde se marcan los nombres tanto de los derivados del petróleo como los de las zonas de influencia.

3. Metodología de jerarquización.

Se jerarquizó la tabulación de los coeficientes de elasticidad mediante una tabla de Burt a partir de una forma disjunta completa, pero los resultados jamás fueron satisfactorios por lo que se decidió construir una tabla de bloques unidimensionales a partir de un arreglo transpuesto de la descripción lógica de los coeficientes de elasticidad de la demanda, pero como aquella resulta simétrica se construyeron arreglos, a los que se les denominó arreglos triángulares de Burt, uno superior y otro inferior de bloques unidimensionales que resultaron muy idóneos para este tipo de estudios.

4. Descripción teórica de los arreglos tabulares.

A continuación, se muestra la construcción teórica de las tablas disjunta completa transpuesta de bloques unidimensionales y las triangulares superior e inferior de Burt-unidimensionales con algunos comentarios.

A partir de una tabla disjunta completa $k(i,q)$ sobre $I \times Q$ es posible construir una tabla transpuesta de descripción lógica en la cual las modalidades de respuesta se convierten en individuos. De manera que:

$$\forall\, i \in I,\ \forall\, q \in Q \quad \text{existe } j \in Q \text{ tal que:}$$

$$(k(i,j) = 1 \wedge (j' \in q;\ j' \neq j) \Rightarrow k(i,j') = 0$$

es decir, para todo elemento i en el conjunto de individuos I y para toda modalidad q en el subconjunto de modalidades Q, existe un elemento j en el subconjunto de las modalidades Q tal que el elemento tabular $k(i,j)$ es igual a 1 y para cualquier otro elemento j' en la modalidad q, con el elemento j' diferente del elemento j implica que el elemento tabular $k(i,j')$ sea cero.

Otra forma de decirlo es la siguiente: cada individuo i posee en cada clase q una y sólo una propiedad j.

A partir de la transpuesta de una tabla lógica k_{IJ} denominada

$K_{II}(i_s, i_p) = \text{Card}\{k(j,i_s) = k(j,i_p) = 1 \quad j,s \in J \ \forall\ p = 1,\ldots,n\}$

$K_{II}(i_s, i_p) = k_{BC}(i_s, i_p)$ es el número de elementos de Q que poseen simultáneamente las propiedades i_s, i_p. El subíndice BC de k significa "Burt completa".

La tabla k_{IxI} es simétrica en relación a su diagonal principal por lo que su descomposición se puede hacer en una matriz triangular superior y en una matriz triangular inferior de la siguiente manera:

$$k_{BC}(i_s, i_p) = k_{TI}(i_s, i_p) + k_{TS}(i_s, i_p) - \{k(i_s, i_p) \mid i_s = i_p\}$$

donde k_{Ti} es la tabla tringular interior y k_{TS} es la tabla tringular superior de la tabla de Burt completa de bloques unidimensionales k_{BC} y como la diagonal de repite, es necesario restarsela para obtener la matriz de Burt de bloques unidimensionales.

En relación a los arreglos triangulares obtenidos a partir de una tabla de Burt de bloques unidimensionales, tenemos que a partir de la transpuesta de una tabla lógica k_{IJ} denominada $k'_{IJ} = k_{JI}$ es posible construir una tabla triangular inferior de Burt de bloques unidimensionales de la siguiente manera:

$$k_{TI}(i_s, i_p) = \text{Card}\{k(j,i_s) = k(j,i_p) = 1 \mid j,s \in J \ \forall \ p = 1, \ldots, k\}$$

$k_{TI}(i_s, i_p)$ es el número de elementos de Q o J que poseen simultáneamente la propiedad i_s, i_p y con las mismas caracteristicas que una matriz de bloques unidimensionales, es decir; no existen diagonales ni extradiagonales de mas de un sólo elemento. Nótese que los valores de la diagonal son los valores de las masas $k(j) = k(i)$.

Los elementos arriba de la diagonal son todos cero pero al algoritmo se le puede invertir y obtener una matriz triangular superior de Burt-unidimensional.

A partir de la transpuesta de una tabla lógica k_{JI} es posible construir una tabla triangular superior de Burt-unidimensional de la siguiente manera:

$$k_{TS}(i_s, i_p) = \text{Card}\{k(j,i_s) = k(j,i_p) = 1 \mid j,k \in J \quad p = k, \ldots, n\}$$

$k_{TS}(i_s, i_p)$ es el número de elementos de Q o J que poseen simultáneamente la propiedad i_s, i_p y con las mismas caracteristicas que una matriz de bloques unidimensionales, es decir; no existen diagonales ni extradiagonales de más de un sólo elemento. Nótese que los valores de la diagonal son los valores de las masas, y que los elementos abajo de la diagonal son todos cero.

Los algoritmos de construcción de las tablas triangulares inferior y superior de Burt a partir de la tabla de bloques unidimensionales difiere únicamente en la toma de valores del subíndice p.

5. Descripción factorialista.

Los parámetros que caracterizan la proyección de las zonas de influencia y la demanda de los cuatro derivados del petróleo al año 2000 en la costa mexicana del Océano Pacífico se obtiene en pocos ejes factoriales.

Los valores propios de dichos ejes son: 0.49169, 0.08537, 0.04533 y 0.02430 con los siguientes porcentajes de inercia 69.17, 12.01, 6.37 y 3.42 respectivamente.

La forma que presenta el primer plano es parabólica, figura 2, característica del proceso de corte en clase, plasmado en una discriminación de clases y principalmente de los derivados del petróleo. En la parte positiva del primer eje encontramos a la gasolina, el diesel y un poco de combustóleo con elasticidades mínimas de 3.0 a 4.0 y máximas de 7.0 a 8.0 para las zonas de Los Mochis, Manzanillo, Navojoa, Ciudad Obregon y la Paz. Para Acapulco y Mexicali en relación al diesel y para La Paz respecto al combustóleo. Para elasticidades que van de 7.0 a 13.0 se tiene únicamente a Guaymas, Mazatlán y Rosarito en relación al diesel. También se presentan aquí las altas elasticidades, aquellas que van más allá de 18.0 en zonas de influencia tales como Rosarito y Mexicali para la gasolina y para Manzanillo en combustóleo.

Como se habrá notado, algunas elasticidades parecen ser bastante grandes pero van de acuerdo a las nuevas zonas industriales que el gobierno mexicano tiene en desarrollo y en perspectiva donde antes no se surtía alguno de estos productos pero si en relación a las necesidades que la industria en crecimiento demanda.

En el segundo eje tenemos a las mismas zonas de influencia con los tres derivados del petróleo en las cuales se ha hecho mención anteriormente, cuyas elasticidades estan por debajo del 2.8 y del 1.0 principalmente en zonas turísticas y urbanas perfectamente cimentadas para productos tales como el diesel en Acapulco, Mazatlán y Manzanillo o combustóleo en Acapulco.

Todo parece indicar una discriminación ya no entre clases sino entre productos derivados y zonas de influencia caracterizadas en turísticas e industriales así como en rangos de elasticidad. Si bien, el esfuerzo del gobierno mexicano por desensentivar el consumo de la gasolina a través de constantes aumentos al precio aumentando la

producción de los derivados propios al transporte de carga y a la industria han tenido buenos resultados este plano manifiesta que a partir de las politicas actuales, la declinación del consumo interno de la gasolina será a principios de la decada de los noventa.

De lo anterior no es nada casual el tener un tercer eje con derivados propios a las nuevas zonas de influencia o industriales en esa parte del país. Lugares tales como Libertad, Rosarito y Topolobampo con una gran elasticidad de la demanda del combustóleo.

6. La jerarquización de los derivados del petróleo.

Con el fin de ver realmente la jerarquía de la demanda de cada derivado del petróleo respecto a las zonas de influencia consideradas en la costa mexicana del Océano Pacífico se cortaron en clases los años-proyección. Se invirtió la matriz y se aplicaron los algoritmos descritos en la sección 4. Se hizo el análisis factorial por derivado como elementos principales, poniendo en suplemento a los restantes tres productos. Los histogramas de valores propios tienen caidas o relaciones sucesivas de valor a valor decreciente, muy proporcionales. El primer valor propio está alrededor del 0.48 así como el 80% de los datos en los primeros cinco ejes factoriales.

Los arboles jerárquicos para los cuatro productos en estudio: gasolina, diesel, combustóleo y kerosenas dados en las figuras 3, 4, 5 y 6 se discriminan en altas y bajas elasticidades asi como en sectores turistico e industrial principalmente.

En el país existen actualmente dos clases de gasolina, la llamada Pemex Nova y la Pemex Extra, siendo la primera la de mayor producción así como la de menor precio. Su discriminación jerárquica es en altas y bajas elasticidades en relación a las zonas de influencia ya mencionadas, según se muestra en la figura 3. Su agrupación está ligada a porcentajes de participación que el PIB del sector turismo tiene con el PIB total nacional medidos en miles de millones de pesos mexicanos de 1960.

Dentro de esa manifiesta discriminación se forman dos grandes clases para cada rango de elasticidad, zonas portuarias de influencia y zonas interiores de influencia basadas en el turismo y en el comercio.

En el rubro "otros productos refinados" relativos a la producción de petróleo y sus derivados, el diesel ocupa en promedio el 38.5% entre 1970 y 1983 así como el 17.5% del total de derivados obtenidos. Muestra también una discriminación jerárquica en altas y bajas elasticidades con una formación a su interior de dos grandes clases en cada rango de elasticidad, denominadas de consumo bajo y estándar, vease la figura 4. Debido a la eliminación y agrupación de clases se

presentan zonas de baja elasticidad dentro de las de alta elasticidad, tales como Colima, Rosarito y Culiacán pero dentro de la rama de bajo consumo.

El combustóleo, vease la figura 5, tiene una participación del 52% como promedio entre 1970 y 1983 en el rubro "otros productos refinados", con una discriminación en altas y bajas elasticidades, pero su agrupación es en relación al PIB de los sectores turismo e industrial.

El último derivado en estudio, kerosenas vease la figura 6, tiene una participación del 8% en el periodo considerado en el rubro ya mencionado. Este producto es una agregación del tractogas, del tractomex, del diáfano, de turbosinas y de la kerosena cruda. La mayoria de sus elasticidades estan entre 0.0 y 1.0 salvo para Acapulco, Ciudad Obregón y Rosarito haciendo que el patrón de altas y bajas elasticidades se conserve. Se nota cierta relación con algunas zonas de influencia relativas al diesel.

Para obtener una buena jerarquización de los años de demanda se construyó una matriz triangular inferior de Burt-unidimensional a partir de una matriz transpuesta de descripción lógica de las zonas de influencia. Acumula en cinco factores el 83.5% de la inercia de la nube contra el 99.9% en el mismo espacio de dimensión cinco pero a partir de un arreglo completo de bloques unidimensionales, cuyos primeros valores propios son muy débiles; del orden de 0.10 y 0.001 respectivamente.

La agregación se hace en tres grandes clases, la primera tiene a casi el tiempo restante del actual gobierno de la República así como a los dos primeros años del próximo gobierno, véase la figura 7. Probablemente se deba a que históricamente en el último año de gobierno 3/ y en los dos primeros se replantean y revisan los logros, metas y objetivos alcanzados y por cumplir; es decir, se programa para que el poder ejecutivo en funciones tenga mínimo cuatro años de realizaciones plenas. Las otras dos clases contienen a la consolidación del poder ejecutivo que estará en funciones entre 1988 y 1994 asi como al que terminará el actual siglo XX.

Las dos últimas clasificaciones jerárquicas que se muestran son: para la figura 8, la jerarquía de los puertos del Pacífico Mexicano en relación a la proyección de los cuatro derivados del petróleo en estudio a partir de los datos iniciales. En la figura 9, una jerarquía de los años en proyección de dichos derivados petroliferos.

7. Conclusiones.

Las funciones factoriales de consumo que expresan la relación entre el

consumo por entidad y su PIB dan a la postre muy buenos coeficientes de elasticidad en base a las funciones de transformación.

Después de muchos intentos por encontrar una buena agregación de cada derivado en la costa mexicana del Océano Pacífico, se ideo la construcción de arreglos triángulares de Burt a partir de arreglos transpuestos disjuntos completos cuyos factores jerarquizan y discriminan las variables o los individuos en estudio, con una formación tipo bloques sobre rangos de elasticidad y zonas de influencia basada en los ingresos o producto interno bruto de dichas zonas.

Los resultados tenidos a partir de la aplicación de los factores obtenidos de los arreglos triangulares en la jerarquización de las zonas de influencia estudiadas y de los años de demanda, proyectan las políticas sexenales que reafirman la soberanía tanto de la nación sobre este recurso no renovable, como de la administración en turno sobre la forma de explotación, refinación y su distribución.

Desde el punto de vista teórico, los arreglos triangulares implementados muestran una relación entre variables o entre individuos, según sea el caso, nada diferente la una de la otra ya que las nubes de puntos $N(I_{TI}) = N(J_{JT})$ y $N(I_{TS}) = N(J_{TI})$ construidas por el análisis factorial estan en el mismo espacio y son transpuestas por lo que la relación e imagen de resultados transpuestos afirman la coherencia de los datos. La interpretación de factores es más dinámica que la realizada a partir de la tabla de bloques unidimensionales ya que existe en ellas más claridad en sus correlaciones además de que los cálculos se efectúan a mayor velocidad lo cual disminuye el tiempo de cálculo.

Queda solamente esperar la aplicación de la técnica propuesta y de las formas tabulares construidas en otro tipo de variables para ver su verdadera valía.

Remarcas:

1/ Véase referencias 2 y 9.

2/ Un análisis factorial sobre este tipo de matrices benefician o fortifican en cierta medida a las n-ésimas variables de una matriz triangular ya sea superior o inferior de Burt-unidimensional.

3/ Los datos trabajados inicialmente son series históricas de la producción real de los derivados y obtenidos de las publicaciones oficiales de Petróleos Mexicanos.

BIBLIOGRAFIA

1. Benzécri, J-P. "Sur le généralisation du tableau de Burt et son analyse par bandes". Les cahiers de L'analyse des Données, vol III, No. 1, pag 33-43, 1982. Paris Dunod.

2. Cramer, J. S. Econometría empírica. Fondo de cultura económica. México, 1973.

3. Casanova del Angel, F. "Un método para pronosticar valores". Acta Mexicana de Ciencia y Tecnología, I.P.N., vol II, No. 6, abril-junio 1984, pag 69-84. México.

4. Casanova del Angel, F. "El análisis factorial y sus enfoques". Boletín de Graduados e Investigación del I.P.N., vol I, No. 4, enero-febrero 1983. México.

5. Cazes, P. "L'analyse de certain tableaux rectangulaires décomposé en blocs: Généralisation des propiétés rencontrés dans l'étude des correspondances multiples. I Definitions et applications a l'analyse canonique des variables qualitatives". Les cahiers de L'analyse des Données, vol V, No. 2, pag 129-130, 1980. Paris Dunod.

6. Centre de Statistique et Informatique Appliquees. Bulletin Technique, vol 2, 1984, Nos. 1-2. Paris France.

7. Goreux, L. M. "Ingresos y consumo de alimentos". Estudios de la FAO sobre economía y estadísticas agrícolas. 1952-1977. Roma 1978.

8. Mbuvil, M. et Maravalle, M. "Typologie des huiles brutes de petrole". Les cahiers de L'analyse des Données, vol IX, No. 3, pags 301-314, 1984. Paris Dunod.

9. Mbuvil, M. "Sur l'evolution de la production mondiale du petrole brut de 1918 a 1979: Trafic mondial du petrole brut de 1970 a 1978. Un exemple d'effect Guttman double". Les Cahiers de L'analyse des Données, vol IX, No. 1, pags 59-66, 1984. Paris Dunod.

10. Secretaría de Agricultura y Recursos Hidráulicos. "El desarrollo agropecuario de México: Pasado y Perspectivas". Informe 1982. Tomo XIII. "Perspectivas de la demanda y de la oferta de productos agropecuarios".

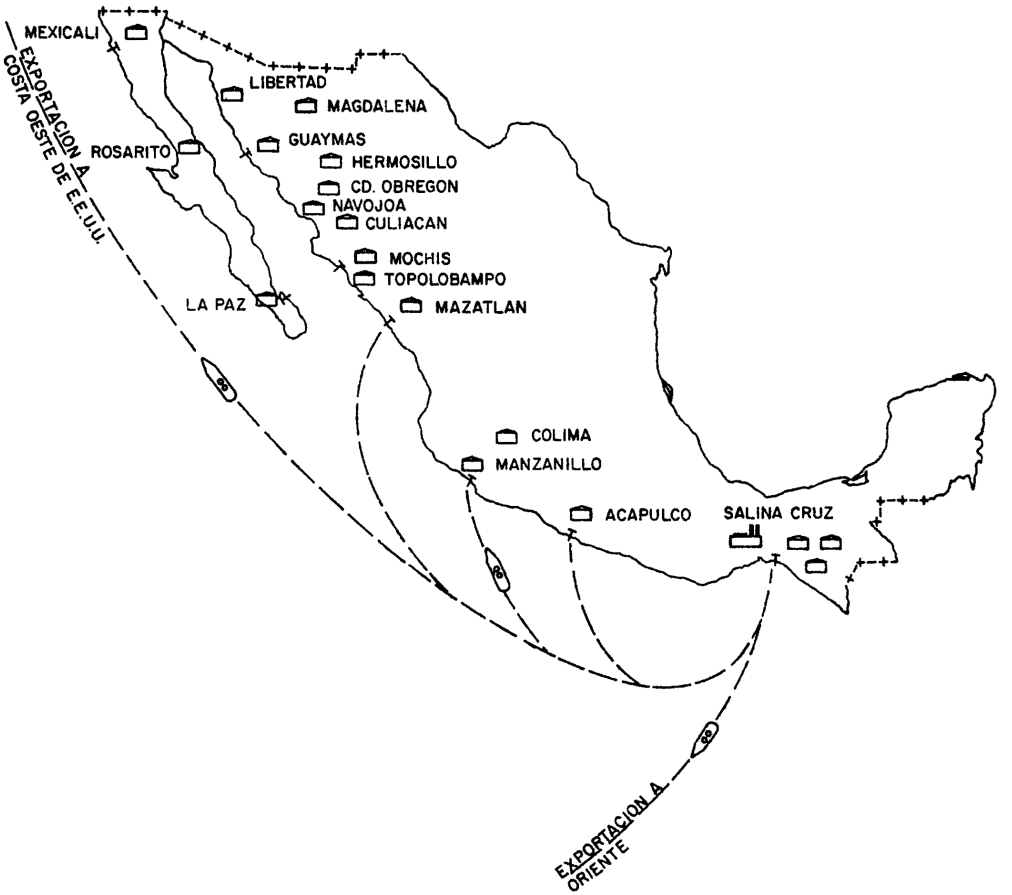


FIGURA 1 : ZONAS DE INFLUENCIA DE DEMANDA DE LOS DERIVADOS DEL PETROLEO EN LA COSTA MEXICANA DEL PACIFICO

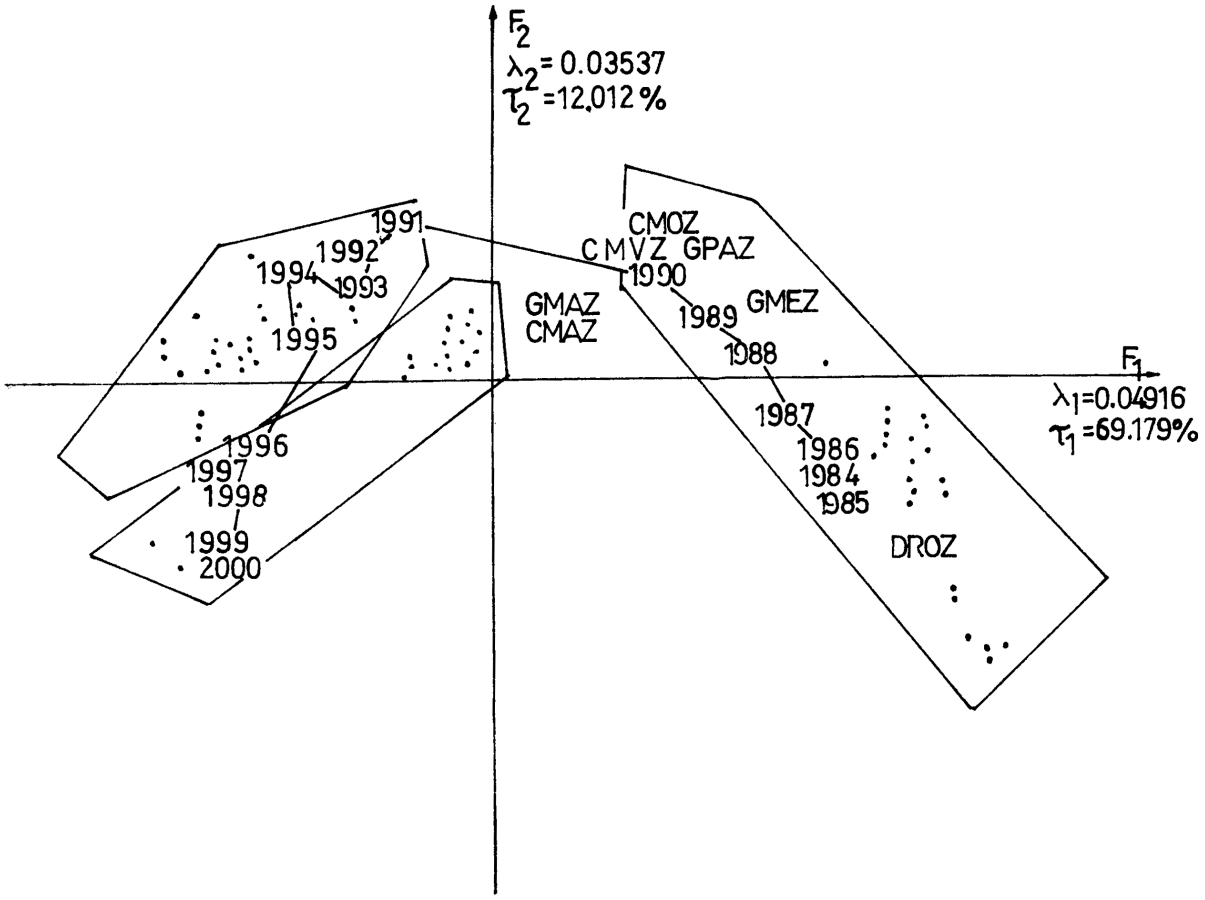


FIGURA 2. Plano factorial 1-2 de los años producción. Obtenida de una tabla transpuesta.

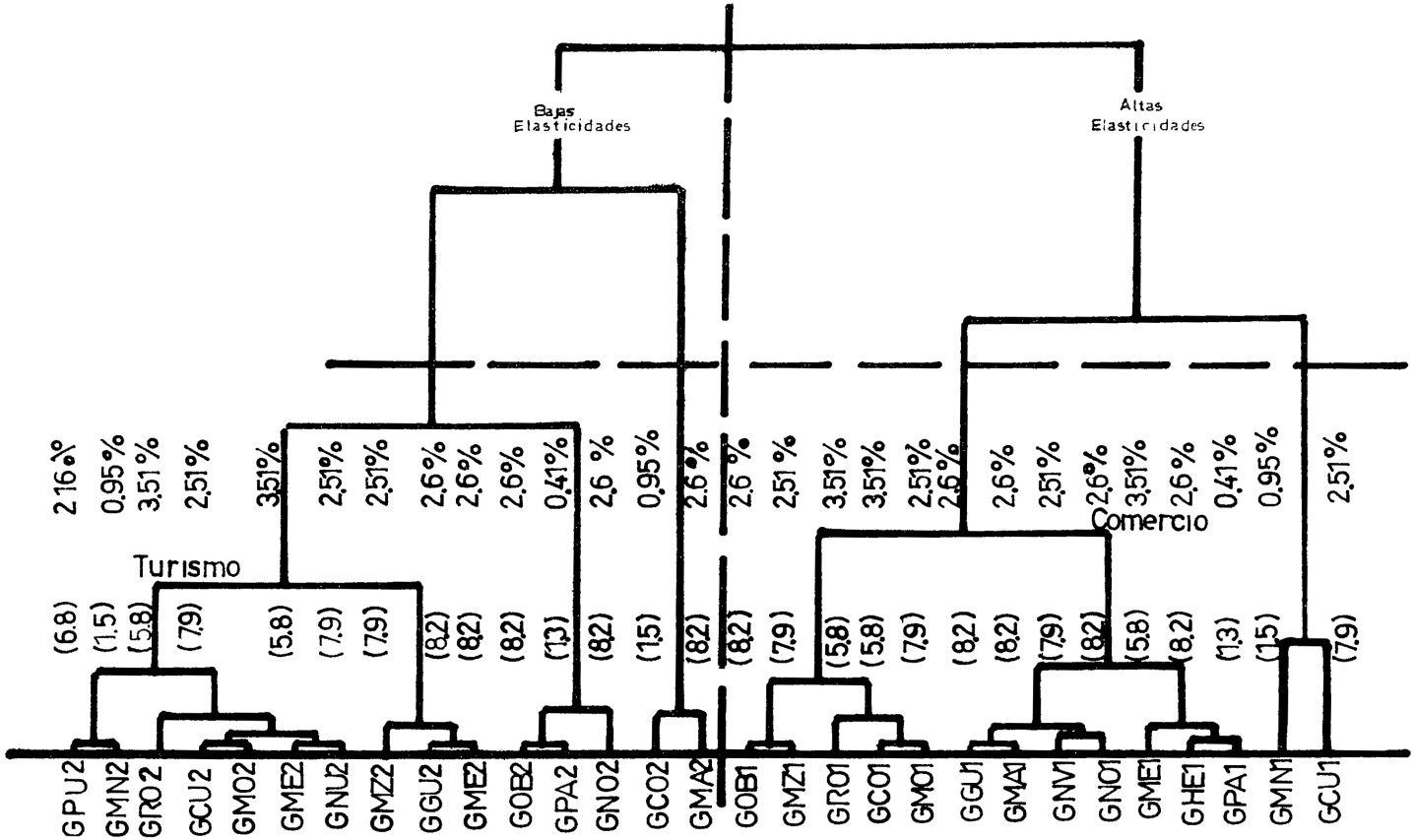


FIGURA 3: AGREGACION JERARQUICA DE LAS GASOLINAS.
El valor entre parentesis indica el PIB del sector servicios estatal en miles de millones a pesos de 1960. El porcentaje es el relativo al total nacional en el sector servicios.

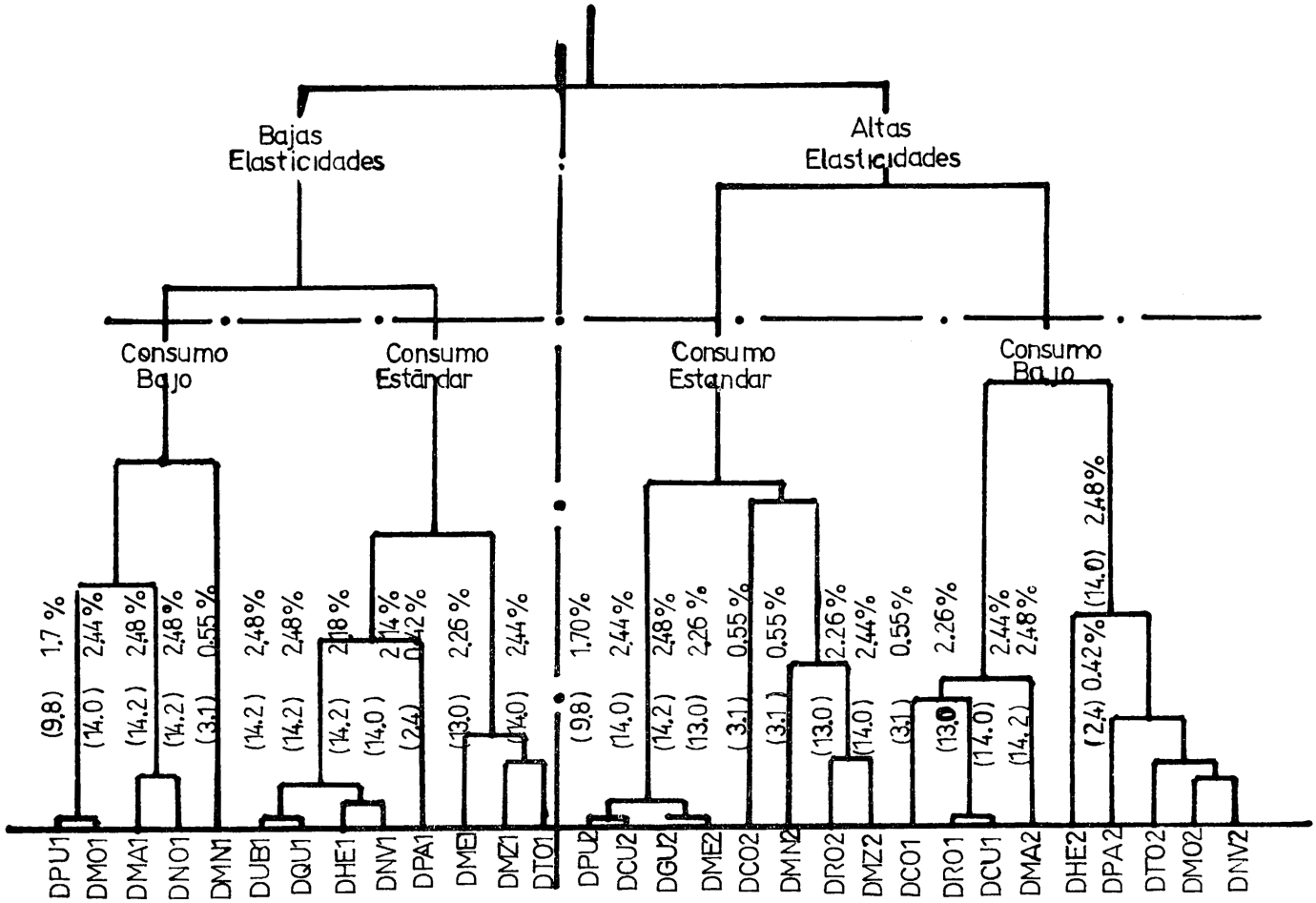


FIGURA 4 AGREGACION JERARQUICA DEL DIESEL.
El valor entre parentesis indica el PIB del sector servicios estatal en miles de millones a pesos de 1960. El porcentaje es el relativo al total nacional en servicios.

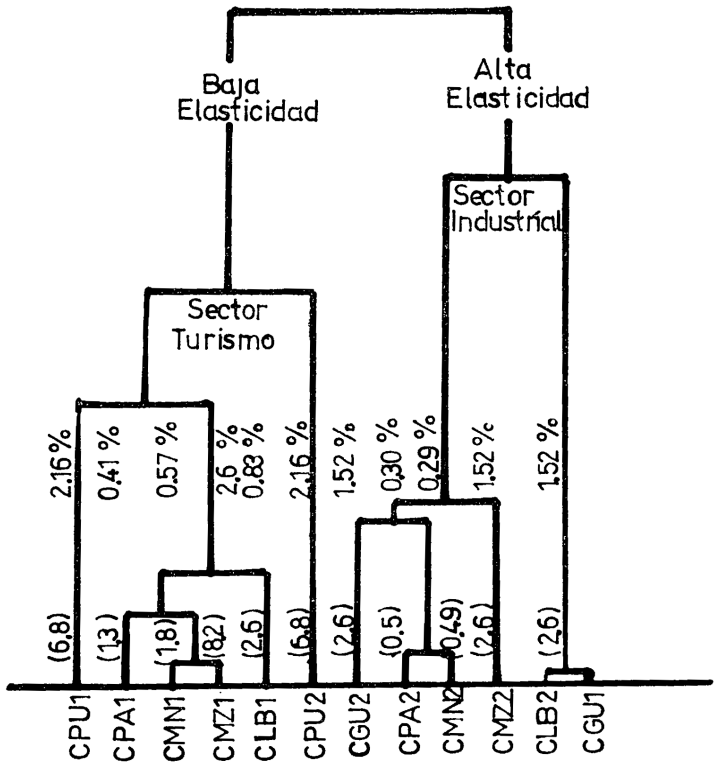


FIGURA 5 AGREGACION JERARQUICA DEL COMBUSTOLEO.
El valor entre parentesis indica el PIB relativo al sector indicado en el arbol y estatal en miles de pesos de 1966.

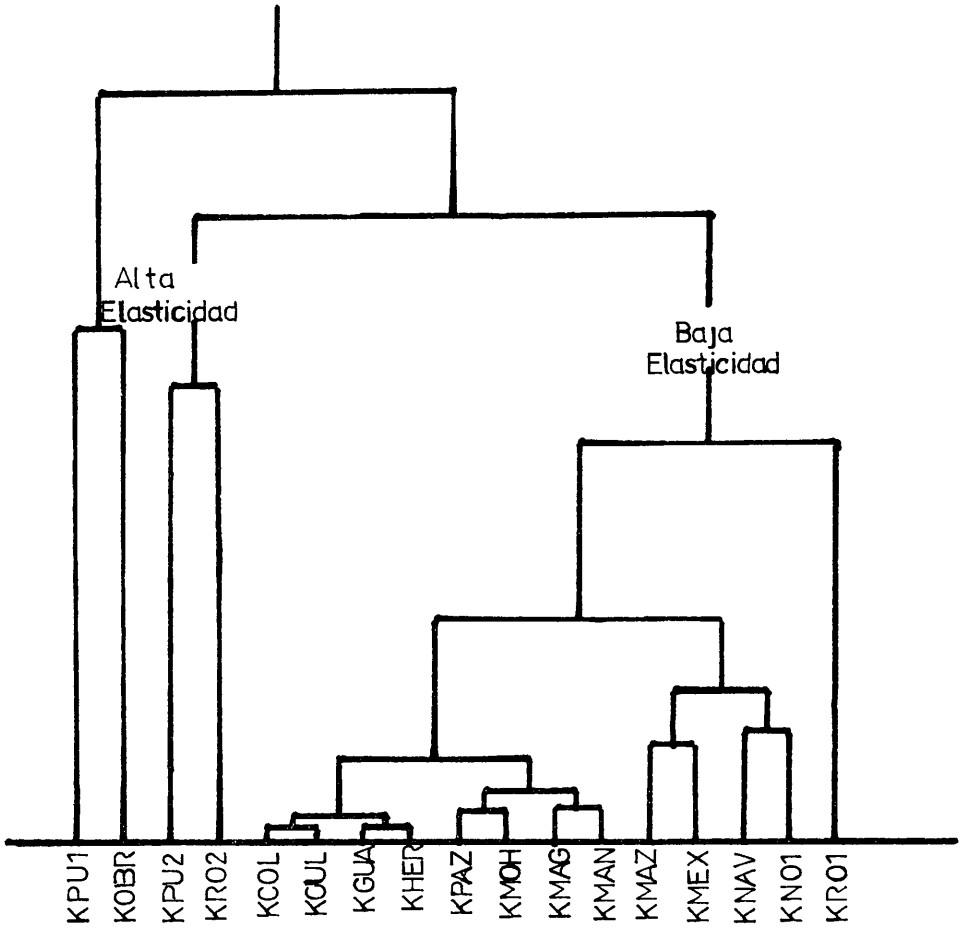


FIGURA 6 AGREGACION JERARQUICA DE LAS KEROSENAS.

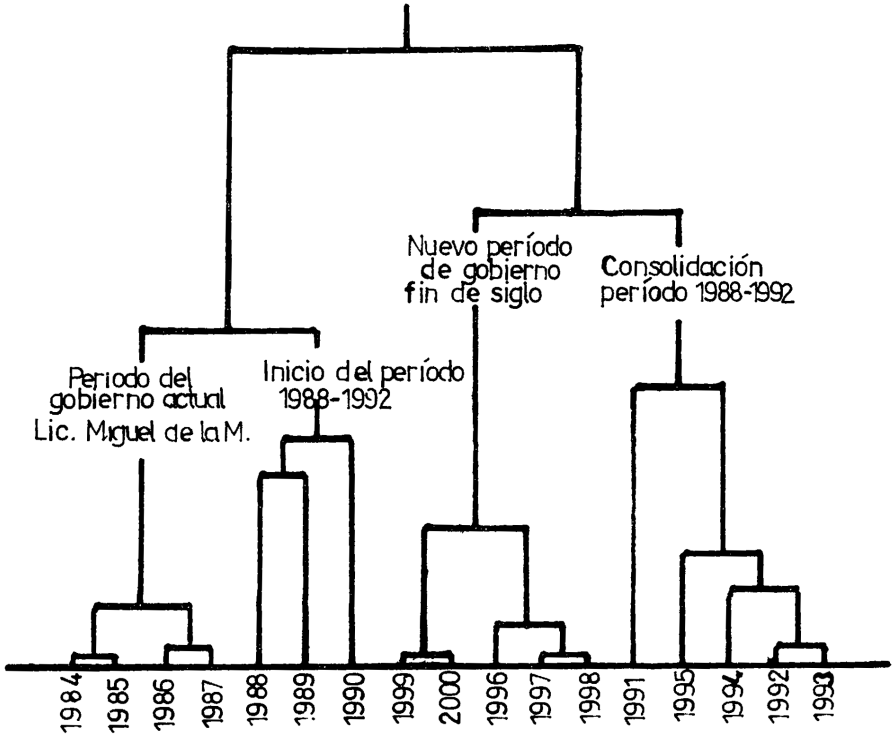


FIGURA 7 JERARQUIA DE LA MATRIZ BURT-UNIDIMENSIONAL DOBLE.

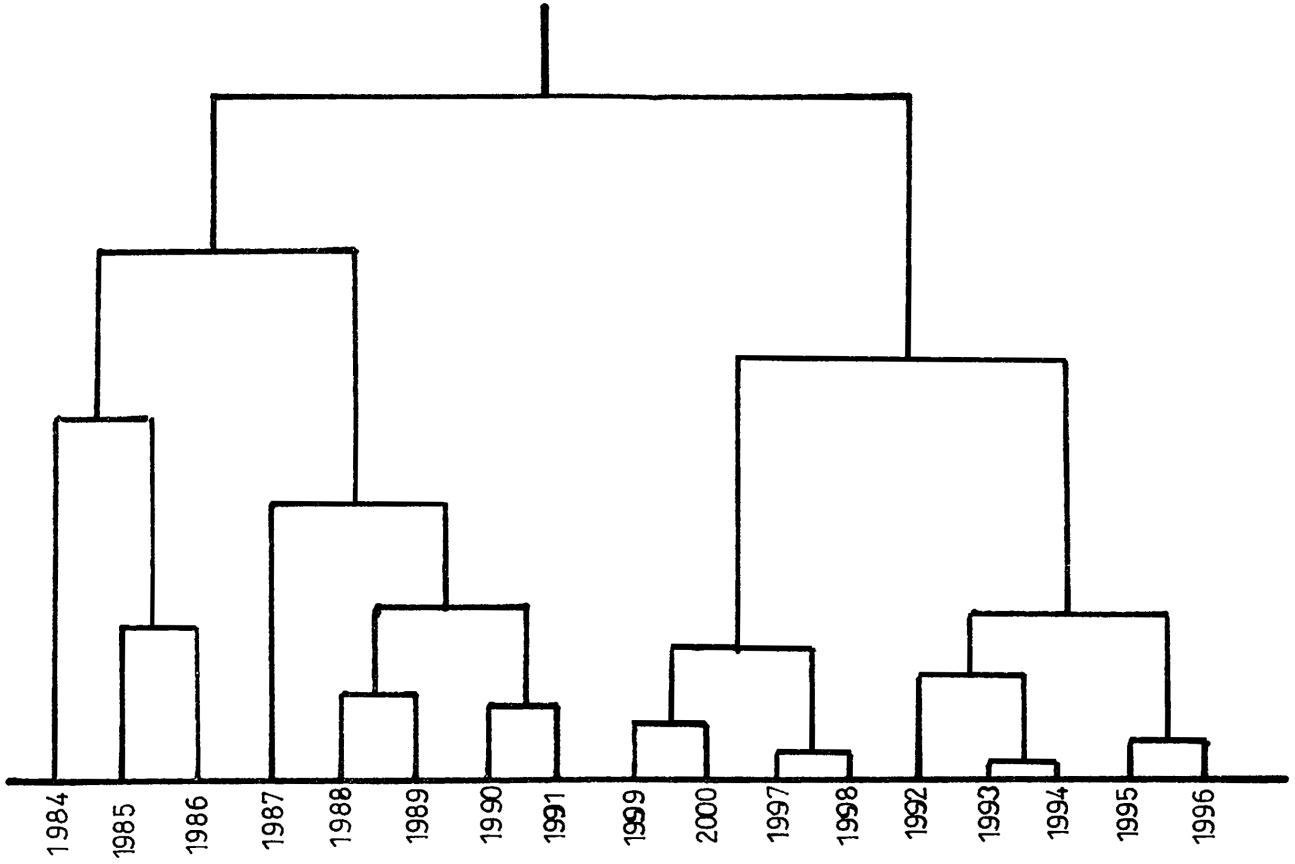


FIGURA 8: JERARQUIA DE LOS AÑOS PROYECCION DE 4 DERIVADOS PRINCIPALES DEL PETROLEO EN EL PACIFICO MEXICANO.

ANALISIS BAYESIANO DE UN ENSAYO DE COCIENTE DE PENDIENTES GENERALIZADO

Mendoza M.

Departamento de Matemáticas, Facultad de Ciencias UNAM
Cd. Universitaria, 04510 México , D. F.

RESUMEN

En este trabajo se presenta una propuesta para el análisis bayesiano del parámetro de interés en una clase de modelos de uso común en el estudio de ensayos biológicos. Esta clase contiene como caso particular el muy conocido ensayo de cociente de pendientes y permite además, la consideración de algunas variantes en que la relación dosis-respuesta no es lineal. Los resultados se obtienen para una familia de distribuciones iniciales que incluye distintas variantes de las llamadas distribuciones mínimo-informativas.

1. INTRODUCCION

En un trabajo previo (Mendoza 1987) se ha considerado un posible análisis bayesiano para el modelo conocido como de cociente de pendientes. Este modelo describe una situación en la que se realiza un experimento a través del cual p dosis $(X_{11},\ldots,X_{1p})$ de un primer estímulo y q dosis de un segundo estímulo $(X_{21},\ldots,X_{2q})$ se aplican para obtener un conjunto $\{Y_{1jk};\ j=1,\ldots,\ p;\ k=1,\ldots,n;\ Y_{2jk}\ ;\ j=1,\ldots,\ q\ ;\ k=1,\ldots,\ n\}$ de $n(p+q)$ observaciones condicionalmente independientes con distribución Normal de varianza σ^2 y tales que

$$E(Y_{1jk}) = \alpha + \beta X_{1j} \qquad k=1,\ldots,n\ ;\ j=1,\ldots,p$$

$$E(Y_{2jk}) = \alpha + \rho\beta X_{2j} \qquad k=1,\ldots,n\ ;\ j=1,\ldots,q \tag{1}$$

Para establecer la potencia relativa de los dos estímulos, basta con producir inferencias sobre el cociente de pendientes $\rho = \rho\beta/\beta$ que contiene toda la información asociada al cociente de cualquier par de dosis equivalentes (dosis, del primer y segundo estímulo respectivamente, que producen la misma respuesta esperada). El procedimiento descrito en Mendoza (1987) se ocupa específicamente de la obtención y el análisis de una distribución mínimo-informativa para ρ, utilizando la técnica propuesta por Bernardo (1979). Entre los resultados que ahí se presentan, se puede mencionar que

si bien la distribución inicial conjunta es impropia, la correspondiente distribución final marginal para ρ es propia con probabilidad uno. Además, es posible establecer una expresión analítica para la densidad final marginal del parámetro de interés y se pueden calcular las modas (a lo más dos). La aplicabilidad de estos resultados tiene limitaciones evidentes. Por una parte, es natural la extensión al caso en que el efecto de las dosis es no lineal y en segundo lugar, queda planteada la posibilidad de generalizar las propiedades de la distribución final marginal cuando se utiliza otra espécificación para la inicial.

En este artículo, se considera una extensión del modelo (1) donde las n(p+q) observaciones son condicionalmente independientes, siguen una distribución Normal con varianza σ^2 y son tales que

$$E(Y_{1jk}) = \alpha + \beta(X_{1j})^{\lambda} \qquad k=1,\ldots,n \;;\; j=1,\ldots,p$$

$$E(Y_{2jk}) = \alpha + \rho\beta(X_{2j})^{\lambda} \qquad k=1,\ldots,n \;;\; j=1,\ldots,q \qquad (2)$$

Esta estructura generaliza el modelo de cociente de pendientes a través del parámetro (positivo) λ y en particular, si $\lambda=1$, reproduce el modelo original. Como es usual, y sin pérdida de generalidad, se incorpora la hipótesis de dosis no negativas ($X_{ij} \geq 0$). De la misma forma en que para el modelo (1) se puede probar fácilmente que la potencia relativa de los estímulos está descrita por el valor de ρ, cuando se considera la estructura (2) es inmediato establecer que la potencia relativa queda determinada por el parámetro de interés $\phi = \rho^{1/\lambda} = (\rho\beta/\beta)^{1/\lambda}$. En general, λ es desconocido y entonces ϕ depende de ρ, el cociente de pendientes, y de λ. En lo que resta de este trabajo, y como una primera aproximación, λ se considera conocido y el análisis de ϕ se desarrolla condicional a λ. En la sección 4 se introduce una aproximación para extender los resultados al caso en que λ es desconocido. En cualquier forma, y también sin pérdida de generalidad, es posible suponer que ρ es positivo en cuyo caso el parámetro de interés está siempre bien definido y es una función uno a uno de ρ.

2. ANALISIS BAYESIANO.

Para producir inferencias bayesianas sobre ϕ, el parámetro de interés, la información obtenida a través del experimento (totalmente contenida en la función de verosimilitud) debe combinarse, vía el torema de Bayes, con la información inicial disponible acerca de los parámetros. Esta información a priori se incorpora a través de la correspondiente densidad inicial y el resultado es la densidad final (o posteriori)

conjunta a partir de la cual se puede obtener, utilizando la transformación e integración adecuadas, la densidad final marginal para ϕ. Explícitamente, si D representa el banco de datos experimentales y se considera λ conocido, se tiene que

$$p(\alpha,\beta,\rho,\sigma \mid D) \propto p(\alpha,\beta,\rho,\sigma)\ p(D \mid \alpha,\beta,\rho,\sigma)$$

y a partir de la conjunta $p(\alpha,\beta,\rho,\sigma \mid D)$ debe de obtenerse la marginal $p(\phi \mid D)$ haciendo uso de la relación $\phi = \rho^{1/\lambda}$. En cualquier caso, la función de verosimilitud para la estructura experimental descrita en (2) está dada por

$$p(D \mid \rho,\alpha,\beta,\sigma) \propto \sigma^{-n(p+q)} \exp\{-[\ \Sigma_j\Sigma_k (Y_{1jk} - \alpha - \beta W_{1j})^2 + \Sigma_j\Sigma_k (Y_{2jk} - \alpha - \rho\beta W_{2j})^2]/(2\sigma^2)\}. \qquad (3)$$

en donde, por facilidad, $W_{ij} = (X_{ij})^{\lambda}$. Para alimentar el mecanismo bayesiano sólo resta la asignación de la densidad inicial conjunta $p(\alpha,\beta,\rho,\sigma)$.

De acuerdo al paradigma bayesiano, la densidad inicial describe el conocimiento con que cuenta el científico sobre los parámetros antes de realizar el experimento. Este conocimiento puede ser de naturaleza subjetiva y personal o puede ser producto de la evidencia experimental previa. Más aún, la densidad inicial puede asignarse de manera que influya poco en los resultados finales. Esto es, puede utilizarse una densidad inicial que permita que los datos experimentales, a través de la verosimilitud, dominen las conclusiones. Este tipo de densidades reciben el nombre de mínimo-informativas o de referencia. En cualquier caso, la representación funcional de la densidad inicial, sin alcanzar trascendencia conceptual, puede resultar de importancia técnica para la obtención de resultados analíticos. En el caso del modelo (2), y en tanto ϕ es una función biyectiva de ρ, puede considerarse que ρ es el parámetro de interés para escribir explícitamente

$$p(\alpha,\beta,\rho,\sigma) = p(\alpha,\beta,\sigma \mid \rho)\ p(\rho) \qquad (4)$$

Ahora bien, puesto que el interés se concentra en ρ y por tanto α, β y σ son parámetros marginales o de ruido , resulta conveniente utilizar una densidad condicionalmente mínimo-informativa para aproximar la densidad $p(\alpha,\beta,\sigma \mid \rho)$. Condicionando a un valor de ρ fijo, la estructura (2) coincide con un modelo de regresión lineal simple y la aplicación de los métodos más comunes para obtener iniciales mínimo-informativas, como por ejemplo el límite de familias conjugadas (DeGroot 1970, cap. 9), la regla propuesta por Jeffreys (Box y Tiao 1973, sec. 1.3) o el análisis de

referencia (Bernardo 1979), produce una expresión del tipo

$$\pi(\alpha,\beta,\sigma \mid \rho) \propto \sigma^{-r} \tag{5}$$

para algún $r > 0$. En consecuencia, una aproximación razonable para la densidad inicial marginal de α,β,ρ,σ puede ser

$$p(\alpha,\beta,\rho,\sigma) \propto \sigma^{-r} p(\rho) \tag{6}$$

en donde $r > 0$ y $p(\rho)$ es la densidad inicial marginal de ρ que describe el conocimiento que el científico posee acerca de ρ antes de efectuar el experimento. Es claro que si se emplea una distribución inicial conjunta de este tipo, la marginal $p(\rho)$ puede ser a su vez mínimo-informativa o bien puede describir la información personal del científico. En cualquier caso, la correspondiente conjunta final está determinada por la expresión

$$\begin{aligned} p(\alpha,\beta,\rho,\sigma \mid D) &\propto \{\sigma^{-r} p(\rho)\}\sigma^{-n(p+q)} \exp\{-[\Sigma_j\Sigma_k(Y_{1jk} - \alpha - \beta W_{1j})^2 \\ &\quad + \Sigma_j\Sigma(Y_{2jk} - \alpha - \rho\beta W_{2j})^2]/(2\sigma^2)\}. \\ &= p(\rho)\,\sigma^{-M} \exp\{-[\Sigma_j\Sigma_k(Y_{1jk} - \alpha - \beta W_{1j})^2 \\ &\quad + \Sigma_j\Sigma_k(Y_{2jk} - \alpha - \rho\beta W_{2j})^2]/(2\sigma^2)\}. \end{aligned} \tag{7}$$

con $W_{ij} = (X_{ij})^{\lambda}$ y $M = r+n(p+q)$. Para obtener la densidad final marginal de ρ basta con integrar en (7) con respecto a α,β y σ. La forma analítica de esta inicial conjunta facilita este proceso. Si se integra respecto a σ, es fácil reconocer el núcleo de una densidad Gama invertida y si posteriormente se considera el caso de α y β, nuevamente sin dificultad se identifica el núcleo de una densidad t de Student bivariada. En resumen, la densidad $p(\rho \mid D)$ se puede obtener sin mayores dificultades y el resultado es el siguiente

$$\begin{aligned} p(\rho \mid D) &= \iiint p(\alpha,\beta,\rho,\sigma \mid D)\, d\alpha\, d\beta\, d\sigma \\ &\propto p(\rho)\{Q(\rho)\}^{(m-1/2)}/\{Q(\rho)S_Y^2 - v[S_{WY1}+\rho S_{WY2}]^2\}^m \end{aligned} \tag{8}$$

en donde $m = M-3$, $v = (p+q)/n$, $Q(\rho) = c_2\rho^2+c_1\rho+c_0$ con

$$c_2 = (p+q)\Sigma_j(W_{2j})^2 - (W_{2.})^2$$

$c_1 = -2W_1 . W_2 .$

$c_0 = (p+q)\Sigma_j (W_{1j})^2 - (W_1 .)^2$

y $W_1 . = \Sigma_j W_{ij}$; i=1, 2. Además,

$S_Y^2 = \Sigma_i \Sigma_j \Sigma_k (Y_{ijk})^2 - (\Sigma_i \Sigma_j \Sigma_k Y_{ijk})^2 / (2m)$

$S_{WY1} = \Sigma_j \Sigma_k Y_{1jk} W_{1j} - (\Sigma_i \Sigma_j \Sigma_k Y_{ijk}) (\Sigma_j W_{1j}) / (p+q)$

$S_{WY2} = \Sigma_j \Sigma_k Y_{2jk} W_{2j} - (\Sigma_i \Sigma_j \Sigma_k Y_{ijk}) (\Sigma_j W_{2j}) / (p+q)$.

La densidad $p(\rho | D)$ está definida para todo $\rho > 0$ y ha sido establecido (Mendoza 1988) que es propia con probabilidad uno y tiene a lo más dos modas, siempre que la inicial $p(\rho)$ sea propia o acotada. La constante de proporcionalidad C, en cualquier caso, debe calcularse numéricamente para obtener la expresión

$$p(\rho|D) = \begin{cases} C\, p(\rho) \{Q(\rho)\}^{(m-1/2)} / \{Q(\rho) S_Y^2 - v\, [S_{WY1} + \rho S_{WY2}]^2\}^m ; & \rho > 0 \\ 0 \qquad \text{otro caso.} \end{cases}$$

Puesto que $\phi = \rho^{1/\lambda}$, se tiene que $\rho = \phi^\lambda$ y el jacobiano de la transformación correspondiente está dado por $\lambda\phi^{(\lambda-1)}$. Entonces, a partir de $p(\rho \mid D)$ se tiene que

$$p(\phi \mid D) \propto p(\phi) [Q(\phi^\lambda)]^{(m-1/2)} \} / \{Q(\phi^\lambda) S_Y^2 - v\, [S_{WY1} + \phi^\lambda S_{WY2}]^2\}^m$$

para $\phi > 0$, en donde $p(\phi)$ representa la densidad inicial para ϕ compatible con $p(\rho)$, esto es, $p(\phi) = p_\rho(\phi^\lambda)\ [\lambda\phi^{(\lambda-1)}]$ con $p_\rho(\phi^\lambda)$ la densidad inicial de ρ evaluada en $\rho = \phi^\lambda$. Como consecuencia de las propiedades de la final $p(\rho \mid D)$ se sigue que $p(\phi \mid D)$ es propia con probabilidad uno siempre que $p(\phi)$ sea propia o acotada. Otras propiedades, más específicas, de esta densidad final dependen de la elección particular de $p(\phi)$ o equivalentemente, de $p(\rho)$. Si se utiliza el procedimiento propuesto por Bernardo (1979) para obtener una densidad inicial conjunta mínimo-informativa cuando ρ es el parámetro de interés, se obtiene

$$\pi(\alpha, \beta, \sigma \mid \rho) \propto \sigma^{-3} \{Q(\rho)\}^{-1/2}$$

que claramente pertenece a la familia de densidades propuestas en este trabajo. De hecho, basta con tomar $r = 3$ y $p(\rho) = \{Q(\rho)\}^{-1/2}$ para establecer este resultado y calcular entonces,

$$\pi(\phi|D) \propto \phi^{(\lambda-1)} [Q(\phi^{\lambda})]^{(m-1)} / \{Q(\phi^{\lambda}) S_Y^2 - v [S_{WY1} + \phi^{\lambda} S_{WY2}]^2\}^m$$

para $\phi > 0$. En la siguiente sección se utiliza esta densidad particular para ilustrar algunos resultados que se pueden obtener con el procedimiento propuesto.

3. EJEMPLO NUMERICO

En esta sección se presenta un ejemplo simulado que exhibe el comportamiento de la densidad $\pi(\phi|D)$ en un problema particular. Para la simulación se utilizó un conjunto de valores parametrales (α=1,β=5,ρ=0.75,σ=1,λ=0.5) que describen una relación dosis-respuesta convexa como se muestra en la Figura 1. El diseño experimental consta de cuatro dosis para cada estímulo (p=q=3) y tres repeticiones (n=3).

Con este arreglo, se generaron dos muestras independientes (D_1 y D_2) que se presentan en la Tabla 1 y se calcularon las correspondientes distribuciones finales $\pi(\phi \mid D_1)$ y $\pi(\phi \mid D_2)$ que se exhiben en la Figura 2. Si se observa que el valor del parámetro de interés está dado por $\rho^2 = (0.75)^2 = 0.5625$, se puede notar, de la misma figura, que las densidades para ambas muestras son unimodales y se concentran en una vecindad del verdadero valor de ϕ. Por otra parte, sin embargo, la variación entre muestras parece considerable en este ejemplo.

4. DISCUSION

El procedimiento descrito en este trabajo puede ser utilizado para el análisis de una amplia clase de modelos para ensayos biológicos del tipo cociente de pendientes. Su aplicación se limita, sin embargo, por la hipótesis de que el valor de λ es conocido. La dificultad técnica del análisis se incrementa considerablemente en el caso en que λ es desconocido y tiene que ser considerado como otro parámetro marginal. Una aproximación, accesible utilizando únicamente los resultados de este trabajo, está basada en la idea de un análisis de sensibilidad de la densidad final $p(\phi|D)$ respecto al posible valor de λ.

En la práctica, suele ocurrir que la información inicial del científico sobre λ establece un intervalo relativamente pequeño en el que toma valores el parámetro. Bajo tales circunstancias, puede ser suficiente con producir una serie

de análisis condicionales para un conjunto finito de valores de λ en el intervalo, obteniendo así una descripción del comportamiento de $p(\phi \mid D)$. Elaborando aún más esta idea, se puede considerar otra aproximación. Si el parámetro λ se discretiza ($\lambda \in \{\lambda_1, \lambda_2, \ldots\}$) y se asigna una densidad inicial conjunta de la forma

$$p(\alpha, \beta, \rho, \sigma, \lambda) = p(\alpha, \beta, \rho, \sigma)\, p(\lambda)$$

incorporando así la hipótesis de independencia a priori entre λ y el resto de los parámetros, la densidad final $p(\lambda \mid D)$ se puede obtener con complicaciones técnicas menores y $p(\phi \mid D)$ se expresa como la mezcla (con pesos dados por $p(\lambda_i \mid D)$; i=1,2,...) de las densidades condicionales que se pueden obtener aplicando el procedimiento descrito en este trabajo.

AGRADECIMIENTOS

Este trabajo ha sido desarrollado con el apoyo del Sistema Nacional de Investigadores de México.

REFERENCIAS

Bernardo, J.M. (1979). Reference posterior distributions for bayesian inference (with discussion). J. R. Statist. Soc., B, 41, 113-147.

Box, G.E.P. and Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Reading, Mass. : Addison-Wesley.

DeGroot, M.H. (1970). Optimal Statistical Decisions. New York : McGraw-Hill.

Mendoza, M. (1987). Análisis bayesiano del cociente de pendientes. Ponencia presentada en el II Foro de Estadística, México, D.F.

Mendoza, M. (1988). Inferences about the ratio of linear combinations of the coefficients in a multiple regression model.In Bayesian Statistics 3, Bernardo, J.M., DeGroot, M.H., Lindley, D.V.and Smith, A.F.M. (eds.) Oxford: Oxford University Press.

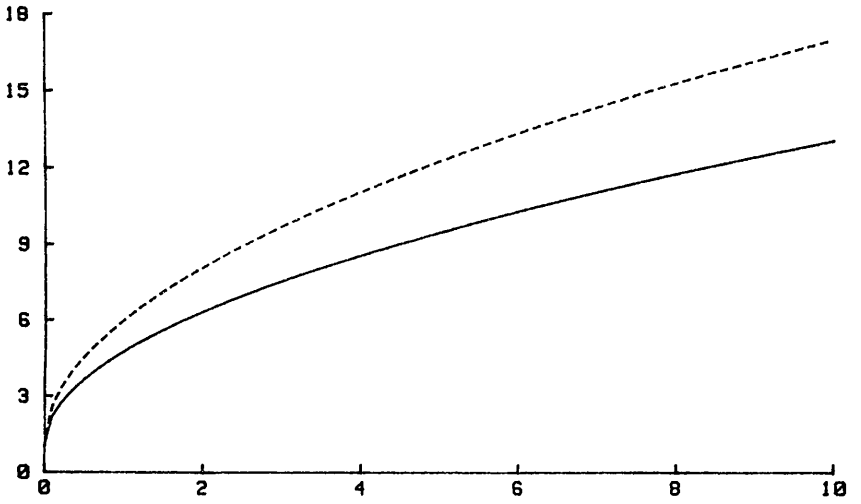


Figura 1. Respuestas esperadas para el ejemplo
(--- : estímulo 1, —— : estímulo 2).

Tabla 1
Datos Simulados (n=3, p=q=4)

Estímulo	1				2			
Dosis	1	4	7	10	2	4	6	8
	6.588	8.593	13.279	16.395	5.122	9.253	11.606	12.765
D_1	6.889	11.473	14.465	15.974	7.273	8.559	11.634	11.050
	5.664	10.692	15.815	17.888	3.614	9.596	10.469	12.059
	4.959	10.980	14.720	17.954	5.678	9.062	9.327	11.679
D_2	6.070	11.176	13.416	17.642	5.170	7.958	13.197	11.104
	5.084	10.215	15.459	17.686	5.545	9.178	8.692	11.553

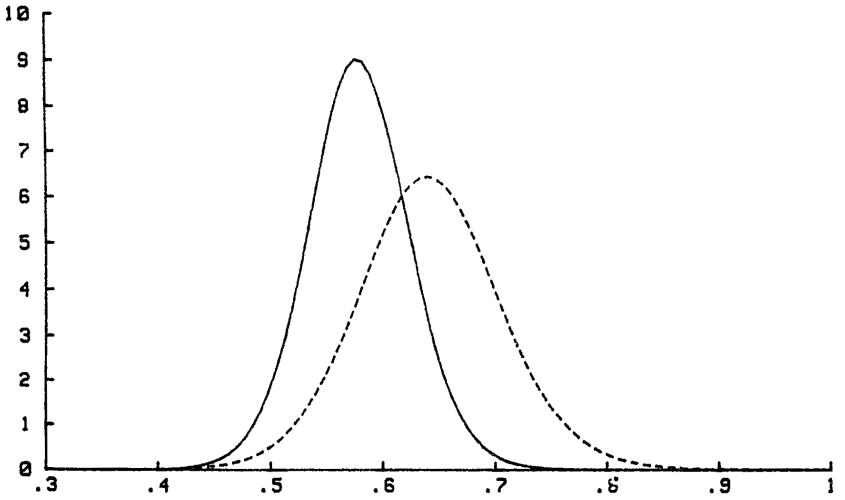


Figura 2. Densidades finales para ϕ (--- : $\pi(\phi/D_1)$, ——: $\pi(\phi/D_2)$).

EFECTO MATEO EN LA PROBLEMATICA CENTRO-PERIFERIA DE LA CIENCIA EN MEXICO

Jaime Jiménez, Miguel Angel Campos, Eloisa Díaz Francés

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM

Apartado Postal 20-726
Administración No. 20
Del. Alvaro Obregón
01000 México, D.F.

RESUMEN

El efecto centro-periferia observado entre los sistemas de ciencia y tecnología de países desarrollados comparados con países menos desarrollados, ha sido un tema de estudio vigente entre los especialistas en los últimos años. Hemos observado y cuantificado el mismo efecto en México. Se define el centro como la zona metropolitana de la ciudad de México; la periferia corresponde al resto del país. En otros trabajos se ha demostrado que en México existe el efecto centro-periferia en la asignación de recursos, la productividad y otras dimensiones de la unidad de investigación. En este estudio se reportan las características del efecto respecto a los cinco grandes campos científicos en que se aglutinan todas las disciplinas (ciencias agropecuarias, ciencias médicas, ciencias de la ingeniería, ciencias exactas y naturales, y ciencias sociales y humanidades). Se concluye que las ciencias médicas y las ciencias exactas y naturales del centro son las que participan más intensamente en la "gran ciencia internacional". Por otro lado, las ciencias agropecuarias del centro y de la periferia, y las ciencias médicas de la periferia manifestan la mayor carencia de recursos para el trabajo de los próximos cuatro a seis años.

INTRODUCCION

En 1985 a solicitud de CONACYT se llevó a cabo una encuesta nacional de 221 unidades de investigación pertenecientes al sistema de ciencia y tecnologia, dentro del marco del proyecto ICSOPRU [1] ("International Comparative Study on the Organization and Performance of Research Units") que es coordinado por la UNESCO a nivel internacional. El objeto de este estudio consiste en comprender mejor el funcionamiento de las unidades de investigación para tomar medidas autocorrectivas que mejoren la organización e incrementen la efectividad del sistema nacional de investigación y desarrollo. En base al Inventario de Instituciones y Recursos Dedicados a las Actividades Cientificas y Tecnológicas del Subsistema de Investigación, que llevó a cabo el CONACYT en 1984, se identificaron en el pais 247 instituciones cuya actividad prioritaria es la investigación cientifica o tecnológica. Como existen diferencias significativas entre las instituciones de la zona metropolitana de la ciudad de México y aquellas localizadas fuera de ella, el universo se estratificó en instituciones pertenecientes al "centro" (zona metropolitana de la ciudad de México), e instituciones pertenecientes a la "periferia" (localizadas en el resto del pais). Adicionalmente, se estratificó al universo, en base a otro criterio, en cuatro grupos de acuerdo con el tipo de institución: académicas públicas, académicas privadas, gobierno federal y otras. Es decir, el universo de instituciones quedó finalmente dividido en ocho estratos según los dos criterios previamente mencionados. Solamente uno de los estratos comprendia un número mayor de instituciones en comparación a los siete restantes: académicas públicas de periferia. Este estrato representa el 47 % del universo de instituciones y abarca en su mayoria a las instituciones de investigación de universidades estatales asi como institutos tecnológicos regionales.

En la primera etapa del muestreo, se censaron todos los estratos a excepción del estrato de académicas públicas de la periferia en donde se realizó un muestreo aleatorio sistemático con fracción muestral de 0.62. Asi, se obtuvo una muestra de 178 instituciones de investigación.

Cada institución se compone a su vez de grupos más pequeños de investigación (departamentos, unidades, "grupos",etc.). En este estudio, se define la unidad de investigación como un grupo de al menos tres personas constituido por el jefe de la unidad, los científicos e ingenieros, y los técnicos, que laboran juntos en cuando menos un proyecto de investigación. El grupo debe tener una expectativa de vida de cuando menos un año. A los miembros de una unidad se les llama "miembros centrales".[2]

En cuanto al universo de unidades de investigación, se observó que había aproximadamente la misma proporción de unidades en el centro que en la periferia. Como la UNESCO requería un mínimo de 200 unidades en la muestra para fines de comparación a nivel internacional, se decidió seleccionar 115 unidades del centro y 115 de la periferia, dando así un margen para poder satisfacer el número mínimo requerido.

En la segunda etapa del muestreo, se procedió a seleccionar unidades pertenecientes a las 178 instituciones seleccionadas en la primera etapa, de la siguiente manera: de cada institución (compuesta por x unidades) se seleccionó aleatoriamente una unidad con probabilidad 1/x.

Con este procedimiento se obtuvieron 178 unidades de investigación (92 del centro y 86 de la periferia). Para completar a las 115 unidades tanto del centro como de la periferia, se volvieron a seleccionar aleatoriamente 52 de las 178 instituciones en la muestra, y de éstas con el mismo procedimiento se seleccionó otra unidad. La muestra final de unidades válidas que respondieron la encuesta consistió de 114 unidades centrales y 107 periféricas.

En trabajos anteriores[3] hemos demostrado el fuerte desbalance que existe en términos de productividad y recursos materiales y humanos entre el centro y la periferia verificándose así el efecto Mateo. El efecto Mateo establece que cuando existe un desbalance entre dos grupos, las diferencias tienden a agrandarse, ocurriendo que el grupo privilegiado lo es cada vez más y el no privilegiado tiende a permanecer en situación desfavorable, perpetuándose así el desbalance.[4] En este trabajo se analiza dicho desbalance a través de las respuestas dadas por los jefes de unidad en términos de los cinco campos científicos generales en que se agrupan todas las disciplinas para su estudio[5], a saber:

* tecnologias y ciencias agropecuarias (TCA),
* tecnologias y ciencias médicas (TCM),
* tecnologias y ciencias de la ingenieria (TCI),
* ciencias exactas y naturales (CEN),
* ciencias sociales y humanidades (CSH).

Debido a que el objetivo del diseño original de muestreo fue contrastar las diferencias de las unidades del centro versus las de la periferia y no analizar un desbalance similar por campo cientifico, cabe aclarar que los resultados que presentaremos a continuación son de carácter exploratorio, ya que al utilizar una muestra disponible y re-estratificar la información obtenida, no podemos conocer la precisión con la que los resultados muestrales describan a los poblacionales en cuanto al aspecto de campo cientifico.

El análisis revela el bajo número de unidades dedicadas a la investigación agropecuaria y médica (6% y 10% respectivamente) en el pais. Asimismo se advierte la carencia generalizada de medios humanos y materiales para el trabajo futuro. Especificamente, las ciencias agropecuarias y las ciencias médicas padecen de una aguda falta de recursos en las unidades periféricas. Finalmente se advierte la orientación internacional de las investigaciones en medicina y en ciencias exactas y naturales sobre todo en las unidades centrales.

ANALISIS

I. DISTRIBUCION DE UNIDADES POR CAMPO CIENTIFICO

Es notorio el desbalance que existe en la proporción de unidades de investigación agropecuaria y de medicina en comparación con los otros campos cientificos. A nivel nacional la investigación agropecuaria representa sólo un 6% del total; asimismo, las unidades médicas alcanzan apenas un 10% del total. Estos valores no son compatibles con las necesidades nacionales de investigación en ambos campos cientificos. A pesar de la fuerte urbanización que se ha observado en el pais, la importancia de las actividades agropecuarias no se puede soslayar, ni postergar la actividad cientifica en un campo estrechamente vinculado con

el problema de alimentación que afecta a los mexicanos. Por otro lado, el pais no ha resuelto problemas de salud a nivel primario, muchos de los cuales, a su vez están estrechamente relacionados con los problemas de nutrición. Es por tanto muy importante fortalecer la investigación en campos donde su aplicación a grandes problemas nacionales es inmediata.

La mayor distribución porcentual se alcanza en el campo de ciencias sociales y humanidades, seguida por el de ciencias exactas y naturales. Ambos campos representan la más antigua tradición cientifica en México. Las ingenierias aparecen en tercer término gracias al impulso que se les ha dado a partir de los años 70.

Tanto en el centro como en la periferia, la distribución de campos cientificos es congruente con la distribución global, salvo en el caso de las ciencias médicas que en la periferia porcentualmente corresponden a la mitad del número de unidades del centro (6% vs 13% respectivamente) y las ingenierias que en la periferia representan el 21% en contraste con el centro, donde representan sólo el 14%.

II. ANTIGUEDAD DE LAS UNIDADES DE INVESTIGACION

La Tabla I muestra que en la periferia las unidades de investigación son de muy reciente creación en comparación con la antigüedad de las unidades centrales. Por otra parte, la tasa de crecimiento de las unidades de investigación en general es mayor en provincia que en la zona metropolitana de la ciudad de México (1.1 y 0.6,respectivamente, en el periodo de 1980-1985 versus 1971-80).

El desarrollo de la ciencia en ciudades de provincia se explica como el resultado de la politica del gobierno federal de impulsar la descentralización de las actividades cientificas y tecnológicas a partir de la década de los 70. Destacan en particular las unidades de provincia dedicadas a las ciencias sociales y humanidades cuya tasa de crecimiento en 1980-85 vs. 1971-80 fue de 1.9, la más alta de todos los campos. El campo cientifico que presentó el crecimiento menor fue el de ciencias médicas del centro con una tasa de 0.3.

III. PRODUCTIVIDAD CIENTIFICA

De la amplia gama de productos del trabajo de investigación reportados por los jefes de unidad, seleccionamos la producción de libros y el número de articulos publicados en el pais y en el extranjero para su anàlisis. Las preguntas se refieren al número de libros, articulos publicados en revistas nacionales y articulos publicados en revistas extranjeras por miembros de la unidad de investigación en los últimos tres años.

LIBROS. Las unidades de instituciones de provincia editaron un número ligeramente mayor de libros por unidad que las unidades del centro en los últimos tres años contabilizados al momento de la encuesta (primavera-verano de 1985): 2.65 en provincia vs. 2.57 en la zona metropolitana de la ciudad de México. Destaca la alta producción de libros de las unidades dedicadas a la ingenieria en la periferia (2.64 libros por unidad), que es notablemente mayor que la correspondiente a las unidades de ingenieria en el centro (0.69 libros por unidad). Este promedio es el más bajo entre las unidades centrales.

Por otro lado, las unidades con mayor producción de libros tanto en el centro como en la periferia corresponden a las ciencias sociales y humanidades con 4.77 y 3.91 libros respectivamente, para el periodo considerado. La menor producción de libros pertenece a las unidades de la periferia dedicadas a la investigación médica (0.67 libros por unidad).

ARTICULOS PUBLICADOS EN MEXICO. En general, las unidades centrales publican más articulos en el pais que las periféricas. Sobresalen por su alta productividad las unidades centrales dedicadas a las ciencias agropecuarias (13.71 por unidad), a las ciencias sociales y humanidades (9.59 por unidad), y a la investigación médica (8.60 por unidad). Es interesante hacer notar que en la periferia, el promedio de publicaciones por unidad es aproximadamente el mismo, independientemente del campo cientifico de que se trate (en la vecindad de 5 articulos por unidad).

ARTICULOS PUBLICADOS EN EL EXTRANJERO. Esta categoria de la productividad cientifica muestra el mismo comportamiento global que las publicaciones nacionales: las unidades del centro publican más en el extranjero que las de la periferia. Mediante esta medida de la productividad se puede apreciar qué campos cientificos mantienen mayores vinculos con la ciencia de otros paises. Resaltan en el centro las unidades de medicina y en ciencias exactas y naturales con un promedio de siete publicaciones por unidad, que contrasta notablemente con el promedio de dos publicaciones por unidad para el resto de los campos cientificos.

En la periferia, nuevamente sobresalen las unidades dedicadas a la ingenieria con un promedio de 3.09 articulos por unidad que es mucho mayor que el promedio en los otros campos cientificos. Sin embargo, se sitúa bastante abajo del promedio de siete publicaciones encontrado para las unidades de medicina y de ciencias exactas y naturales localizadas en el centro del pais.

Las unidades periféricas con más baja producción de articulos en el extranjero pertenecen a las ciencias agropecuarias y a las ciencias médicas, con un promedio de 0.67 articulos por unidad por campo cientifico.

IV. ADECUACION DE RECURSOS PARA EL TRABAJO FUTURO

Los jefes de unidad respondieron varias preguntas referentes a la adecuación de los recursos para el trabajo cientifico de los siguientes cuatro a seis años. El entrevistado podia optar por una de las siguientes respuestas:

* los recursos están lejos de ser adecuados
* los recursos necesitan ser reforzados
* los recursos son plenamente adecuados

Para el análisis se acumuló la frecuencia de las primeras dos respuestas caracterizada por la proposición: "los recursos necesitan por lo menos ser reforzados", de manera que los porcentajes que se presentan a continuación se refieren a esta categoria combinada.

En forma general, observamos que tanto en el centro como en la periferia, cuando menos el 80% de los jefes de unidad en cada campo científico manifiestan que sus recursos necesitan por lo menos ser reforzados.

En el aspecto de recursos humanos (científicos e ingenieros y técnicos) las unidades periféricas se encuentran en un situacion más desfavorable que las centrales. Asimismo, en cuanto a equipo especializado y acceso a la literatura científica nacional e internacional, las unidades periféricas manifiestan mayor inadecuación de recursos que las centrales.

Paradójicamente, las unidades centrales de ingeniería, ciencias exactas y naturales, y ciencias sociales y humanidades reportaron mayor carencia de equipo no especializado que las unidades en la periferia. Una posible explicación consiste en suponer que las unidades del centro cuentan con equipo especializado pero que el apoyo para dicho equipo es deficiente.

En cuanto a medios para el tratamiento de datos por computadora, se observa una situación igualmente desfavorable tanto para el centro como para la periferia en todos los campos científicos.

Resalta el hecho de que las unidades en medicina de la periferia presentan en los tres tipos de equipo (especializado, no especializado y de computación) una situación mucho más desfavorable que las unidades en medicina del centro.

En cuanto al acceso a la literatura científica nacional e internacional, observamos una situación más desfavorable para la periferia que para el centro en todos los campos a excepción de las ciencias exactas y naturales en donde es igualmente desfavorable en el centro que en la periferia.

Sintetizando, los campos científicos en donde reiteradamente un porcentaje mayor de jefes de unidad manifiestan que sus recursos son inadecuados son las ciencias agropecuarias del centro y de la periferia y las

médicas de la periferia. Estas últimas son las unidades que parecen estar operando en las peores condiciones, en comparación con el resto, de acuerdo a lo manifestado por los jefes de unidad.

CONCLUSIONES

A juzgar por el número de unidades dedicadas a cada campo científico, es patente el desbalance que existe en contra de las ciencias agropecuarias y las médicas a nivel global. Este desbalance no es congruente con las carencias que el país confronta en materia de alimentación y de salud pública.

En otro aspecto, se observa que la ciencia en la periferia es más "joven" que en el centro. El crecimiento en la década de los 70 fue uniforme tanto en el centro como en la periferia. Sin embargo, a partir de 1981, el crecimiento en número de unidades en instituciones de provincia es aproximadamente el doble que en instituciones de la zona metropolitana de la ciudad de México, evidenciando el intento del gobierno por desconcentrar las actividades científicas del país.

La edición de libros es ligeramente mayor en unidades de provincia que en unidades de la ciudad de México. La publicación de artículos tanto en el país como en el extranjero es mayor en el centro que en la periferia. Destaca la publicación de artículos en el extranjero en ciencias médicas y ciencias exactas y naturales que evidencia su vínculo con la "gran ciencia" internacional. En la periferia sólo las ciencias de la ingenierías reportan una publicación apreciable de artículos en revistas extranjeras.

Es notorio el contraste entre las unidades en ciencias médicas del centro y la periferia. La producción científica de las unidades centrales es alta y está orientada hacia el extranjero, lo que hace suponer que seleccionan sus temas de investigación entre los vigentes de la "gran ciencia" internacional. En cambio las unidades periféricas tienen la producción más baja de libros y artículos en revistas extranjeras, y su producción de artículos en revistas nacionales está en el promedio de las periféricas.

Asimismo, la producción científica escrita de las unidades agropecuarias en la periferia es muy baja comparada con la del centro.

En cuanto a la disponibilidad de recursos para el trabajo futuro, tanto el centro como la periferia manifiestan una gran necesidad de reforzarlos. Los campos que sobresalen en carencia de recursos son las ciencias agropecuarias del centro y la periferia, y las ciencias médicas de la periferia.

Aunque ha habido un aumento en el número de unidades de investigación en el interior del país, los datos permiten afirmar que no han tenido el debido respaldo en infraestructura que les permita ubicarse como unidades de primera categoría. En otras palabras, se han fundado unidades "secundarias" de acuerdo a la clasificación usada en un trabajo anterior[6] y se verifica que la condición de "secundaria" en virtud del efecto Mateo, se hace permanente, como se afirma en dicho estudio.

Nuestra recomendación se sintetiza de la manera siguiente: se debe continuar el crecimiento de la ciencia en el interior de la república, y no en la ciudad de México. Es necesario aumentar el número y mejorar la calidad de unidades de investigación en ciencias agropecuarias y ciencias médicas en la periferia. Es urgente aumentar el apoyo en infraestructura humana y de equipo en todo el país. Finalmente, al fundar nuevos centros de investigación, se les debe dotar con los medios que permitan ubicarlos como unidades de primer nivel desde su inicio. La omisión de este esfuerzo condena a dichos centros a permanecer en un segundo nivel, siendo extremadamente difícil lograr que con el tiempo cambien de categoría.

NOTAS Y REFERENCIAS:

1. ICSOPRU es un proyecto que en México fue patrocinado por el gobierno federal, a través del Consejo Nacional de Ciencia y Tecnologia.

2. Se administraron cuatro cuestionarios, tres de ellos formulados por la UNESCO y uno adicional formulado por el equipo de trabajo en México. Con el objeto de llevar a cabo estudios comparativos de carácter internacional, los cuestionarios de la UNESCO tienen el mismo contenido para todos los paises participantes. El cuestionario adicional se refiere a tópicos que conciernen solamente a México y fueron administrados a todo el personal que contestó la encuesta. Los cuestionarios de la UNESCO están dirigidos a tres niveles jerárquicos en la institución:

* a los directores de la institución a que pertenece la unidad de investigación muestreada,
* a los jefes de unidad,
* a los miembros centrales de la unidad.

3. Jimenez J., M.A.Campos, J.Dominguez , L. Romano; "Center-Periphery Analysis of Research and Development Resource Allocation: Preliminary Results of ICSOPRU in Mexico"; Comunicaciones Técnicas, IIMAS, UNAM, Serie naranja, investigaciones; No. 436 México, 1986.

4. Referencia: Merton, R.;"The Mathew Effect in Science"; Science, V.159 (pp. 56-63); enero, 1968.

5. Clasificación utilizada en las publicaciones del Consejo Nacional de Ciencia y Tecnologia.

6. Jimenez J., M.A. Navarro, M.W. Rees;"Scientific Research Areas in Mexico: Growth Patterns in the Late Seventies"; Scientometrics, V.9, Nos 5-6 (pp.209-223);1986.

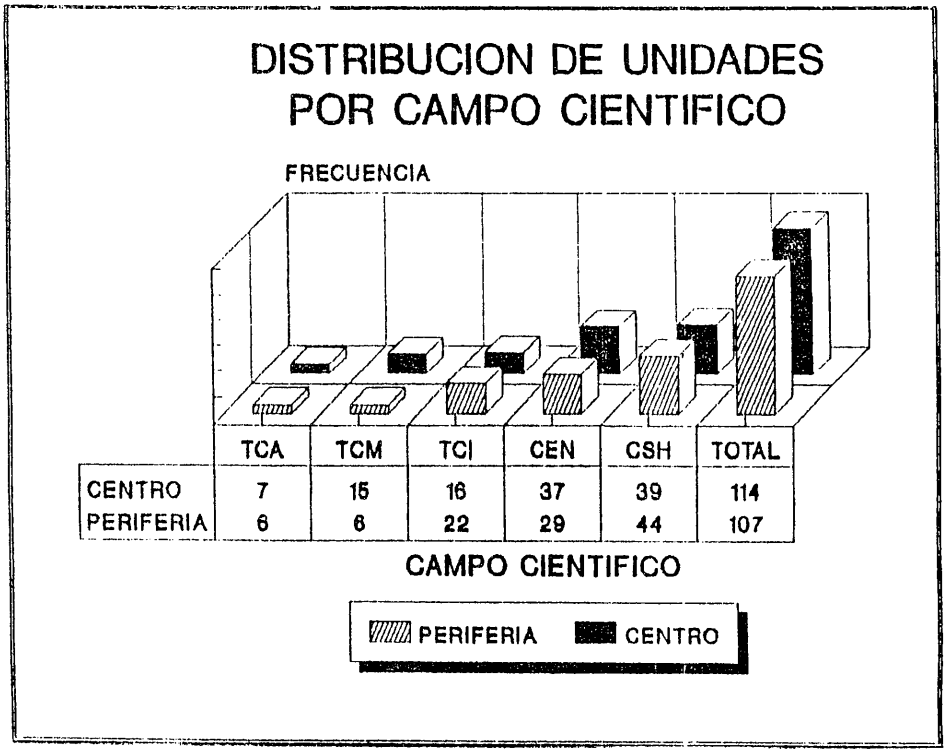


TABLA II. ANTIGUEDAD DE LAS UNIDADES DE INVESTIGACION

CENTRO	TCA	TCM	TCI	CEN	CSH	GLOBAL
Antes de 1950	0	2	0	2	1	5
1951-1970	0	3	4	11	9	27
1971-1980	4	8	8	15	17	52
1981-1985	2	2	4	9	12	29
TASA CRECIMIENTO (1981-85/1971-80)	(.5)	(.3)	(.5)	(.6)	(.7)	(.6)
PERIFERIA						
Antes de 1950	0	0	0	0	0	0
1951-1970	0	0	3	1	1	5
1971-1980	2	2	13	17	15	49
1981-1985	4	3	6	11	28	52
TASA CRECIMIENTO (1981-85/1971-80)	(2)	(1.5)	(.5)	(.6)	(1.9)	(1.1)

COMPORTAMIENTO AGRONOMICO DE OCHO GENOTIPOS DE JITOMATE PARA MERCADO FRESCO EN EL ESTADO DE TABASCO

Vargas Ch. D.
Aranda L. S.
CIFAP-TABASCO, Apdo. Postal 17.
86400, Huimanguillo, Tabasco.

RESUMEN

Durante la época de nortes se evaluaron ocho genotipos de jitomate para consumo fresco en dos localidades, tres fechas de siembra, utilizando un diseño en parcelas divididas. El mercado fresco contempla tres categorías: exportación, nacional y regional, se evaluó el rendimiento agronómico para cada calidad en un máximo de ocho cortes durante el ciclo de desarrollo. Para cada calidad se realizó el análisis de varianza y prueba de Tukey para el rendimiento acumulado al corte ocho, así como el análisis de perfiles para los cortes uno al ocho. De acuerdo a los resultados, la localidad con textura media obtuvo los mejores rendimientos, la siembra durante el mes de enero superó a la de febrero. Las mejores variedades para la región son: Floradade para exportación, Roma VF para mercado nacional y Homestead 24 y 500 para mercado regional.

INTRODUCCION

El jitomate en el estado de Tabasco, durante 1986 se sembró en una superficie de 141 ha, con rendimientos de 15 ton/ha. El producto se ha comercializado al mercado local de autoconsumo, la gran mayoría lo siembra en predios de 0.5 a 1 ha. El sureste mexicano presenta condiciones agroecológicas favorables que indican la factibilidad de producir diversas especies hortícolas. Sin embargo, en esta región se carece de tradición hortícola, ya que es una actividad relativamente nueva.

En Tabasco, existen dos regiones muy promisorias para la producción de hortalizas y son: la región de los Ríos y la Chontalpa. Desde el punto de vista social, incrementar la horticultura en el estado, permite generar demanda de mano de obra en el campo; en términos de economía son fuente de ingreso de divisas al país por concepto de exportación.

Finalmente el aprovechamiento de la humedad residual, producto de la precipitación pluvial registrada durante

verano-otoño, es posible que prospere el jitomate durante la época invernal. Ello representa la oportunidad de comercializar el producto hacia el mercado de exportación de la costa este de los Estados Unidos de Norteamérica ya que no produce hortalizas en invierno.

El objetivo del presente trabajo es determinar la localidad, la época de siembra y las variedades que se adapten mejor y obtengan los mejores rendimientos en un lapso de tres años (1986-1989).

REVISION DE LITERATURA

La literatura sobre el cultivo del jitomate en regiones tropicales no es muy abundante. De hecho la investigación que se realiza en Tabasco es muy reciente.

Lara (1985) realizó un experimento sobre cinco fechas de siembra en jitomate de la variedad Floradade. Las fechas de siembra fueron: 8 de noviembre, 28 de noviembre, 18 de diciembre, 7 de enero y 28 de enero. Evaluó fruto sano, fruto dañado (rajaduras, pudriciones de sol, porciento de amarre, número de frutos, número de flores, días de floración e incidencia de plagas y enfermedades). Concluye preliminarmente que las fechas evaluadas, la mejor fue la del 8 de noviembre.

Quevedo (1985) evaluó 11 variedades de jitomate, las variedades fueron ACE 55 VF, Floradade, Floradel, Homestead 500, Pacesetter 490, Petomech II, Redstone, Riñón, Roma VF, Walter y Winner. Concluye que las variedades Redstone y Roma VF, fueron más resistentes a plagas y enfermedades. El principal problema fue las pudriciones y las variedades más susceptibles fueron: Petomech II y Riñón. Se evaluó el rendimiento de fruto dañado.

Casillas (1985) realzó un trabajo sobre densidades de plantación de jitomate variedad Walter en la Chontalpa, Tabasco. Encontró que la mejor distancia para la variedad de ciclo determinado fue de 0.30 m entre plantas y 1.5 m entre surcos.

Vásquez (1985) realizó un trabajo, en la Chontalpa, Tabasco sobre fertidización con nitrógeno, fósforo y potasio y su influencia en la calidad del fruto. Concluye que si hay un efecto por la fertilización e influye en la calidad del producto en cuanto al color y tamaño.

Respecto al análisis estadístico se realizó mediante el análisis de varianza univariado y el analisis de perfiles, tratado por Morrison (1978) y Jhonson & Wichern (1982).

MATERIALES Y METODOS

El presente experimento se llevó a cabo en dos localidades, "Los Pinos" y "CEFAPHUI". Ambas tienen una precipitación media anual de 2300 mm y una temperatura promedio de 26 C. La primera localidad se caracteriza por ser un suelo de textura pesada, el segundo es media. Las fechas de siembra son tres: 5 de enero, 25 de enero y 20 de febrero. Las variedades a probar son: Floradade, ACE-55 VF, Redstone, Roma VF, Pacesetter, Winner, Homestead 24 y Homestead 500.

El diseño experimental utilizado en cada localidad fue el de parcelas divididas. La parcela grande es la fecha de siembra y la menor son las variedades. Se utilizaron cuatro repeticiones y el tamaño de la unidad experimental fue de 8 M2 en la primera y 4 M2 en la segunda. El modelo univariado es el siguiente:

$$Y_{ijklm} = M + L_i + B_{j(i)} + F_k + LF_{ik} + BF_{jk(i)} + V_l + LV_{il} + BV_{jl(i)} + FV_{kl} + LFV_{ikl} + E_{m(ijkl)}$$

$$i = 1,2 \quad j = 1,\ldots,4 \quad l = 1,\ldots,8 \quad m = 1,\ldots,4$$

Donde Y_{ijklm} es el rendimiento acumulado hasta el corte ocho de la i-ésima localidad, j-ésimo bloque, k-ésima fecha de siembra, l-ésima variedad, m-ésima repetición.

El análisis estadístico consistió en su primera fase en un análisis de varianza y prueba de Tukey para el rendimiento acumulado hasta el corte ocho. En su segunda fase consistió en analizar el análisis de perfiles, realizando la prueba de paralelismo entre cortes, para la localidad CEFAPHUI.

La prueba de hipótesis de paralelismo plantea probar:

$$H_{01} : M^t \tau = 0 ,$$

donde

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & \ddots & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}_{pxq-1} \qquad \tau = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \mu_{i3} \\ \cdots \\ \mu_{i7} \\ \mu_{ip} \end{bmatrix}_{px1}$$

La hipótesis de paralelismo es análoga al análisis de varianza multivariado para un criterio de clasificación para las p-1 diferencias de las observaciones de las respuestas adyacentes de cada corte; esto se traduce en la forma de los elementos de matriz hipótesis y de la matriz error utilizadas en el criterio de Wilks. Consideremos las matrices B y E las cuales son (pxp). Donde B= $[b_{rs}]$y E= $[e_{rs}]$

$$[b_{rs}] = \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{1}{n} y_{ijr} y_{ijs} - \frac{1}{n} G_r G_s$$

$$[e_{rs}] = \sum_{j=1}^{k} \sum_{i=1}^{n} y_{ijr} y_{ijs} - \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{1}{n} y_{ijr} y_{ijs}$$

donde y_{ijr}= es la i-ésima observación de la respuesta r bajo el tratamiento j-ésimo y

$$G_r = \sum_{j=1}^{k} \sum_{i=1}^{n} y_{ijr},$$

utilizadas en la versión multivariada del diseño completamente al azar. Ahora para nuestro caso se define a B^* y E^* , matrices de (p-1)x(p-1) como las matrices hipótesis y error del análisis de perfiles, que son

$$B^* = [b^*_{rs}] \text{ y } E^* = [e^*_{rs}]$$

Los elementos b^*_{rs} y e^*_{rs} se pueden reescribir como las matrices hipótesis y error del análisis de perfiles, que son

$$B^* = [b^*_{rs}] \quad y \quad E^* = [e^*_{rs}].$$

Los elementos b^{*}_{rs} y e^{*}_{rs} se pueden reescribir como

$$b^{*}_{rs} = b_{rs} - b_{r+1,\,s} - b_{r,\,s+1} + b_{r+1,\,s+1}$$
$$e^{*}_{rs} = e_{rs} - e_{r+1,\,s} - e_{r,\,s+1} + e_{r+1,\,s+1} .$$

Tales ecuaciones manifiestan el efecto de los contrastes establecidos para respuestas consecutivas de cada tratamiento y una forma de asociar a las matrices B y E con B^{*} y E^{*} es mediante la matriz M, ya definida, y tendremos que

$$B^{*} = M^{t} B M$$
$$E^{*} = M^{t} E M .$$

Similar al caso multivariado de un criterio de clasificación, el estadístico de prueba para la hipótesis de paralelísmo es Λ de Wilks, donde

$$\Lambda = \frac{|E^{*}|}{|E^{*}+B^{*}|} = \frac{|M^{t}EM|}{|M^{t}(E+B)M|} ,$$

bajo H_{01}, el estadístico Λ de Wilks se distribuye como $\Lambda(m, p-1, q)$ donde $m = h-1$, $q = t-1$ y el criterio de prueba es, rechace H_{01}, para valores bajos Λ o en su caso, utilizar la aproximación de Bartlet, la que en el caso de un diseño completamente al azar es

$$-\{n-1-\tfrac{1}{2}(p-1+t)\}\ \ln\Lambda \approx X^{2}_{((p-1)(t-1))}$$

y el criterio de decisión es rechazar H_{01}, si el valor transformado de Λ es mayor que $\Lambda((p-1)(t-1);\alpha)$ donde α es el nivel de significancia de la prueba.

RESULTADOS Y DISCUSION

El análisis de varianza para el rendimiento acumulado al corte ocho reporta significancia para los efectos localidad, fecha de siembra y variedad ($P<.001$) para las calidades exportación, nacional y regional (Cuadro I). Los mejores rendimientos se observaron para la localidad CEFAPHUI, siendo de 37.4 ton/ha para la suma de las tres calidades durante los ocho cortes. La fecha de siembra que reportó los ,mayores rendimientos fue el 5 de enero con 38.9 ton/ha para la producción total. Las variedades con mayor producción acumulada para las tres calidades fueron Roma VF y Floradade con 35.8 y 31.2 ton/ha, respectivamente.

La gráfica de medias para el rendimiento acumulado vs. el número de corte por localidades, muestra la tendencia de incremento sostenido en la localidad CEFAPHU (Fig. 1); el crecimiento se da hasta el corte cinco, a partir de éste, para la localidad "Los Pinos" no hay producción. Las calidades Nacional y Regional tienen un comportamiento similar de crecimiento, solo que el rendimiento de la última es mayor que las calidades restantes, la textura del suelo para CEFAPHUI es franco y permite obtener rendiminetos altos.

Para observar el efecto de las fechas de siembra la Fig. 2 muestra las medias de rendimiento acumulado para los ocho cortes. A partir del corte cinco las fechas 5 y 25 de enero, presentan los rendimientos acumulados más altos para la calidad exportación. El comportamiento agronómico para las calidades restantes es similar.

En cuanto al efecto de variedades, se tiene un comportamiento distinto para las calidades exportación, nacional y regional. Para la calidad exportación las variedades Floradade, Roma VF y Pacesetter reportan rendimientos entre 10 y 12 ton/ha (Fig. 3). Sin embargo Folaradade tiene un intervalo amplio de variación, su rendimiento oscila entre las 5 y 22 ton/ha.

Para la calidad nacional el comportamiento es diferente. En general los rendimientos son inferiores, comparado con las calidades restantes. Las variedades más prometedoras son Winner y Floradade con rendimientos de 10 ton/ha. Por último para la calidad regional, destacan la Roma VF, Homestead 500 y Homestead 24 con rendimientos acumulados de 13.8 a 16.2 ton/ha.

La prueba de paralelismo para las variedades reporta una Λ de Wilks significativa ($P<.001$), para calidad exportación $\Lambda=.2001$, para la nacional $\Lambda=.1779$ y la regional $\Lambda=.1523$. Por lo que el rendimiento agronómico acumulado durante los ocho cortes no tiene un comportamiento

paralelo; es decir, la tasa de crecimiento en producción no es la misma entre variedades.

Para las tres calidades las tendencias de producción acumulada se mantienen paralelas hasta el corte cuatro, ahí ocurre un cambio en la producción y algunas variedades dejan de producir de manera sostenida y otras incrementan su producción. Por ejemplo para la calidad exportación la variedad Floradade manifiesta de manera sostenida su producción desde el primer hasta el séptimo corte, en tanto que la Pacesetter deja de manifestar su tendencia a partir del corte cuatro.

CONCLUSIONES

Los suelos francos en el estado de Tabasco son adecuados para el cultivo de jitomate, sus características permiten que exista una adecuada disponibilidad de nutrientes, buen drenaje, lo cual permite un manejo con menos dificultades que los suelos con textura pesada.

En la época de nortes las siembras realizadas durante el mes de enero permiten que el jitomate tenga una una floración y amarre de frutos justo antes de que inicie la época de secas. Las siembras tardías, en febrero, no permiten un buen amarre de frutos.

En cuanto a las variedades se observa que para la calidad exportación la Florarade obtiene rendimientos altos, sumando las tres calidades su producción total es de 31.2 ton/ha pero con una gran variación, para la calidad nacional es la Roma VF con un total de 35.8 y finalmente para la regional es la Homestead 24 y 500 con un total de 27.9 y 27.8 ton/ha, respestivamente.

BIBLIOGRAFIA

1. - Casillas, E.J.C., (1986). Determinación de la densidad óptima de plantación en tomate (*Lycopersicum esculentum Mill*) variedad Walter en la Chontalpa, Tabasco. Tesis de licenciatura. Colegio Superior de Agricultura Tropical.

2. - Jhonson, R.A. y Wichern D.W. (1982). Applied Multivariate Statistical Analysis. Prentice-Hall Inc.

2. - Lara, P.H. (1985). Evaluación del rendimiento del cultivo de tomate (*Lycopersicum esculentum Mill*)

establecido en cinco fechas de siembra en la región de la Chontalpa, Tabasco. Tesis de licenciatura. Colegio Superior de Agricultura Tropical.

3. - Morrison, D.F. (1978). Multivariate Statistical Methods. McGraw Hill, 2nd. Edition.

4. - Quevedo, C.F. (1985). Evaluación del rendimiento de 11 variedades de tomate (*Lycopersicum esculentum Schrad*) en la región de la Chontalpa, Tabasco. Tesis de licenciatura. Colegio Superior de Agricultura Tropical.

5. - Vásquez, A.P. (1986). Dosis óptima económica de fertilización N-P-K y su influencia en la calidad del tomate (*Lycopersicum esculentum Mill*) en suelos de la Chontalpa, Tabasco. Tesis de licenciatura. Colegio Superior de Agricultura Tropical.

CUADRO I. ANALISIS DE VARIANZA PARA EL RENDIMIENTO ACUMULADO AL CORTE OCHO 1_/

FV.	G. L.	Exportacion	Nacional	Regional
L_i	1	1858.1 **	1785.3 **	1832.6 **
$B_{j(i)}$	6	176.5 **	132.7 **	217.9 **
F_K	2	960.8 **	528.1 **	1037.8 **
$L*F_{iK}$	2	775.9 **	397.7 **	1134.8 **
$B*F_{jK(i)}$	12	21.9 **	22.4 *	13.9
V_l	7	133.9 **	41.3 *	224.3 **
$L*V_{il}$	7	3.4	3.7	62.6 **
$B*V_{jl(i)}$	42	10.6	8.5	25.2
$F*V_{kl}$	14	19.4 **	11.6	24.6
$L*F*V_{ikl}$	14	17.0 *	9.4	25.3
$E_{m(ijkl)}$	84	8.4	9.6	17.4

1_/ Cuadrados medios,
* Significativo P<.05
** Significativo P<.001

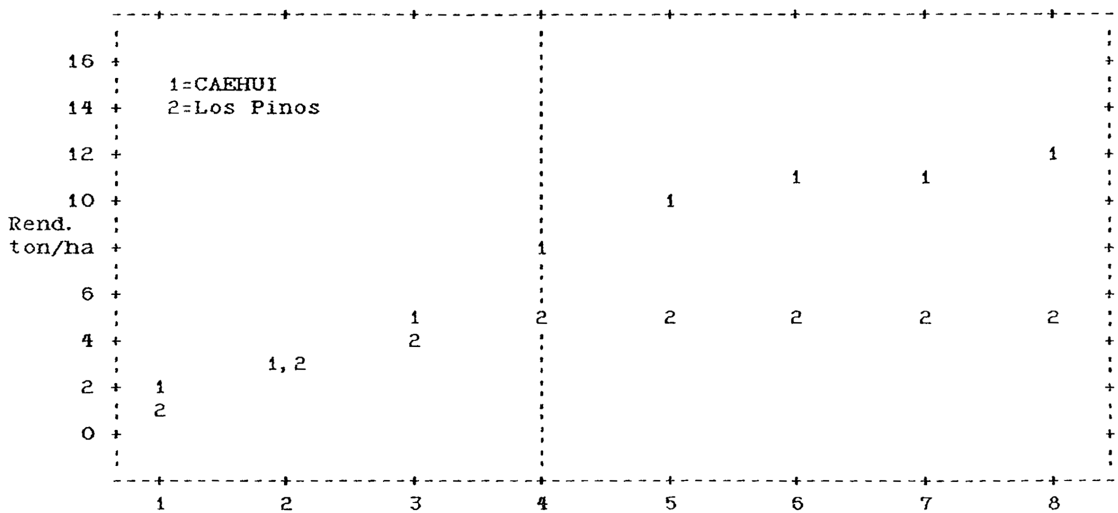


Fig1. Medias de rendimiento acumulado en dos Localidades. CIFAPTAB 1988.
Calidad Exportación

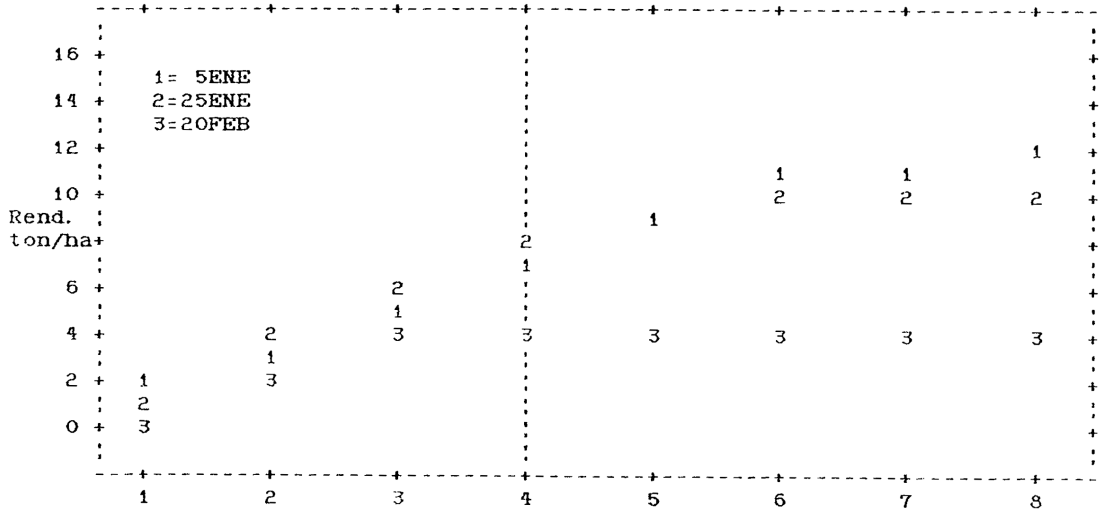
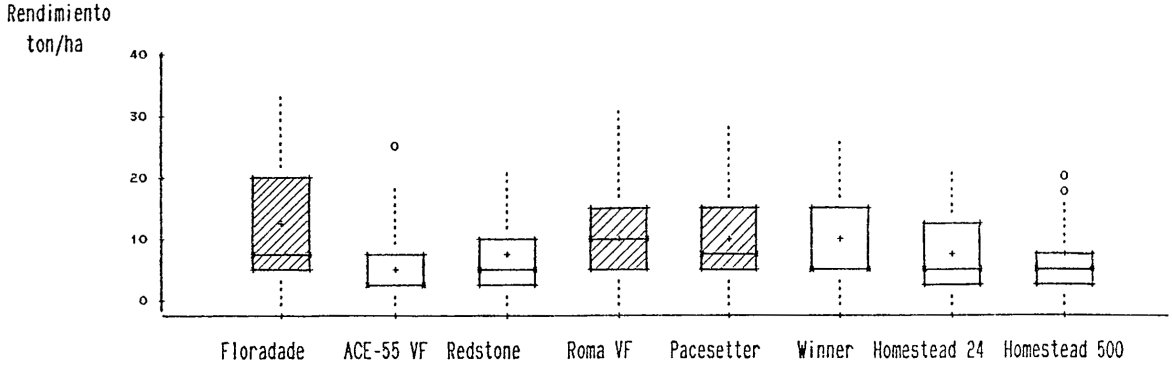


Fig2. Medias de rendimiento acumulado para fechas de siembra. CIFAPTAB 1988
(a) Calidad Exportación

RENDIMIENTO ACUMULADO PARA VARIEDADES DE JITOMATE
CALIDAD EXPORTACION. CIFAPTAB 1988.

Gráficas Esquemáticas



PRUEBA DE COMPARACIONES MULTIPLES DE TUKEY

VARIEDAD	RENDIMIENTO	
Floradade	12.1	a
Roma VF	11.1	a b
Pacesetter	10.3	a b
Winner	8.9	b c
Redstone	7.3	c d
Homestead 24	6.9	c d
Homestead 500	6.3	c d
ACE-55 VF	5.7	d

CARACTERIZACION DE PRODUCTORES DE CACAO EN EL ESTADO DE TABASCO MEDIANTE EL ANALISIS DE CORRESPONDENCIAS

Rodríguez G. F., Vargas Ch. D.
CIFAP INIFAP Apdo postal 17
Huimanguillo, Tab.

R E S U M E N

En Octubre de 1985 en el Estado de Tabasco se les aplicó un cuestionario a los productores de Cacao, con el objeto de conocer los medios de comunicación que prefieren en la adopción de nuevas prácticas agrícolas.

La importancia de la presente investigación estriba en la Técnica de Análisis de un cuestionario usando el Análisis de Correspondencias.

Para analizar la información obtenida, primero se dividió la población de cacaoteros en tres estratos denominados: pequeños, medianos y grandes productores de acuerdo a la superficie cultivada con Cacao.

Los pequeños productores constituyen el 87% de la población. La gráfica del Análisis de correspondencias revela que estos productores no realizan todas las labores cuando están en contacto con el Extensionista y no prefieren la radio como fuente de información agrícola.

Para los medianos productores el Análisis de correspondencias revela que cuando prefieren la radio y no tienen contacto con el Extensionista, asocian rendimientos altos. También hay correlación entre producción alta y lectura de periódicos locales.

El 6% de los Cacaoteros son grandes productores, el Análisis de Correspondencias revela que la totalidad de estos señores tienen rendimientos altos, realizan todas las labores culturales y desean seguir cultivando el Cacao criollo tambien asocian a estas características poca experiencia, la cual compensan asistiendo a exposiciónes cerca de su localidad y leyendo periódicos nacionales. Son personas maduras, nativos de la región, además no prefieren la radio como fuente de información agrícola y en su lugar prefieren leer folletos agrícolas. Debido a que la mayoria revela beneficios cuando prefieren la radio concluimos que el mejor medio de comunicacion para transferir la nueva tecnologia a los cacaoteros es la radio.
A raiz de esto es necesario hacer un estudio mas especifico de la radio, los programas que desean escuchar y los temas que les interesan.

I N T R O D U C C I O N

Actualmente se cultivan aproximadamente 4.7 millones de hectáreas de cacao, principalmente en América, Africa, Malasia y Oceanía entre los 20°de latitud Norte y Sur.

Hasta 1986, América produjo el 85% de la producción mundial, al cual bajó debido principalmente a tres enfermedades pudrición negra de la mazorca, Moniliasis y escoba de bruja. Hoy en día, América sólo produce el 33 % de la producción Mundial de cacao y México ocupa el octavo lugar produciendolo principalmente en el estado de Tabasco. El cual cuenta con una superficie de 25,267 Km2 donde actualmente se cultivan 44 mil has, las cuales en 1984 produjeron alrededor de 30 mil toneladas con un valor de 9600 millones de pesos, representando la principal fuente de ingresos para más de 16 mil familias Tabasqueñas.

Entre las condiciones climatológicas que permiten desarrollar mejor el cacao, se encuentran, una temperatura mínima de 20°C y máxima de 25° y no menos de 1500 mm anuales de lluvia. No tolera más de 3 meses de estación seca, aunque puede auxiliarse con riegos. La altura máxima favorable es de 700 metros sobre el nivel del mar, si bien pueden encontrarse plantaciones a mayores alturas ademas de que tolera una gran variedad de suelos, siempre que tengan buen drenaje. Por lo anterior, solo en algunos municipios de este estado se produce este producto, entre ellos los más importantes son: Cardenas, Comalcalco, Cunduacán, Huimanguillo, Jalpa de Méndez, Paraiso y Teapa.

Actualmente los productores Tabasqueños han formados sociedades de crédito y asociaciones agrícolas que les han permitido comprar infraestructuras avanzadas que les aumentan los beneficios al vender sus productos en la región o inclusive explotarlos a otros países. Sin embargo, en las plantaciones los productores enfrentan grandes problemas tales como: Pudrición negra de la mazorca, el mal del machete, antracnósis, uso de variedades poco rendidoras de mala calidad, el salivazo, el pulgón negro, el tripps, malas hierbas, desconocimiento de fertilización , forma de poda, etc. El INIFAP (Instituto Nacional de Investigaciones Forestales y Agropecuarias) realiza investigaciónes para resolver estos problemas y posteriormente transfiere a los productores las experiencias usando distintos medios de comunicación. Deseando mejorar la eficiencia en la transferencia de tecnología en octubre de 1985 aplicó un cuestionario a los productores con el objeto de conocer los medio de comunicación que prefieren, su situación socioeconomica y sus necesidades.

La importancia del presente trabaja estriba en la aplicación del Análisis de correspondencias como herramienta estadistica de Análisis de un cuestionario.

ANTECEDENTES DE LOS MEDIOS DE DIFUSION

Willson y Gallup (1960), detectaron que el grado de adaptación de la nueva tecnologia por parte de los productores se ve afectada por el nivel de educación, el tamaño de la parcela, nivel socioeconomico y contacto con el extensionista. Además que a mayor edad del productor había menor interes en adaptar las nuevas prácticas.

Bohlen et .al (1961) comparando la tecnología que adoptan y la de los productores dicen que la adoptada tiene ciertas ventajas. Por ejemplo: que es compactible con el cultivo, más compleja y que al principio la aplican a nivel experimental.

Martinez (1963) descubrió que la mayoria de los productores adoptantes del maiz hibrido en Guanajuato fueron los que contaban con mayores fuentes de información.

Por su parte canizales y Myren (1967) encontraron que los campesinos que leian periodicos eran los que poseian más de 15 hectáreas.

Según Real y Boulen (1986) la adopción de la nueva tecnología está en funsión de la edad, además que los primeros en aceptar el cambio; son los productores más jovenes de posición social alta, con mayor educación y dedicación especializada en su actividad.

Roger et al, citado por Zambrano (1975) encontró que la edad no está relacionada con las innovaciones.

Fierro y Quiroz (1980), señalan que los medios impresos son menos efectivos que los audivisuales, además de que permiten controlar la velocidad y ritmo de lectura.

EL ANALISIS DE CORRESPONDENCIAS (A N C O R R)

Se define principalmente como una tecnica descriptiva multidimensional que permite análizar un gran conjunto de variables discreta al mismo tiempo, despliega renglones y columnas de una tabla de contingencias de doble entrada como puntos en correspondencia sobre espacios vectoriales de menor dimensión (Greenacre y Urba, (1984)). Al proyectar los espacios vectoriales de renglones y los de las columnas en un solo espacio, obtiene una grafica conjunta.

ANTECEDENTES DEL A N C O R R .

Los principios del ANCORR fueron desarrollados por Benzécri (1964) y por Lebart & Tabart (1977). Aunque existen trabajos como son los de Richardson y Kuder (1933), Hirschfield (1935), Horst (1935), Fisher (1940) y Burt (1950) que señalan a Hirschfield como el padre del ANCORR. Los trabajos más actuales son los realizados por Hill (1974), Benzécri (1976), Greencre (1984) y Vargas (1986).

Esta técnica ha sido aplicada en el análisis de encuestas por conesa et al (1975), Alvarez, (1980) y recientemente Vargas (1986); tambien se ha aplicado en ecología hatheway (1971), Ibañez y Seguin (1972), Greencre y Urba (1984), Hill (1973 y 1974), Orloci (1975); en psicología Nishsisato (1980) y en otras áreas como sociología, medicina, linguistica y Antropología, Escofier (1969), Benzecri et al (1973) y Benzecri (1980).

MATERIALES Y METODOS.

Metodo de Muestreo.

Se consideró a la población de productores de cacao ubicados en los municipios de Cárdenas, Cunduacán, Huimanguillo, Jalpa de Méndez, Paraiso y Teapa.

La Unidad Estadística de muestro está representada por cada productor de cacao en esta región.

El diseño de muestreo fue el estratificado con afijación proporcional de Neyman, tomando como criterio de estratificación la superficie sembrada de cacao.

La población de cacaoteros se agrupo en 3 estratos :

Estrato 1. - Productores con menos de 5 has. sembradas de cacao.

Estrato 2. - Productores que tienen entre 5 y 10 has sembradas con cacao.

Estrato 3. - Productores que tienen más de 10 has cultivadas de cacao

Inicialmente se aplicó una encuesta piloto de 18 cuestionarios, con el objeto de determinar la variabilidad.

Posteriormente se procedió a estimar el tamaño de la muestra óptima adecuada, conociendo que el número de productores en cada estrato es de : 10800 1100 y 500 para los estratos uno, dos y tres respectivamente. Estableciendo un limite para el error de estimación B = 0.245 a a un nivel de confianza del 95% obtuvimos una presición en la estimación de D=0.015 has.

El tamaño de la muestra está dada por la forma.

$$n = \frac{\left\{ \sum_{i=1}^{3} N_i \sigma_i \right\}^2}{N^2 D + \sum_{i=1}^{3} N_i \sigma_i^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

donde i es el indice del estrato, i=1, 2, 3 .
N= 12400 en el tamaño total de población.
D representa la precisión de estimación.
y S_i es la desviación típica del estrato i.

El tamaño de muestra resultó 120 sin embargo, para mayor cobertura se encuestaron 129 de la siguiente forma: 112, 11 y 5 cuestionarios para lo estrato uno, dos y tres respectivamente.

VARIABLES A MEDIR.

Se midieron dos grupos de variables, las del primer grupo se refieren a los medios de comunicación (variables dependientes) y las del segundo grupo a las caracteristicas socioeconomicas del productor (variables independientes) y se listan a continuación:

Medios de comunicación: Periodicos, revistas agropecuarias, folletos agrícolas, Radio, Televisión, demostraciones agricolas, Exposiciones agropecuarias, prefiere los folletos, prefiere la radio prefiere al extencionista y prefiere otros medio.

Caracteristicas socioeconomicas: Edad , Años de cultivar cacao, tiempo de vivir en la zona, contacto con el extencionista, educación, vivienda con luz electrica, casa con paredes de material, piso de cemento, casa con baño, casa con más de dos cuartos, tamaño total de las parcelas, superficie sembrada, producción, viajan a México, Veracruz y Mérida, viajan a Villahermosa, viajan a Cardenas, Comalcalco Huimanguillo y Coatzacoalcos, viajan a paraiso, cunduacan, Jalpa y Tenosique, usan drenes, controlan enfermedades, fertilizan, podan, desean seguir cultivando el cacao criollo.

Los resultados se vaciaron en tres matrices de frecuencias (una para cada estrato), cada una pone en correspondencias a los renglones y columnnas donde los renglones son medios de comunicación y las columnas son las caracteristicas socioeconomicas.

DESCOMPOSICION DE LA ENERCIA TOTAL EN DOS EJES PRINCIPALES

La variación inherente en todas las variables la repartimos en dos ejes factoriales, considerando de cada eje las siguientes caracteristicas:

1.-LA CONTRIBUCION ABSOLUTA.- expresa la parte de variación que toma un elemento explicado por un factor α permitiendo saber cuales variables son las responsables de la construcción de un factor.

Sean CTR_α (i) las contribuciones del α factor en el i-elemento para los renglones y CTR_α (j) para las columnas.

$$CTR_\alpha(i) = r(i) * f_\alpha^2(i) \quad \ldots\ldots\ldots\ldots\ldots (2)$$

$$CTR_\alpha(j) = c(i) * g_\alpha^2(j) \quad \ldots\ldots\ldots\ldots\ldots (3)$$

donde f_α y g_α son elementos del i-ésimo vector de las matrices F y G.

2. - CONTRIBUCIONES RELATIVAS. - Expresan la variación que toma un factor para explicar la dispersión de un elemento al centro de gravedad.

Denotemos con $COR_\alpha(i)$ y con $COR_\alpha(j)$ a las contribuciones relativas del vector en el i-renglon y en la j-columna respectivamente.

$$COR_\alpha(i) = F_\alpha^2(i) \,/\, d^2\,(i,\ c) \quad \ldots\ldots\ldots (4)$$

$$COR_\alpha(j) = G_\alpha^2(j) \,/\, d^2\,(i,\ r) \quad \ldots\ldots\ldots (5)$$

donde $d^2(i,c)$ y $d^2\ (j,r)$ denotan la distancia ji-cuadrada del punto i al gravientro c y de la j-columna al promedio de las columnas respectivamente.

3. - CALIDAD. - Expresa en que medida un punto se encuentra explicado por los factores y representa la suma acumulada de contribuciónes relativas.

Para el caso de α dimensiónes la calidad de representación de i elementos es:

$$CALD_\alpha(i) = COR_1(i) + COR_2(i) + \ldots + COR_\alpha(i) \ldots\ldots\ldots (6)$$

Para representar a la j columna tenemos:

$$CALD_\alpha(j) = COR_1(j) + COR_2(j) + \ldots + COR_\alpha(j) \ldots\ldots\ldots (7)$$

4. - LA INERCIA RELATIVA. - representa la variación total del grupo de los renglones en relación al centro de gravedad c, y se define como:

$$INR\ (i) = r(i)\ \ f_\alpha^2(i) \,/\, \sum_{i=1}^{m} r(i)\ d^2(i,c) \ldots\ldots\ldots\ldots (8)$$

De igual modo la inercia relativa del j-elemento en el grupo de columnas esta dado por:

$$INR\ (i) = c(j)\ \ g_\alpha^2(j) \,/\, \sum_{j=1}^{n} c(j)\ d^2(j,r) \ldots\ldots\ldots\ldots (9)$$

ANALISIS ESTADISTICO

Se parte de dos grupos X y Y con variables indicadoras de la presencia o ausencia de alguna caracteristica de interés.
Las matrices de varianza y covarianza se definen por:

$$V_{xx}=1/n\ X^t\ X \quad \ldots\ldots\ldots\ldots (10)$$

$$V_{yy}=1/n\ Y^t\ Y \quad \ldots\ldots\ldots\ldots (11)$$

Para los renglones y columnas, respectivamente.

Las matrices de productos cruzados estan dadas por:

$$V_{xy}=1/n\ X^t\ X. \quad \ldots\ldots\ldots\ldots (12)$$

$$V_{yx}=1/n\ Y^t\ Y. \quad \ldots\ldots\ldots\ldots (13)$$

El análsis canónico propuesto por H. Hottelling (1936) propone resolver el sistema de ecuación siguiente.

$$V_{xx}^{-1}\ V_{xy}\ V_{yy}^{-1}\ V_{yx}\ U = \lambda U. \quad \ldots\ldots\ldots\ldots\ldots (14)$$

$$V_{yy}^{-1}\ V_{yx}\ V_{xx}^{-1}\ V_{xy}\ v = \lambda v. \quad \ldots\ldots\ldots\ldots\ldots (15)$$

Sea : $A=V^{-1}{}_{xx}\ Vxy\ V^{-1}{}_{yy}\ V_{yx}$

Entonces en el análsis canonico basta con resolver la ecuación caracteristica: $|A-\lambda I|$ ya que la solución es la misma para ecuaciones (14) y (15).

La solución al polinomio caracteristico resultante genera los valores caracteristicos

Para calcular las coordenadas de los renglones y columnas definimos una matriz de frecuencia relativas.

$$P= P(i,j)\ /\sum_{j=1}^{m}\ \sum_{i=1}^{n}(i,j) \quad \ldots\ldots\ldots\ldots\ldots (16)$$

donde m=número de renglones y n=número de columnas Los totales por renglón los podemos calcular de la siguiente

Los totales por renglón los podemos calcular de la siguiente manera:

$$r(i) = \sum_{j=1}^{m} p(i,j) \quad \ldots\ldots\ldots\ldots (17)$$

Donde $i = 1, 2, 3, \ldots, n$
Los totales por columna se calculan con:

$$r(j) = \sum_{i=1}^{n} p(i,j) \quad \ldots\ldots\ldots\ldots (18)$$

donde $j = 1, 2, 3, \ldots, m$

Una vez definidos los dos conjuntos de puntos tanto para renglones como para columnas, lo que prosigue es ajustar un subespacio para cada nube de puntos, con la condición de que cada subespacio tenga una correlación máxima. Los pasos que se siguen en este caso pueden consultarse en Greenacre (1984) o Bouroche & Saporta (1980).

Las coodenadas principales para los perfiles columnas estan dados por :

$$F = QU \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (19)$$

Donde U es la matriz formada por los vectores normalizados.
La matriz Q esta definida como:

$$Q = P(i,j) / r(i) \sqrt{c(j)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (20)$$

Del mismo modo, las coordenadas principales de los perfiles renglón son:

$$G = \hat{Q}t \; \hat{U} \quad \ldots\ldots\ldots\ldots\ldots\ldots (21)$$

$$\text{donde } \hat{U} = \bar{Q} \; U \; \Delta^{-1/2} \quad \ldots\ldots\ldots\ldots\ldots\ldots (22)$$

$$\bar{Q} = \frac{P(i,j)}{\{r(i)\,c(j)\}^{1/2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (23)$$

$$\hat{Q} = \frac{P(i,j)}{r(I)\; r(j)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (24)$$

$\Delta^{-1/2}$ es la matriz diagonal que contiene la inversa de la raiz cuadrada de los valores caracteristicos.

RESULTADOS Y DISCUSION

Los rsultados del ANCORR son mostrados en las graficas 1, 2 y 3 (ver anexo) para el estrato 1 2 y 3 respectivamente.

ESTRATO 1.- La grafica 1 muestra que los pequeños productores que no realizan la mayoria de las labores culturales tienden a estar en contacto con el extencionista y no preferir la radio como fuente de información agricola esto se debe quiza a lo que dice Bohlen et al 1961, que la nueva tecnologia la aplican al principio con recelo al principio al nivel experimental. Por otro lado el escaso nivel educativo les evita leer documentos y como estan en continuo contacto con la radio, adoptan la tecnologia recomendada por este medio ademas que les es mas comodo escuchar que leer.

ESTRATO 2.- Observando cuidadosamente la grafica 2 notamos que los medianos productores con rendimientos altos son los que prefieren la radio como fuente de información agricola (lo cual coincide de cierto modo con el estrato anterior) ademas de que leen periodicos locales y no viajan a México Veracruz y Mérida esto sucede por que tienen un nivel mas alto de educaión.

ESTRATO 3.- En la grafica 3 el ANCORR nos muestra que los producciones altas, realización de las labores culturales y deseo de cultivar el cacao criollo son caracteristicas dominantes de los grandes peoductores los cuales se asocian fuertemente con caracteristicas como no prefiere la radio como fuente de información agricola, prefieren leer los folletos agricolas, leer periodicos nacionales y asistir a exposiciones agropecuaria, lo que hace pensar que los conocimientos los obtiene por medio de leer folletos agricolas, periodicos nacionales y de asistir aexposiciones agropecuarias. Estas cuestiones suceden por que en este estrato la mayoria de los productores son personas con un nivel de educacion alto.

CONCLUSIONES.

Los pequeños y medianos productores adoptan con más facilidad la nueva tecnologia que se transmite por medio de la radio que con el contacto con el extencionista ya que cuando prefieren la radio como fuente de información agricola tienden mejorar sus rendimientos o labores en la plantación.

Los grandes productores son personas capacitadas que adoptan mejor la tecnologia a travez de los medios escritos tales como folletos, periodicos nacionales y exposiciones agropecuarias.

El medio de comunicacion que prefiere la mayoria de los productores de cacao del estado de Tabasco es la radio.

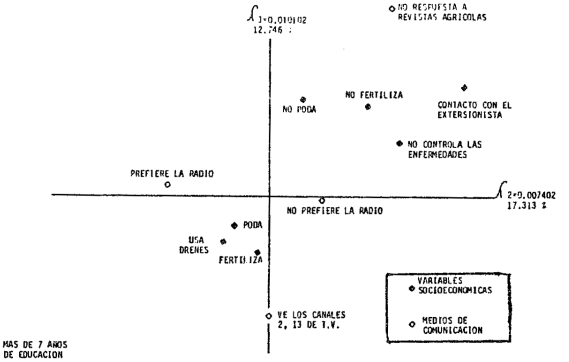
Es neceario hacer un estudio mas especifico sobre la radio y sobre los programas agropecuarios que desean escuchar los productores.

BIBLIOGRAFIA

1. -Alvarez, S. M. (1980) Una encuesta en la rama agropecuaria como instrumento para la investigación. (tesis doctoral). Instituto Superior de Ciencias Agropecuarias de la Habana. La Habana, Cuba, 130 pp.
2. -Benzécri, J. P and collaborators, (1973) L'Analyse des Donnees Tome 2: L'Analyse des Correspondances Ed. Dunod, Paris. France.
3. -Benzécri J. P. (1964) Cour de Linguistique Mathematique Publication Multigraphie Faculte des sciences de Rennes, France.
4. -Benzécri J. P. (1976) Hiestorie del' analyse des donnees. Les caniers de L' analyse des donnes (1) pp. 9-32, 101-120 y 343-376.
5. -Benzécri J. P. (1980) Analyse des correspondances: Exposee elementaire. Ed. Dunod, Paris.
6. -Bouroche, J. M. Saporta, G. (1980) L' analyse des donnes. Coleccion que sais-je? Presses Universitaires de France. p. 125.
7. -Bohlen, et al. (1961) Adopters of new farm ideas; Characteristics and comunications behavior. Agricultural Services. pp. 3-7. Ames, Iowa.
8. -Burt, C. (1950) The factorial analysis of cualitative data Bitain Jornal Statistical Psichilogy Bol. III (3) : 165-185.
9. -Conesa, A. P. et al. (1975) E'tude globale de la culture de la betterave a sucre sur le perimetre du Haut-Cheliff. I. annales Agronomiques, 26, 709-740. II. Analyse en Regresion. Annales Agronomiques. 27 (1) p. 61-84.
11. -Escofier-Cordier, B. (1965). L'Analyse des Correspondances. (tesis publicada en 1969). Cahiers du Bureau Universitaire Recherche Operatinnelle, No. 13.
12. -Fisher, R. A. (1940) The presicion of discriminante fuctions. Ann Eugen Lond 10:422-429.
13. -Fierro, G. y Quiroz, D. (1980) Estudio Comparativo de tres formas de comunicación para la transferencia de tecnología. Instituto Colombiano Agropecuario (ICA) Boletín de investigación No. 60. pp. 11-12. Colombia.
14. -Galizales, A. y Myren, T. (1967) Difunsión de la información Agrícola en el Valle del Yaqui. Secretaría de Agricultura y Ganadería. Instituto Nacional de Investigaciones (Folleto Técnico No. 52). México.

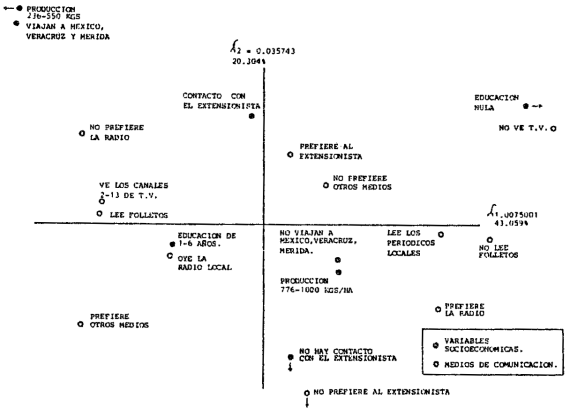
15. -Greenacre , M.J. and E.S. Urba (1984) Theory and applications of correspondance analysis Academic press. New York.
16. -Hatheway, W.H. (1971) Contingency table analysis of rain forest vegetation. In "Statistical Ecology", vol. 3: Many Species Populations, Ecosistems and systems Analisis (Patil, G.P., Pielou. E.C. and Waters. W.E., eds). Pennsylavania State University Park, Pennsylvania.
17. -H. Hotelling. Relations between two sets of variables. Revista Biometrika (1936) vol. 28.
18. -Hirschfield, H.O. (1935) A connetion between correlation and contingency. Proc Camb. phil. soc. 31:520-524 .
19. -Hill, H.O. (1973) Correspondance Annalysis: a neglected multivariate method of ordenation journal of ecology 61:237-251.
20. -Horst, P. (1935) Measuring Complex Attitudes. Journal Social Psichology 6, 369-374.
21. -Ibañez, F. y Seguin, G. (1972) Etude du cycle annuel du zoo plancton d'Abidjan. Comparison de plusieurs methodes d'analyse multivariable: composantes principales, correspondances, coordenees principales. Invest. Pesq. 36: 81-108
22. -Lebart, L., Morineau, A et Tabard, N. (1977) Techniques de la description statistique: methodes et logiciels pour l'analyse des grands tableaux. Paris, France; Ed. Dunod.
23. -Martinez, R. (1963) Difusión y adopción del maíz hibrido en cuatro municipios del estado de Guanajuato. tesis profesional. Escuela Nacional de agricultur. pp. 52-53. Chapingo Mexico.
24. -Nishisato, S. (1980) Analysis of Categorical Data: Dual Scalig and its Aplications. Toronto: University of Toronto Press.
25. -Orloci L. (1985) Multivariate analysis in vegetation research Dr. W. Junk B.W. Trhe Hauge Ix.
26. -Real, C.M. Boulen, J.M. (1968) Como aceptan los agricultores nuevas ideas. Bogotá IICA-CIRA, 228 p.
27. -Richardson, M. y Kuder, G.F. (1933) Making a rating scale that measures. Personnel J. 12: 36-40.
28. -Vargas, Ch. D. (1986) Aplicación del Analisis de correspondencias en el estudio de la interacción Medio ambiente-Vegetación, en el valle de Apatzingan. (tesis diplomada en estadistica aplicada). IIMAS, UNAM 107p. Mexico.
29. -Willson, M. C. y G. Gallup. (1960). Metodos de enseñanza en extensión y otros factores que influyen en la adopción de prácticas y de la economia del hogar. Centro Regional de Ayuda Técnica. Mexico. 347p.

ANEXO

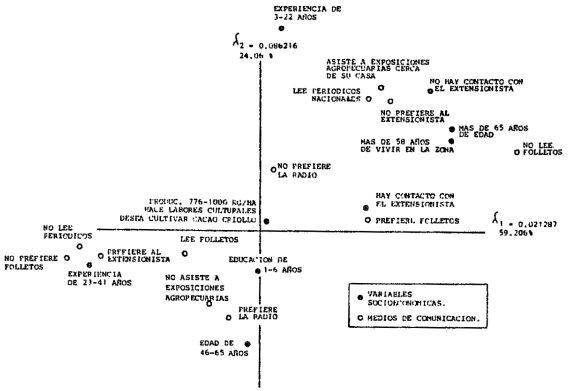


Grafica 4.- Representación bidimensional óptima mediante en análisis de correspondencias del estrato uno de productores de cacao, (Eje dos contra eje tres) CIFAP-TAB. 1988

Gráfica 2. Representación bidimensional óptima mediante el Análisis de Correspondencia del estrato dos de cacaoteros, (eje uno contra eje dos). CIFAP-TAB, INIFAP, SARH. 1985.



Gráfica 3. Representación bidimensional óptima mediante el Análisis de Correspondencia del estrato tres de productores de cacao. (eje uno contra eje dos). CIFAP-TAB, IMIFAP,SARH. 1985.



DISEÑO Y ANALISIS DE EXPERIMENTOS PARA DETERMINAR PATRONES DE CONDUCTA EN PRIMATES NO HUMANOS.[1]

Domínguez L. E.[2]
Facultad de Estadística (LINAE)
Universidad Veracruzana.
Ave. Xalapa esq. A. Camacho
Xalapa, Ver., México.

RESUMEN:

En este trabajo se presenta el diseño y análisis de un experimento realizado para determinar patrones conductuales, en cuanto a conductas agonísticas y afiliativas, y patrones de alimentación en monos aulladores. Los experimentos consideran dos grupos de monos en cautiverio y los métodos de registro de información se basan en la observación directa. La alimentación de los monos durante el experimento se basa en diferentes dietas de frutos cultivados y se pretende caracterizar la preferencia por alguno de los frutos. Se comparan intergrupal e intragrupalmente los patrones conductuales y de alimentación, en esta fase a nivel exploratorio. Se presentan algunas conclusiones importantes.

1 Como parte de los resultados del Proyecto de Translocación de poblaciones de mono aullador Alouatta palliata.

2 Alumna Investigadora del Laboratorio de Investigación y Asesoria Estadística. Universidad Veracruzana.

INTRODUCCION:

El presente trabajo forma parte de un proyecto que se lleva a cabo conjuntamente entre investigadores de la Estación de Primatología del Centro de Investigaciones Biológicas y el LINAE, ambos departamentos de la Universidad Veracruzana.

La inquietud surgió debido a que la colonización del Sureste de México está determinando la fragmentación, perturbación y desaparición de los bosques tropicales. Esta actividad humana está propiciando la desaparición de poblaciones silvestres de plantas y animales, así como el establecimiento de pequeñas áreas aisladas de bosque tropical, con diversos grados de perturbación, que son usadas como refugio natural por numerosas poblaciones silvestres de animales que huyen de la acción transformadora del hombre. Por tal razón, es posible encontrar poblaciones silvestres de monos aisladas en pequeños fragmentos de bosque tropical, donde sus expectativas de sobrevivencia son muy bajas, debido principalmente al empobrecimiento ecológico del ámbito hogareño. Son varias las amenazas que sufre estas poblaciones aisladas, como la de ser presa fácil para los cazadores que se dedican al tráfico de animales como mascotas, o bien que utilizan la carne y la piel para su comercialización.

De acuerdo con SEDUE, son tres las especies de monos mexicanos que se encuentran en peligro de extinción: Ateles geoffroyi (Mono araña), Alouatta palliata y Alouatta pigra (Mono aullador).

Las posibilidades de sobrevivencia para estas poblaciones de monos son reducidas. Ante esta situación, la Translocación de poblaciones silvestres de monos amenazadas en su habitat a áreas protegidas, constituye la única alternativa de supervivencia para estos animales.

DESARROLLO.

I.- EXPERIMENTOS DE TRANSLOCACION.

La TRANSLOCACION implica la transferencia de poblaciones silvestres de animales de un área a otra con un mínimo de tiempo en cautiverio y considera las siguientes etapas:

1.- Identificación de poblaciones silvestres de monos con bajas expectativas de sobrevivencia en habitat perturbado.
2.- Selección de poblaciones apropiadas para la translocación.
3.- Estudios prospectivos de las poblaciones candidatas para la captura y transferencia.
4.- Captura de las poblaciones de monos seleccionadas, bajo rigurosas medidas de seguridad.
5.- Estudios clínicos de los animales capturados.
6.- Estudios del comportamiento individual y social de los animales ya en cautiverio, así como de sus patrones alimenticios.
7.- Estudios ecológicos, etológicos y fisiológicos sobre las poblaciones silvestres bajo condiciones de cautiverio.
8.- Análisis de las áreas candidatas para la introducción o reintroducción, considerando fundamentalmente la capacidad forestal del ecosistema, así como la presencia humana en el área y garantías de conservación de la vida silvestre en la zona.
9.- Liberación de los monos translocados, la cual deberá realizarse bajo cuidadosa supervisión.
10.- Realización de estudios sobre ocupación y utilización del ámbito hogareño por los animales translocados. (Rodríguez, García y Canales.1987).

El proyecto mencionado en el primer párrafo considera la translocación de dos grupos de mono aullador (Alouatta palliata) y el trabajo que se presenta a continuación forma parte de una serie de estudios que se le realizaron a los dos grupos de monos en cautiverio. Los animales estuvieron, aproximadamente un año, en jaulas independientes y especialmente diseñadas, ubicadas en la Isla de Totogochillo, en la Laguna de Catemaco. Los resultados de este estudio fundamentarán hipótesis de trabajo para estudios de comportamiento de Alouatta palliata en semilibertad.

II.- OBJETIVOS DE ESTE ESTUDIO.

II.1.- OBJETIVO GENERAL:

Mediante este trabajo se pretende explorar la relación existente entre el consumo de las frutas, y los desplazamientos conductuales, ya sean agonísticos o afiliativos; así como las diferencias entre animales y entre grupos, referentes a las variables y las relaciones.

II.2.- OBJETIVOS PARTICULARES:

1.- Establecer patrones de consumo y de preferencia con respecto a los diferentes tipos de frutas cultivadas, para cada animal y para cada grupo.

2.- Establecer índices de conductas agonísticas para cada animal y para cada grupo.

3.- Establecer las relaciones existentes entre consumo de alimento y conductas emitidas.

III.- DESCRIPCION DEL EXPERIMENTO.

Se consideraron dos grupos de mono aullador (Alouatta palliata) en cautiverio. El primero compuesto por un macho y seis hembras, y el segundo, por un macho y cuatro hembras. El método de resgistro que se utilizó fué el de observación directa ("Ad libitum"), mediante focales de 5 minutos por animal, durante media hora en dos turnos (10:00 hrs. y 17:00 hrs.) diarios a lo largo de 30 días consecutivos. Además se necesitó de un sistema de video para registrar y contabilizar los eventos conductuales que se presentaron.

La dieta escogida para el experimento consistió en diferentes raciones de tres tipos de frutas cultivadas: piña, plátano y sandía, pretendiéndose caracterizar con esto la preferencia por alguna de estas. La conformación de la dieta fué única a lo largo del experimento.

La selección de los monos para la observación se realizó de manera aleatoria y con reemplazo, en cada grupo y en cada turno.

IV.- HIPOTESIS DE INTERES.

Se tiene como hipótesis que no hay diferencia en los patrones de consumo alimenticio entre los dos grupos pero que sí la hay entre los individuos. El investigador considera que los machos de ambas jaulas presentan un consumo mayor que cualquiera de las hembras, así como también, que las conductas que más se registran durante el aprovisionamiento son las agresivas por parte de los animales dominantes, y en menor grado, las defensivas. Con esto último se pretende ver si hay algún tipo de orden jerárquico que rija a esta especie de monos. Cabe señalar que hasta la fecha no se tienen resultados reportados respecto al comportamiento y dieta de esta especie.

V.- VARIABLES CONSIDERADAS.

Las variables consideradas para el experimento en cada turno son las siguientes: la cantidad de piña, plátano y sandía consumida por animal durante los 5 minutos de observación, así como el número de eventos de tipo agonístico (ataque y amenaza) y de tipo defensivo (apaciguamiento y huída) que emite y recibe cada individuo.

METODOLOGIA, RESULTADOS Y CONCLUSIONES.

Se obtuvieron sumas generales de las variables de mayor interés, como la cantidad de alimento consumido y los desplazamientos conductuales que registraron los animales durante el experimento. Se hicieron gráficas comparativas de estas variables, entre animales y entre grupos. Posteriormente se obtuvieron correlaciones simples entre 5 variables (consumo de piña, plátano, sandía, número de conductas agresivas y defensivas), por animal y por jaula, de lo que se siguen las conclusiones siguientes, al evaluar la significancia de las correlaciones probando la hipótesis Ho: $\rho=0$ contra Ha: $\rho\neq0$. (Yamane 1979).

De las Correlaciones, tenemos:

El Macho de la jaula 1, come más piña a medida de que consume más sandía; además, mientras más come plátano, emite mayor número de conductas defensivas. Podemos decir con esto que compite más con las hembras por el consumo de plátano.

Para "la Güera", hembra joven del grupo 1, la única correlación significativa que presenta se refiere a las conductas que emite, ya que agrede a los demás al mismo tiempo que se defiende de ellos. Es una hembra muy competitiva en todos los roles.

"Mony" es una hembra adulta, también de la jaula 1, cuyas correlaciones significativas nos permiten decir que, conforme consume más plátano, incrementa el consumo de la sandía, siendo más agresiva cuando consume el primero. En cuanto a conductas, se defiende en igual forma que ataca.

"Luisa", es otra hembra adulta de la jaula 1, presenta lo siguiente: conforme aumenta el consumo de piña, aumenta también el de plátano y agrede más cuando más se alimenta de éste último.

"Petra", también integrante del grupo 1, no presenta ninguna correlación significativa, lo que quiere decir que su patrón de alimentación y conductual es muy aleatorio.

"Paty" es la hembra más dominante de este grupo (el 1) y su patrón de comportamiento nos dice que mientras consume más piña, agrede en gran medida a sus compañeras.

La última hembra de esta jaula es "María" que, según los resultados, disminuye su consumo de sandía conforme aumenta el de piña y es atacada más cuando consume ésta última.

Del Macho de la jaula 2, la correlación más importante es la nos indica que consume más piña emitiento más conductas agresivas.

La segunda integrante de este grupo es "Chabela" la cual, conforme aumenta su consumo de piña, decrementa el de plátano, y sus conductas de agresividad y defensa están correlacionadas positivamente.

Otra hembra es "La Niña", que al comer más plátano,agrede y a su vez se defiende más.

"La Pinta", es una hembra joven que integra también este grupo. La única correlación significativa nos dice que al consumir mayor cantidad de plátano, ataca más a sus compañeras.

Y por último tenemos a "Santamaría" que, al igual que "Petra", la mona de la jaula 1, presenta un patrón de consumo alimenticio y de conductas muy aleatorio.

En general, para ambos grupos tenemos:

En el grupo 1, compuesto por los 7 animales citados primeramente, se puede observar que a medida de que se consume más piña, el consumo de plátano y sandía decrementa, y cuando consumen plátano, emiten más conductas defensivas.

Para el grupo 2, el de 5 individuos, se presenta lo siguiente: conforme es mayor el consumo de piña, disminuye el de plátano y de sandía, detectándose además que entre más se atacan, más se defienden.

De las Gráficas Comparativas, se tiene que:

En las gráficas de consumo total de frutas, podemos apreciar, para la jaula 1, (Fig. 1), que la más consumida es la piña, seguida del plátano y la sandía, observándose una preferencia más o menos similar. El macho de esta jaula es el que consume mayor cantidad de piña y plátano pero no sandía, ya que de ésta, son las hembras la que la consumen más. En la gráfica del grupo 2, (Fig. 2), se nota claramente el orden de preferencia que tienen los monos, siendo en primer lugar, la piña, después la sandía y por último el plátano. Para este grupo, el macho es el que consume más plátano y más sandía, pero es "la Niña" la que come más piña.

Del segundo tipo de gráficas comparativas que se obtuvieron, las de conductas recibidas por animal y por grupo, de la jaula 1 (Fig. 3), se tiene que hay una gran cantidad de conductas amenazantes recibidas por casi todos los animales, excepto por el macho, y de las defensivas, son las de apaciguamiento, siendo "Paty" la que mayor número recibe por parte de sus compañeras. El patrón conductual que presenta el grupo 2 (Fig. d), nos dice que el apaciguamiento es la conducta que más se observa en los 5 individuos, y es a la "Pinta" y al Macho a los que más apaciguan las otras hembras. A diferencia del grupo anterior, aquí se contabilizaron mayor número de huídas.

El último tipo de gráficas es el de las conductas observadas en los dos turnos para ambas jaulas, teniéndose, de la primera, (Fig. 5), que es en el turno 2 donde mayor actividad conductual se registra, habiendo, para amenaza y apaciguamiento, una cantidad más o menos similar de conductas observadas, y para el grupo 2, (Fig. 6), dicha actividad se da mayormente en el turno 1.

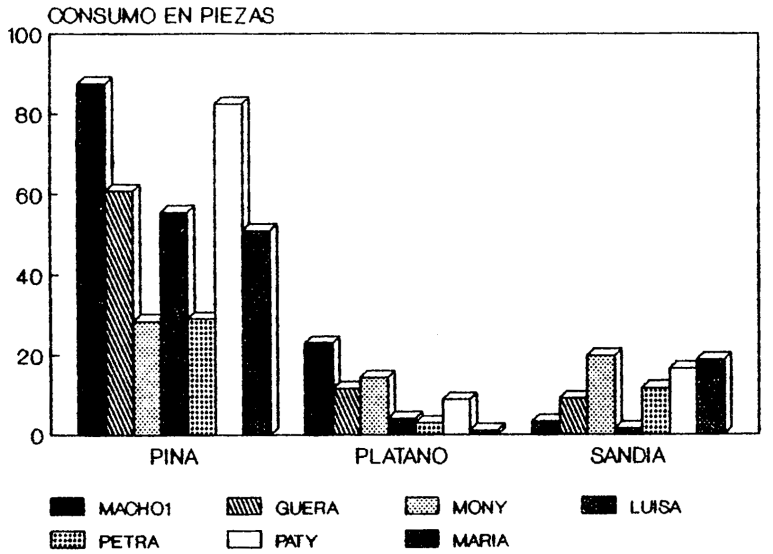


Fig. 1. Consumo total de frutas por animal para la Jaula 1

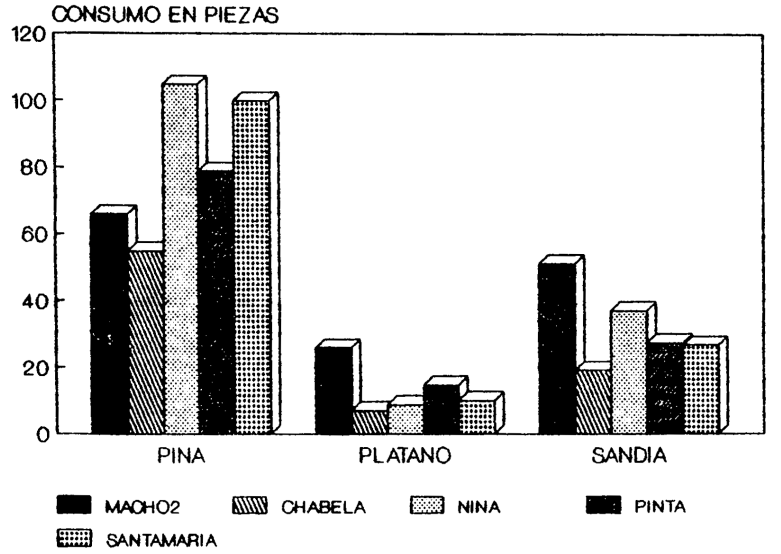


Fig. 2. Consumo total de frutas por animal para la Jaula 2

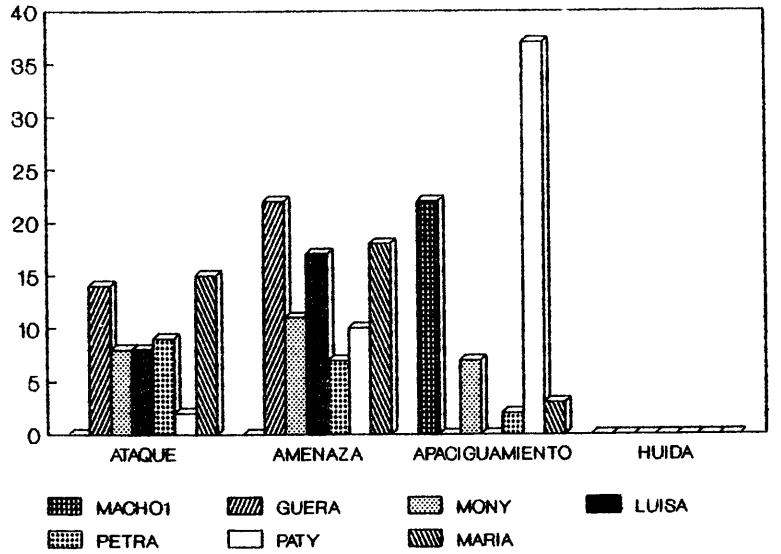


Fig. 3. Conductas recibidas por animal para la Jaula 1.

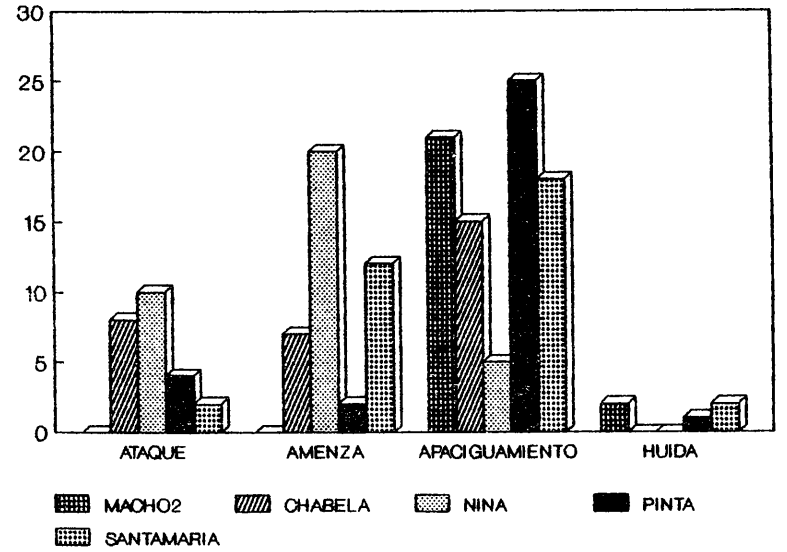


Fig. 4. Conductas recibidas por animal para la Jaula 2.

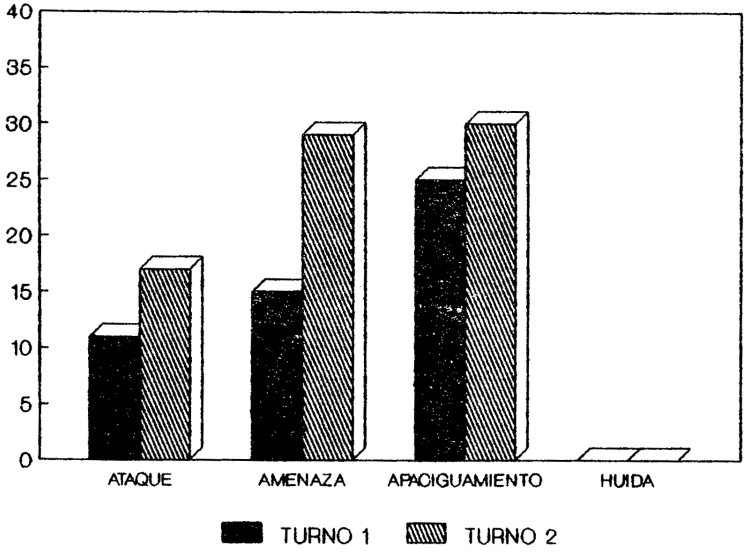


Fig. 5. Conductas observadas por turno para la Jaula 1.

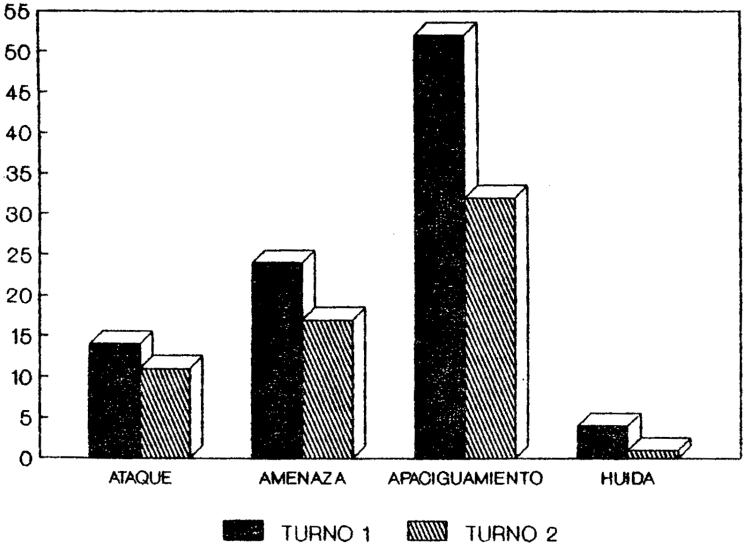


Fig. 6. Conductas observadas por turno para la Jaula 2.

OBSERVACIONES FINALES.

El estudio de secuencias conductuales se realizará a través de Cadenas Markovianas. Actualmente se revisan los videos tomados para obtener la información necesaria para el modelaje.

AGRADECIMIENTOS.

Agradezco al Biólogo Ernesto Rodríguez Luna y al Personal del Centro de Primatología su apoyo y colaboración para la realización de este trabajo, así como a Mario Miguel Ojeda por su dedicación , asesoramiento y paciencia para la revisión de las fases de este proyecto. La Información para este trabajo fué recabada conjuntamente por la autora, y Cristina cuevas, por lo que agradezco su colaboración.

BIBLIOGRAFIA.

Rodríguez Luna E., Fa. J. E., García Orduña F., Silva López G. y Canales Espinoza D. 1987. Primate conservaqtion in Mexico.Primateconservation,8;114-118.

Subcommitee on Conservation of Natural Populations, Commitee on Nonhuman Primates, Division of Biological Sciences, Assembly of Life Sciences and National Research Council. 1981. TechniquesfortheStudyofPrimatePopulationEcology. National Academy Press, Washington, D. C.

Krebs J. R. & Davies N. B. 1987. AnIntroductiontoBehaviural Ecology. Sinauer Associates, Inc. Publishers Sunderland Massachusetts.

Hinde A. R. 1983. Primate Social Relationships. Sinauer Associates Inc. Publishers, Sunderland, Massachusettes.

Yamane T. 1979. Estadística.Harla.

UN PROCEDIMIENTO PRACTICO PARA ESTUDIAR LA CORRELACION ENTRE DOS VARIABLES CIRCULARES Y ENTRE UNA VARIABLE CIRCULAR Y UNA LINEAL.

I. OSEGUERA Y M. M. OJEDA
Fac. de Est. (LINAE) Fac. de Est. (LINAE)

RESUMEN

Se dan algunas observaciones de tipo práctico para estudiar la correlación entre dos variables angulares y entre una circular y una lineal. Se hacen algunas observaciónes sobre el estimador de varianza jackknife de Tukey para estos casos.

INTRODUCCION

Jonhson y Wherly (1977) mencionan que, existen varias medidas muestrales no-paramétricas para medir la dependencia en variables que involucran medidas angulares.

La correlación canónica técnica del análisis multivariado ayuda a unificar una cantidad substancial de trabajos anteriores sobre correlación angular.

Aquí se utilizan los criterios propuestos por Jonhson y Wherly (1977) donde aparecen expresiones no cerradas para la varianza del coeficiente de correlación entre una variable angular y una lineal (ρ_{AL}) y para la del coeficiente de correlación entre dos variables angulares (ρ_{AA}). Con respecto a esto aquí se ha aplicado el método de jackknife, con lo que se propone una metodología alternativa y de fácil implementación.

CORRELACION ENTRE DOS VARIABLES CIRCULARES

Sea $(\theta_1,\theta_2)'$ un vector aleatorio bivariado de variables aleatorias circulares. Y sea el vector $(\underset{\sim}{X}_1',\underset{\sim}{X}_2') = (SEN\ \theta_1, COS\ \theta_1, SEN\ \theta_2, COS\ \theta_2)$ el correspondiente vector tetravariado. Tomemos a $(X_1 : X_2) = X$ como la matriz de datos y sea

$$S = \begin{bmatrix} S_{11} & S_{12} \\ \cdots & \cdots \\ S_{21} & S_{22} \end{bmatrix} = \left[\begin{array}{cc:cc} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ \hdashline s_{31} & s_{32} & s_{33} & s_{34} \\ s_{41} & s_{42} & s_{43} & s_{44} \end{array}\right]$$

la matriz de varianzas y covarianzas asociada a X particionada de tal manera que S_{11} sea la matriz de varianzas y covarianzas de $\underset{\sim}{X}_1$, S_{22} la de $\underset{\sim}{X}_2$ y S_{12} la matriz de covarianzas entre los elementos de $\underset{\sim}{X}_1$ y $\underset{\sim}{X}_2$.

El coeficiente de correlación muestral entre las variables circulares es la correlación canónica principal; es decir la correlación entre

$$V_1 = u_1'\underset{\sim}{X}_1 \quad y \quad W_1 = u_2'\underset{\sim}{X}_2$$

y no es otra cosa que la raiz cuadrada del primer valor característico asociado a cualquiera de las matrices A_1 o A_2 donde $A_1 = S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}$ y $A_2 = S_{22}^{-1} S_{21} S_{11}^{-1} S_{12}$, donde u_1 es el primer vector característico normalizado asociado a A_1 y u_2 es el primer vector característico asociado a A_2. Respecto de esto Jonhson y Wherly (1977) mencionan "Una característica importante práctica del método de correlación canónica para hallar la correlación entre variables circulares es que uno puede calcularla usando programas estadísticos estandar. Para correlación canónica".

Realmente la correlación entre variables angulares se define como $\rho_A = \underset{\alpha_1,\alpha_2\in(0,2\pi)}{Sup}\ \rho(\cos(\theta_1-\alpha_1), Cos(\theta_2-\alpha_2))$ y se muestra que este problema puede ser traducido al problema definido antes, el de hallar la correlación canónica entre $\underset{\sim}{X}_1$ y $\underset{\sim}{X}_2$ (mencionado por Jonhson y Wherly (1977)).

Los ángulos α_1 y α_2, los cuales definen la máxima correlación, pueden hallarse transformando a α_1 y α_2 a coordenadas polares.

la correlación angular es una correlación rotacional, por lo tanto la correlación perfecta sucede si y solo si θ_1 es una rotación con reflexión sobre θ_2; es decir

$$\rho_A = 1 \quad \text{implica que} \quad (\theta_1-\alpha_1) = \pm(\theta_2-\alpha_2)(\text{mod } 2\pi)$$

Como aspecto importante debemos señalar que relación perfecta no implica correlación perfecta, pero independencia implica $\rho_A = 0$.

Puede mostrarse siguiendo algunos pasos algebraicos que

$$r_A = [(-C_2 + (C_2^2 - 4C_1C_3)^{1/2})/2C_1]^{1/2}$$

donde

$$C_1 = (S_{12}^2 - S_{11}S_{22})(S_{34}^2 - S_{33}S_{44}),$$

$$C_2 = -(S_{11}S_{23} - S_{12}S_{13})(S_{44}S_{23} - S_{24}S_{34}) - (S_{33}S_{24} - S_{23}S_{34})$$
$$-(S_{22}S_{13} - S_{12}S_{23})(S_{44}S_{13} - S_{14}S_{34}) - (S_{22}S_{14} - S_{12}S_{24})$$
$$(S_{11}S_{24} - S_{12}S_{14})$$
$$(S_{33}S_{14} - S_{13}S_{34})$$

$$C_3 = (S_{12}S_{24} - S_{14}S_{23})^2$$

CORRELACION ENTRE UNA VARIABLE CIRCULAR Y UNA LINEAL

Sea (θ, Y) un vector aleatorio bivariado y sea $(\cos\theta, \operatorname{Sen}\theta, Y)$ el vector trivariado formado al transformar la media angular en su seno y coseno. Tomemos a $(X_1 : X_2) = X$ como la matriz de datos y sea

$$S = \begin{bmatrix} S_{11} & S_{12} \\ \cdots & \cdots \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ \cdots & \cdots & \cdots \\ S_{11} & S_{32} & S_{33} \end{bmatrix}$$

la matriz de varianzas y covarianzas entre $(\cos\theta, \operatorname{Sen}\theta, Y)$. Entonces el coeficiente de correlación canónica requerido entre $(\cos\theta, \operatorname{Sen}\theta)$ y Y es la raiz cuadrada positiva de

$$r_{AL}^2 = (S_{12}S_{23}^2 + S_{22}S_{13}^2 - 2S_{12}S_{23}S_{31})/(S_{33}(S_{11}S_{22} - S_{12}^2))$$

Se puede mostrar que esta es una solución equivalente para el problema de hallar el coeficiente de correlación lineal angular estimado, dado que $\rho_{AL} = \max_{\alpha_1, \alpha_2 \in [0, 2\pi)} \rho(\cos(\theta - \alpha), Y)$.

También puede verse a la correlación lineal-angular como el coeficiente de correlación canónica principal entre (Cos θ, Sen θ) y Y.

Ejemplo

En un estudio que tiene como objetivo predecir velocidad y dirección del viento en superficie a través de información de velocidad y dirección del viento a 500mlb. Se estudio la correlación entre la velocidad y dirección del viento en superficie; se tienen las siguientes observaciones.

Dir. del viento en dg.	338.4	357.5	337.3	343.8	333.5	342.3	348.5
Velocidad del viento.	7.54	1.76	9.43	4.71	5.41	7.30	2.90
Dir. del viento en dg.	356.1	198.4	190.5	353.6	183.8	181.8	190.9
Velocidad del viento.	2.30	1.99	1.63	6.08	1.68	1.94	2.29
Dir. del viento en dg.	187.4	356.4	337.3	337.3	358.9		
Velocidad del viento.	1.65	4.00	11.18	7.41	4.63		

EL JACKKNIFE PARA ESTOS CASOS

El método jackknife para estimar la varianza de un parámetro de interés resulta ser un método de fácil implementación. Este método puede proveernos de estimaciónes no paramétricas para la varianza y el sesgo.

Sea $\hat{\theta}(X1, X2, \ldots, Xn)$ un estadístico de interés, donde X_1, X_2, X_n son variable aleatorias independientes e identicamente distribuidas, y $\theta(X_1, X_2, \ldots, X_n)$ es invariante bajo permutaciónes de los argumentos.

El estimador de varianza jackknife, $\hat{Var}\ \hat{\theta}(X_1, X_2, \ldots, X_n)$, es definido en términos de las cantidades $\theta_{(i)} \equiv \theta(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, el valor de θ cuando X_i es excluido de la muestra de la siguiente forma:

$$\hat{Var}\ \hat{\theta}(X_1, X_2, \ldots, X_n) = ((n-1)/n) \sum_{i=1}^{n} [\hat{\theta}_{(i)} - \hat{\theta}_{(.)}]^2$$

donde $$\hat{\theta}_{(.)} \equiv \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$$

que son estimaciones propuestas en [2].

Efron y Stein (1981) mencionan que el estimador de varianza jackknife es conservador al estimar la varianza; es decir $E(Var\ \hat{\theta}(X_1, X_2, \ldots, X_{n-1})) \geq Var\ \hat{\theta}(X_1, X_2, \ldots, X_{n-1})$, para cualquier función simétrica $\hat{\theta}(X_1, X_2, \ldots, X_{n-1})$; y se menciona que de echo ni la simetría ni la distribución idéntica para X_i son esenciales.

También se hace la observación de que el estimador de varianza jackknife es más conservador cuando $\hat{\theta}(X_1, X_2, \ldots, X_n)$ es una función cuadrática o de mayor orden.

Con estos resultados construimos un programa computacional que dado uno de los dos problemas mencionados evalua el estimador jakknife y el estimador de la varianza. Asumiendo los resultados de la teoría asintótica hemos construido intervalos de confianza.

A continuación se dan los intervalos de confianza obtenidos mediante este método de estimación de varianza para los ejemplos dados por John y Wherly (1977).

Para la correlación lineal-angular $r_{LA} = 0.72$ el intervalo de confianza obtenido es: $\alpha = 0.05$ $0.476 < \rho_A < 0.969$

Para la correlación Angular-Angular $r_{AA} = 0.57$ el intervalo de confianza obtenido es: $\alpha = 0.05$ $0.2058 < \rho_A < 0.9272$

Para el ejemplo en este artículo $r_{AL} = 0.817$ el intervalo de confianza obtenido es $\alpha = 0.05$ $.7116 < \rho_{AL} < 0.9231$

RESULTADOS Y CONCLUSIONES

Después de haber observado los intervalos de confianza obtenidos por Jhon y Wherly (1977) y los obtenidos mediante el metodo propuesto aquí, puede observarse empiricamente la eficiencia de este último.

Como se mencionó en (4) el estimador de varianza jackknife sobreestima la varianza verdadera; aun así los intervalos obtenidos aqui son más pequeños que los obtenidos por John y Wherly

La corrección de sesgo no se realizó en esta aplicación pués la naturaleza de los datos dan a la varianza un sesgo muy pequeño.

REFERENCIAS

[1] RICHARD, A. J. Y THOMAS, WHERLY (1977)Measures and Models for Angular Correlation and Angular-Linear Correlation. Journal of the Royal Statistical Society, Ser. B, 39, 222-229.

[2] Efron, B. y Stein, C. (1981) Jackknife Estimator Variance. The Annals of statistics, Vol. 9, No. 3, 586-596.

[3] Louis-Paul Rivest (1982) Some Statistical Methods for Bivariate circular Data. Journal of the Royal Statistical Socity, Ser. B, 44, No.1, 81-90.

DIAGNOSTICO Y ESTIMACION EN MODELOS DE REGRESION LINEAL MULTIPLE APLICADOS A PROBLEMAS METEOROLOGICOS.[2]

OJEDA M. M.[1]
Fac. de Estadística (LINAE)
Av. Xalapa esq. A. Camacho
Xalapa, Ver. Mexico.

RESUMEN:

Se presentan los resultados de modelar la dirección y la velocidad del viento en superficie a partir de la dirección y la velocidad observada a 500 mlb., considerando las condiciones meteorológicas (Gradiente térmico, Situación sinóptica, Nubosidad, Estabilidad atmosférica) y como variable indicadora el periódo de observación (diurno y nocturno). El modelaje se realizó a través de un modelo de regresión lineal múltiple bivariado y también a través de modelos univariados. Se presentan la exploración de supuestos, la detección de puntos de influencia, las estimaciones de mínimos cuadrados ordinarios y las ecuaciones resultantes de aplicar procedimientos de estimación robusta [Mosteller y Tukey (1977)] ponderando las observaciones. Algunas conclusiones y recomendaciones generales se incluyen con el propósito de ubicar el avance de esta investigación.

2 Como parte de un proyecto a cargo de Adalberto Tejeda, Ana Delia Contreras (Fac. de Física U. V.), María Cristina Ortiz, Isaías Oseguera y Mario Miguel Ojeda (LINAE, Fac. de Estadística U. V.).

1 Maestro de la Facultad de Estadística y Coordinador de Investigación del Laboratorio de Investigación y Asesoría Estadística de la Universidad Veracruzana.

INTRODUCCION.

Es de gran interés en meteorología contar con modelos que permitan predecir la velocidad y la dirección del viento en la superficie de una región dada a partir de la velocidad y la dirección a 500 mlb., y considerando condiciones meteorológicas generales. Esto en razón de que la información sobre las condiciones meteorológicas generales y la dirección y la velocidad del viento a 500 mlb. se pueden conocer con bastante precisión hasta con 72 horas de anticipación [Jhonson y Kalma (1986), Cervantes (1987), Tejeda y Cervantes (1988)].

En trabajos anteriores [Tejeda y Cervantes (1988)] se ha predicho la dirección del viento en una región dada a través de modelos de regresión lineal múltiple, considerando condiciones meteorológicas generales y dirección del viento en altura a 800 mlb., obteniendo resultados bastante satisfactorios. La línea metodológica de este trabajo se sigue de Jhonson y Kalma (1986). Sin embargo, la dirección y la velocidad del viento son dos variables que observan correlaciones significativas desde el punto de vista estadisitico [Oseguera (1988)], y además se tiene una explicación física-meteorológica para este hecho [Contreras (1988)].

VARIABLES, DATOS Y METODOLOGIA.

En la región de Laguna Verde, Ver. se tienen 6 estaciones meteorológicas que registran cada hora la velocidad y la dirección del viento en superficie, entre otras variables de interés para la meteorología. Mediante un estudio preliminar [Contreras (1988)] se ha determinado que un promedio espacio-temporal por cada turno (diurno y nocturno) es un buen indicador de la dirección y la velocidad en superficie para esa región. Para el caso de la variable dirección se ha tomado el promedio circular [Mardia (1972)] El interés de predecir la dirección y la velocidad del viento en la región de Laguna Verde nace del hecho de que estas variables son parámetros en el modelaje de la difusión de contaminantes radioactivos, que se requiere realizar para prevenir accidentes en la Planta Nucleoeléctrica.

Denotemos a las variables dirección y velocidad del viento en superficie por Y_1 y Y_2 respectivamente. Se tienen también para cada uno de dos turnos, registros de las siguientes variables: X_1=

Dirección del viento a 500 mlb., X_2= Periodo (0= diurno, 1= nocturno), X_3= Estabilidad atomosférica, X_4= Velocidad del viento a 500 mlb., X_5= Situación sinóptica, X_6= Nubosidad y X_7= Gradiente térmico.

Sea (Y!X) la matriz de datos de orden 288 x 10, donde $X=[1!X_{(1)}!...!X_{(7)}]$. Las variables de dirección dadas en grados fueron linearizadas a través de la siguiente transformación:

$$\theta' = \begin{cases} \theta \text{ si } 0 \leq \theta \leq 180 \\ \theta - 360 \text{ si } 180 \leq \theta \leq 360 \end{cases}$$

considerando la convención de que viento del norte implica $\theta = 0$ y viento del este implica $\theta = 90$, y así sucesivamente.

Entonces podemos postular el problema de modelaje a través del modelo lineal general $Y = XB+E$, donde $B = [B_{(1)}!B_{(2)}]$ es la matriz de parámetros y $E=[E_1',E_2',...E_n']$ es la matriz de errores.

Los supuestos se pueden resumir en:

$$E_i \sim N_2(0,\Sigma) \quad i = 1,2,...,n$$

con distribuciones independientes.

El estimador de mínimos cuadrados es [Seber (1985)]:

$$\hat{B} = (X'X)^{-1}X'Y = CY$$

Sea $H=XC$, entonces $\hat{Y}=[\hat{Y}_{(1)}!\hat{Y}_{(2)}]=HY$. Sea $\hat{E}=[\hat{E}_{(1)}!\hat{E}_{(2)}]=Y-\hat{Y}$.

Una vez dados los datos (Y!X) bajo la situación observacional especificada y dado el esquema teórico de modelaje presentado, desearíamos saber las cualidades del modelo ajustado. Algunas preguntas interesantes a contestar son:

a) Qué tan razonables son las suposiciones para esta situación y para estos datos particularmente?

b) Qué tanta precisión tiene el modelo para estimar? Qué tan estable es?

c) Son necesarias todas las variables independientes para lograr el grado observado de bondad en las estimaciones?

d) Alguna o algunas observaciones particulares afectan de manera significativa la estructura (parámetros estimados) del modelo para los datos dados?

El diagnóstico en modelos de regresión tiene que ver con las respuestas a estas preguntas, aunque tal vez las técnicas de diagnóstico en el sentido de la definición de Besley, Kuh y Welsch (1980) se refieren mas especificamente a las preguntas en b) y d).

Nosotros usamos $\hat{E}$ y $\hat{Y}$ para explorar la normalidad y la homocedasticidad [Gunst y Mason (1980)] y $H=(h_{ij})$ para explorar la influencia de las observaciones, según el criterio de que Y_i es un punto de influencia si $(2p/n) < h_{ii}$ [Hocking (1983)]. A demás obtuvimos indicadores de multicolinealidad y verificamos la estabilidad de las estimaciones usando la descomposición espectral de (X'X) [Besley, Kuh y Welsh (1980)].

Se realizaron gráficas de valores residuales contra valores predichos con el criterio de mínimos cuadrados y una gráfica correlograma para estudiar la distribución conjunta de los residuales [Anscombe y Tukey (1963)].

Obtuvimos modelos generales como aproximaciones preliminares. Además, dado que el grado de precisión al predecir la dirección es mayor, se ha implementado el algoritmo de mínimos cuadrados ponderados usando tres criterios de ponderación descritos en Mosteller y Tukey (1987) pag. 341-351, considerando cinco iteraciones, para obtener estimaciones robustas mejoradas de la dirección en superficie.

RESULTADOS.

Las matrices de sumas de cuadrados y productos cruzados del error y total, respectivamente, del modelo ajustado a través de mínimos cuadrados ordinarios, son como siguen:

$$T=\begin{bmatrix} 2605078.59 & \\ 3067.724 & 761.54 \end{bmatrix} \qquad \hat{E}'\hat{E}=\begin{bmatrix} 953329.84 & \\ -3266.36 & 641.796 \end{bmatrix}$$

Las correlaciones múltiples y los coeficientes de determinación, simple y ajustada respectivamente, se muestran en la tabla siguiente:

	Variable Y_1	Variable Y_2
r	0.796	0.397
r^2	0.634	0.157
r^2_{aj}	0.623	0.131

Tabla 1. Coeficientes de correlación múltiple, de determinación y de determinación ajustados para el modelo a través de mínimos cuadrados ordinarios.

Al analizar los residuales encontramos posibles puntos de influencia [Ver figura 1 y figura 2], y sospechamos de fallas en los supuestos básicos del modelo. A través del criterio de la entrada de la matriz H, eliminamos algunas observaciones y volvimos a ajustar el modelo a través de mínimos cuadrados ordinarios, obteniendo que los resultados no mejoraban sustancialmente. También verificamos los indicadores de multicolinealidad (factores de inflación de varianza y valores característicos de la matriz $(X'X)^{-1}$), sin encontrar indicios de problemas de inestabilidad de las estimaciones por esta razón.

La gráfica 1 nos muestra dos nuves de puntos en razón de la variable indicadora turno, la cual define esencialmente dos modelos. Aparentemente no existe evidencia de tendencias o patrones específicos en los residuales, salvo la identificación de puntos extremos.

En la gráfica 2 podemos identificar una violación clara del supuesto de homocedasticidad y posiblemente dos distribuciones de los errores mezcladas, una con alta variabilidad correspondiente a las velocidades altas y otra con baja dispersión correspondiente a las velocidades menores.

La gráfica 3 nos muestra la figura de, aparentemente, dos distribuciones bivariadas con puntos "fuera de contexto" y con correlaciones, la primera negativa y la segunda positiva. Puede pensarse de ver la figura, que las anomalías de la distribución de los residuales de la velocidad impactan a la distribución bivariada. También la influyen los puntos extremos a la izquierda y a la derecha sobre los residuales de la dirección.

Usando el método de mínimos cuadrados ponderados considerando el criterio de ponderación doble de Tukey, el método de la aplanadora y la ponderación según el seno de Andrews, ajustamos un modelo para predecir la dirección, realizando cinco iteraciones obtuvimos los resultados que se muestran en la tabla siguiente:

TERMINO	MCO	METODO DE TUKEY	SENO DE ANDREWS	METODO DE LA APLANADORA
CONSTANTE	-102.1	-110.95	-111.03	-111.17
X_1	-.03	-.02	-.02	-.02
X_2	154.79	156.01	155.91	152.77
X_3	5.55	7.07	7.09	7.23
X_4	-.01	-.07	-.06	-.04
X_5	.168	1.83	1.86	1.84
X_6	-28.5	-29.5	-29.5	-31.8
X_7	-3.9	-5.4	-5.4	-5.8

Tabla 2. Coeficientes estimados en el modelo de regresión para predecir la dirección usando métodos de ajuste robusto con cinco iteraciones.

No se realizó el ajuste bivariado usando los métosos de mínimos cuadrados ponderados mencionados antes en razón de que no se cuenta con un paquete computacional que permita este procedimiento. Actualmente se diseña un programa computacional para sortear esta dificultad.

COMENTARIOS FINALES.

La metodología de series de tiempo no se usó porque las observaciones no se hicieron diarias en todos los casos. Aún el caracter del modelaje y los resultados se consideran preliminares, actualmente se diseña un estudio observacional con propositos de verificar la bondadad de los modelos propuestos y avanzar más en los aspectos metodológicos de la predicción. Se han obtenido evaluaciones preliminares de los modelos encontrando consistencia con los resultados de otras investigaciones [Contreras (1988)] y se han planteado fases avanzadas de modelaje, las cuales consideran otras variables meteorológicas y descartan las que aquí no mostraron relevancia.

Considerando algunas observaciones respecto de la metodología aquí presentada, sobre una versión preliminar [Mendoza (1988)], se ha retomado la realización del diagnostico sobre modelos alternativos considerando variables circulares.

Se ha realizado recientemente una valoración completa de los avances, tanto desde el punto de vista meteorológico como estadístico [Tejeda y Ojeda (1988)] y se han planteado nuevas metas en el proyecto de investigación que costituye el contexto del presente artículo.

AGRADECIMIENTO: Agradezco la colaboración valiosa y los comentarios del Dr. Carlos Mendoza.

BIBLIOGRAFIA.

Anscombe F. J. and Tukey J. W. (1963) **The examination and analysis of residuals**; Technometrics Vol. 5 p.p. 141-160.

Atkinson A. C. (1985) **Plots, transformations and regression**; Clarendon Press.

Beckman R. J. and Cook R. D. (1983) **Outliers**; Technometrics Vol. 25 p.p. 119-149.

Besley D. A., Kuh E. and Welsh R. E.(1980) **Regression diagnostics:identifying influential data and sources of collinearity**; Wiley.

Cervantes P. J. (1987) **Una relación entre la dirección del viento a 850 mlb. y la dirección del viento en la zona costera central de Veracruz**; Tesis Facultad de Física U.V.

Contreras A. (1988) **Estudio de la relación estadística del viento en superficie y el viento a 500 mlb.**; Tesis en preparación.

Gunst R. F. and Mason R. L. (1980) **Regression analysis and its applications: a data oriented approach**. Marcel Dekker.

Hocking R. R. (1983) **Developments in linear regresion methodology: 1959-1982**; Technometrics, Vol. 25 p.p. 219-230.

Johnson M. E. and Kalma J. D. (1984) **A study of the dependence of surface wind direction on the gradient wind**; Jour. Met. vol 114 p. p. 257-269.

Mardia K. V. (1972) **Statistics of Directinal data** Academic Press.

Mendoza C. (1988) **Comentarios sobre una versión preliminar de este artículo.** disponible con el autor.

Mosteller F. and Tukey J. W. (1977) **Data analysis and regression**; Addison-Wesley.

Oseguera I. (1988) **Una aplicación de análisis de correlación en variables circulares**; Reporte técnico LINAE.

Seber G. A. F. (1984) **Multivariate observations**; Wiley.

Tejeda A. y Contreras J. A. (1988) **Modelos de regresion para predecir direccion de viento en superficie**; enviado a la revista Heuristica. Dep. de Ingeniería U. del Valla Cali, Colombia.

Tejeda A. y ojeda M. M. (1988) **Evaluación de modelaje de la relacion viento en superficie-viento en altura.** versión disponible con el autor (enviada para posible publicacion).

Weisberg S (1980) **Applied linear regression**; Wiley.

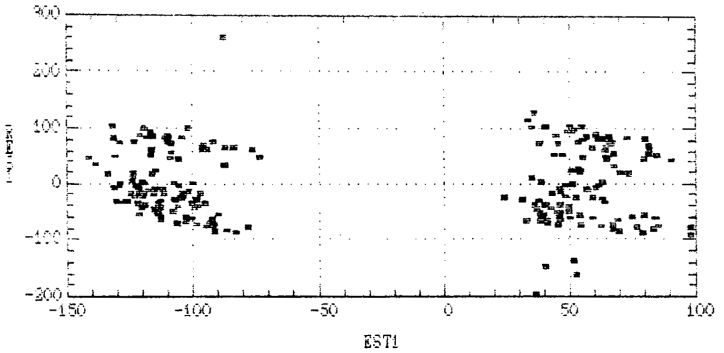


Figura 1. Gráfica de residuales contra valores ajustados al predecir la dirección usando mínimos cuadrados ordinarios.

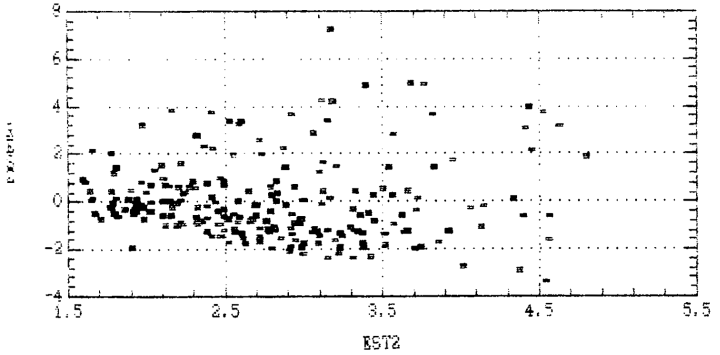


Figura 2. Gráfica de residuales contra valores ajustados al predecir la velocidad usando mínimos cuadrados ordinarios.

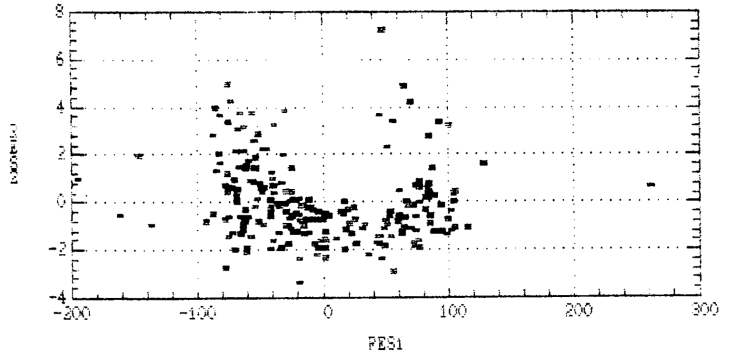


Figura 3. Gráfica de residuales de una variable contra residuales de la otra.

COMPORTAMIENTO DE METODOS ROBUSTOS EN REGRESION DENTRO DE LA FAMILIA DE DISTRIBUCIONES ESTABLES

Jorge Domínguez D.
y
Víctor Pérez-Abreu

Centro de Investigación en Matemáticas, A. C.
A. P. 402, 36000 - Guanajuato, Gto.

Resumen

En este trabajo se reportan los resultados de un estudio de simulación para investigar el comportamiento de algunos métodos de estimación robusta en regresión lineal cuando los errores pertenecen a la familia de distribuciones estables simétricas.

1. INTRODUCCION

En los últimos veinte años las distribuciones estables han sido frecuentemente usadas en Economía, entre otras cosas, como modelos estocásticos del comportamiento del precio de acciones y análisis económicos (ver por ejemplo Akgiray y Booth (1988)). El interés en estas distribuciones se debe principalmente a las siguientes propiedades: (a) sólo las leyes estables tienen dominio de atracción y por lo tanto generalizan el teorema central del límite; (b) las distribuciones estables pertenecen a su propio dominio de atracción, lo que representa una condición de estabilidad, y (c) excepto la distribución normal, todas las otras distribuciones estables tienen colas pesadas, es decir, observaciones grandes ocurren con probabilidad alta.

Una variable aleatoria es estable si el logaritmo de su función característica es de la forma

$$log\ \phi(t) = i\mu t - c|t|^{\alpha}\{1+\beta(t/|t|)\omega(|t|,\alpha)\}$$

en donde $-\infty < \mu < \infty$, $c > 0$, $|\beta| \leq 1$, $0 < \alpha \leq 2$ y

$$\omega(|t|,\alpha) = \begin{cases} tan(\pi\alpha/2) & \text{si } \alpha \neq 1 \\ (2/\pi)log|t| & \text{si } \alpha = 1. \end{cases}$$

Los parámetros μ y c son medidas de localización y escala respectivamente y β es una medida de asimetría. Si $\beta < 0$ la distribución es sesgada a la derecha, para $\beta > 0$ es sesgada a la izquierda y $\beta = 0$ indica que la distribución es simétrica alrededor de μ. El parámetro α se llama índice de estabilidad y está asociado a la probabilidad de las colas. Para α pequeño las colas tienden a ser más pesadas y si α es cercano a 2 éstas son menos pesadas. Si $0 < \alpha < 2$ los momentos de orden $k \geq \alpha$ de la distribución no existen (en particular la varianza no existe), y sólo para $\alpha = 2$, que corresponde al caso de la distribución normal, existen todos los momentos. Una descripción detallada de las distribuciones estables y sus propiedades se puede encontrar en Feller (1971), Fama y Roll (1971), Du Mouchel (1971) y Gómez Arias (1988).

La interpretación anterior de los parámetros de una distribución estable muestra que la familia de leyes estables ofrece una amplia gama de modelos probabilísticos completamente caracterizados por localización, escala, sesgo y comportamiento de las colas. Este hecho y las propiedades mencionadas al principio de este trabajo sugieren que las distribuciones estables sean usadas en estudios de robusticidad como modelos alternativos a la violación de la suposición de normalidad, principalmente en lo relativo a asimetría y colas pesadas. En esta dirección Gómez Arias y Pérez-Abreu (1987) estudian la robusticidad de la estadística t-student dentro de la familia de distribuciones estables.

El propósito del presente trabajo es investigar el comportamiento de algunos métodos robustos de estimación en regresión lineal cuando se supone que el error pertenece a la familia de distribuciones estables simétricas ($\beta = 0$). Heiler (1981) reporta un estudio de simulación similar en el que no se consideran distribuciones estables excepto la distribución normal y la Cauchy. Además hemos incluido un M-estimador

adicional (ψ = *tanh*) el cual da estimadores especialmente bien comportados cuando se suponen errores estables simétricos con colas muy pesadas. Recientemente Caroll y Welsh (1988) han estudiado regresión robusta con errores asimétricos de varianza finita. En un trabajo posterior reportaremos sobre el comportamiento de M-estimadores en regresión lineal cuando se consideran errores estables asimétricos, así como el caso de regresión no lineal con errores estables diversos.

En la Sección 2 de este trabajo recordamos brevemente a los M-estimadores en regresión y mencionamos los correspondientes que incluimos en el estudio . La Sección 3 presenta los diseños usados y sus principales características, así como los criterios empleados para comparar a los estimadores. En la Sección 4 presentamos las tablas con los resultados de la simulaciones realizadas y finalmente en la Sección 5 aparecen las conclusiones del estudio.

2. M-ESTIMADORES EN REGRESION

En esta sección mencionamos de manera breve el método de M-estimadores en regresión así como los estimadores empleados en el estudio. Para una mayor referencia del tema el lector es remitido al trabajo de Gracia-Medrano (1984). Consideremos el modelo lineal

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \tag{2.1}$$

donde $X_{(n\times p)}$ es la matriz diseño, $\beta_{(p\times 1)}$ es el vector de parámetros y $\underline{\varepsilon}$ es el vector de errores. En analogía con el método de mínimos cuadrados se define el M-estimador de $\underline{\beta}$ asociado a la función ψ como el valor $\underline{\beta}$ que resuelve el sistema de ecuaciones

$$\sum_{i=1}^{n} \psi((Y_i - X_i\underline{\beta})/\delta)X_{ij} = 0 \quad j = 1,2,\ldots,p \tag{2.2}$$

en donde δ es un estimador de escala.

El estimador de mínimos cuadrados corresponde al caso $\psi(x) = x$. Un buen procedimiento robusto debe estar asociado a una función ψ que produzca estimadores que no sean fuertemente influenciados por observaciones discrepantes.

Una forma de resolver la ecuación (2.2) es mediante un método iterativo que usa mínimos cuadrados ponderados. Así, si $\underline{\beta}_1$ es una estimación inicial de $\underline{\beta}$ (el estimador de mínimos cuadrados es usado en este trabajo)

$$\hat{\underline{\beta}}_2 = (X'WX)^{-1}X'WY \tag{2.3}$$

en donde $W = diag(W_1,\ldots,W_n)$ y

$$W_i = \frac{\psi((Y_i - X_i\hat{\beta}_1)/\Delta)}{(Y_i - X_i\hat{\beta}_1)/\Delta} \qquad i = 1,\ldots,n \tag{2.4}$$

y el procedimiento continúa en forma iterativa. Los estimadores de escala que han sido propuestos en la literatura son

$$\Delta = \text{mediana } |u_i - med(u_i)|/.6745 \tag{2.5}$$

y

$$\Delta = \frac{1}{2}(q(.75)-q(.25)) \tag{2.6}$$

en donde q(.75) y q(.25) son el tercer y primer percentil muestral respectivamente.

Los M-estimadores que se usan en el presente trabajo son los asociados a las siguientes funciones ψ

Andrews (AN) $\psi(x) = \begin{cases} \text{sen}(x/k) & |x| \le k\pi \\ 0 & |x| > k\pi \end{cases}$ $k = 1.5$ (2.7)

Hampel (HA) $\psi(x) = \text{sign}(x) \begin{cases} |x| & |x| < a \\ a & a < |x| < b \\ \frac{c-|x|}{c-b}a & b \le |x| < c \\ 0 & c \le |x| \end{cases}$ (2.8)

con a = 1.5, b = 5.0, c = 8.0

Tukey (TU) $\psi(x) = \begin{cases} x(1-(x/k)^2)^2, & |x| \le k \\ 0 & |x| > k \end{cases}$ $k = 6.0$ (2.9)

Huber (HU) $\psi(x) = \begin{cases} x & |x| \le k \\ k \text{ sign } x, & |x| > k \end{cases}$ con k = 1.0 (2.10)

Tangh (TA) $\psi(x) = \tanh(x)$

Mínimos Cuadrados (MC) $\psi(x) = x$ (2.11)

Los cuatro primeros M-estimadores han sido bastante usados en estimación robusta y en especial en el estudio de Heiler (1981). El quinto ha sido recientemente usado en regresión no lineal robusta por Aguirre Torres et al (1989).

3. LOS DISEÑOS

Los diseños empleados en este trabajo son los propuestos por Heiler (1981) y que a continuación describimos:

Diseño 1:

$$Y_i = 1 + .5x_{2i} + e_i \quad n = 20 \qquad x_{2i} = -.95 + .1(i-1) \tag{3.1}$$

Diseño 2:

$$Y_i = 1 + .5x_{2i} + e_i \quad n = 40 \qquad x_{2i} = -.342 + .00064(i(i-1)) \tag{3.2}$$

Diseño 3:

$$Y_i = 1+.5x_{2i}+.25x_{3i}+e_i \ , \ n = 30 \qquad x_{2i} = (2i-31)/30 \qquad (3.3)$$
$$x_{3i} = .60148+(i-1)(i-3)/225$$

Diseño 4:

$$Y_i = 1+.5x_{2i}+.25x_{3i}+e_i \ , \ n = 30 \qquad x_{2i} = -.34435+.001149i(i-1) \qquad (3.4)$$
$$x_{3i} = (x_{2i}+.34465)^2-.21374$$

El diseño 1 tiene por característica la de ser un modelo lineal cuyos valores de X_{2i} se distribuyen simétricamente alrededor de cero. El diseño 2 es altamente desbalanceado. Las variables X_{2i} y X_{3i} en el diseño 3 son ortogonales, mientras que en el diseño 4 presentan una alta colinealidad.

Generamos los errores estables ε_i usando el método de simulación de variables aleatorias estables propuesto por Chambers, Mallows y Stuck (1976).

Si $\beta = 0$ (caso simétrico), la expresión para la función característica mostrada en la introducción se reduce a:

$$log\phi(t) = i\mu t - c|t|^{\alpha} \ , \qquad (3.5)$$

en donde $-\infty < \mu < +\infty$, $c > 0$ y $0 < \alpha \leq 2$.

En nuestro trabajo α toma valores entre 1 y 2. Esto debido a que se quiere estudiar el comportamiento de los estimadores, por un lado con valores de α cercanos a 2, esto es próximos a una distribución normal (colas no pesadas), y por otro nos vamos alejando de 2 hasta aproximarnos a $\alpha = 1$, el caso de la distribución de Cauchy (colas demasiado pesadas). En todos los casos $\mu = 0$ y $c = 1$.

El criterio empleado para evaluar los estimadores es la desviación cuadrática media de la linea de regresión o el plano de regresión según corresponda, este se expresa por:

$$DCM = \frac{1}{H} \sum_{j=1}^{H} \left[\sum_{i}^{n} (\hat{Y}_{ij}-Y_i^0)^2 \right] \qquad (3.6)$$

donde Y_i^0 es alguno de los cuatro diseños sin el término ε_i y $\hat{Y}_{ij}$ es el estimador correspondiente de Y_i^0 en la j-ésima repetición, n representa el tamaño de la muestra para cada uno de los diseños, y H es el número de simulaciones que se llevan a cabo, que en nuestro caso es igual a 250.

Usando esta información se construye un criterio para evaluar la eficiencia de los estimadores

$$DEF = 1 - \frac{DCM \text{ (el mejor estimador)}}{DCM} \ . \qquad (3.7)$$

4. RESULTADOS

	Diseño 1		Diseño 2		Diseño 3		Diseño 4	
	DEF	DCM	DEF	DCM	DEF	DCM	DEF	DCM
α=2.0								
MC	.316	12.371	.489	20.4	.311	18.157	.297	18.07
AN	0.0680	9.078	.145	12.189	.061	13.313	0.061	13.54
HA	0.0682	9.076	.138	12.087	.048	13.146	0.031	13.12
TU	0.	8.457	0	10.423	0	12.506	0	12.71
HU	0.136	9.789	.101	11.60	.127	14.32	0.040	13.24
TH	0.17861	10.296	.312	15.158	.169	15.05	0.083	13.87
α=1.98								
MC	.412	13.756	.556	22.448	.423	20.52	.392	20.70
AN	.066	8.657	.142	11.618	.062	12.66	.059	13.37
HA	.083	8.815	.146	11.681	.069	12.75	.013	12.75
TU	0	8.085	0	9.973	0	11.87	0	12.586
HU	.164	9.674	.133	11.5	.161	14.154	-	-
TH	.205	10.173	.332	14.925	.201	14.365	-	-
α=1.95								
MC	.526	16.05	.637	25.740	.230	24.358	.520	24.908
AN	.067	5.456	.142	10.873	.060	11.965	0	11.96
HA	.097	8.444	.163	11.138	.075	12.158	.006	12.038
TU	0	7.612	0	9.326	0	11.243	.020	12.2
HU	.177	9.48	.165	11.172	.190	13.872	.016	12.16
TH	.241	10.04	.362	14.609	.232	14.645	.046	12.50
α=1.90								
MC	.658	20.486	.740	32.019	.657	31.76		
AN	.071	7.541	.150	9.802	.0204	11.107		
HA	.118	7.939	.191	10.301	.053	11.487		
TU	0	7.002	0	8.335	0	10.88		
HU	.234	9.14	.219	10.67	.188	13.396		
TH	.285	9.795	.409	14.11	.240	14.319		

	Diseño 1		Diseño 2		Diseño 3		Diseño 4	
	DEF	DCM	DEF	DCM	DEF	DCM	DEF	DCM
α=1.80								
MC	.819	32.823	.863	44.289	.816	52.61	.830	56.115
AN	.088	6.524	.150	7.93	0.0	9.65	.068	10.213
HA	.165	7.125	.244	8.914	0.069	10.369	.069	10.219
TU	0	5.4949	0	6.74	0.014	9.79	0	9.517
HU	.237	7.80	.296	9.58	.226	12.46	.170	11.46
TA	.369	9.42	.486	13.11	.297	13.72	.258	12.83
α=1.7								
MC	.894	53.29	.9285	77.92	.892	87.89	.904	95.665
AN	0	5.656	.147	6.53	-	-	0.055	9.769
HA	.127	6.479	.273	7.66	0	9.45	.007	9.296
TU	.130	6.503	0	5.57	-	-	0	9.230
HU	.284	7.90	.346	8.52	.187	11.62		-
TA	.369	8.96	.539	12.07	.281	13.136		-
α=1.6								
MC	.934	90.10	.963	129.71	.938	152.61	.949	168.97
AN	.151	7.012	.130	5.469	-	-	.077	9.42
HA	0	5.952	.273	5.55	0	9.45	0	8.69
TU	.309	8.614	0	4.76	-	-	.037	9.02
HU	.139	6.92	.378	7.66	.126	10.84	-	-
TH	.282	8.29	.566	10.96	.281	13.136	-	-
α=1.5								
MC	.962	161.96	.981	237.75	.975	281.23	.974	315.9
AN	0.449	10.29	.079	4.8	.048	7.52	.129	9.48
HA	0	5.668	.843	28.07		-	0	8.26
TU	.457	10.538	0	4.42	0	7.16	.120	9.39
HU	.283	7.91	.427	7.72	.392	11.78	.210	10.46
TH	.310	8.209	.555	9.931	.410	12.130	.252	11.05

	Diseño 1		Diseño 2		Diseño 3		Diseño 4	
	DEF	DCM	DEF	DCM	DEF	DCM	DEF	DCM
α=1.1.								
MC	.998	5232	.999	7681	.998	9200		1061
AN	.905	74.9	.978	372.2	.999	411296		3713
HA	0	7.12	.530	16.889	.103	14.66		249.5
TU	.907	76.7	.984	491.25	.99	411276		36235.1
HU	.999	142854	.999	419845	.999	53836		
TH	.196	8.86	0	7.932	0	13.449		

5. CONCLUSIONES

El método de mínimos cuadrados se ve muy afectado cuando nos alejamos de la distribución normal, aún el caso α = 1.98 que está muy cercano a la normal.

El estimador robusto obtenido usando la función de Tukey resulta ser bueno para valores de $1.8 \leq \alpha \leq 2.0$ y esencialmente en todos los modelos; además, para $\alpha \leq 1.7$ tiene un comportamiento adecuado en el diseño 2, salvo en α = 1.1 .

A medida que nos alejamos de la normal $\alpha \leq 1.7$, Hampel resulta ser el mejor, exceptuando el diseño 2.

El estimador robusto cuya función ψ es la tangente hiperbólica observó un comportamiento homogéneo en todos los casos y se comporta extremadamente bien cuando α es cercano a uno.

Algunas consideraciones de nuestros resultados con los obtenidos por Heiler(1981) son los siguientes:

Heiler considera el caso en que los errores ε_i se distribuyen como una normal y la dispersión (varianza) es muy pequeña (bondadosa). En tal situación el estimador de MC resulta ser el mejor, como era de esperarse. En el presente trabajo no consideramos este aspecto, sin embargo nuestros resultados son muy similiares a los de Heiler cuando se considera una distribución normal con varianza grande. En nuestro estudio resalta TU.

Para la distribución de Cauchy, HU tuvo un comportamiento regular en el estudio de Heiler, mientras que en el nuestro se vió muy afectado cuando se está cercano a la Cauchy. Aunque no tenemos referencia de Heiler para HA, esta función fue la mejor en nuestro estudio para el diseño 1 y regular para los otros. Sin embargo cabe resaltar que la tangente hiperbólica es el mejor estimador cuando los ε_i siguen una distribución cercana a la Cauchy, caso no considerado por Heiler.

Por el momento no podemos comparar nuestros resultados con los de Heiler de manera más fidedigna porque nos falta considerar otros estimadores de escala, ver expresión (2.2). Debemos tener en cuenta que en este trabajo aún no estudiamos el caso no simétrico, lo cual forma parte de una segunda etapa.

REFERENCIAS

Aguirre Torres, V., Gallant, R. y Domínguez, J. (1989). On Choosing Between Two Nonlinear Models Estimated Robustly, Comm. in Statist. Sim & Com. Vol. B 18 No. 1.

Akgiray V. y Booth, G. (1988). The Stable-Law of Stock Returns. J. of Business & Econ. Statistics, Vol. 6, No. 1. 51-57.

Carroll, A. y Welsh, A. (1988). A Note on Asymmetry and Robustness in Linear Regression. The American Statistician, Vol. 42, No. 4 Pag. 285-287

Chambers, J. M., Mallows, C. L. y Stuck, B. W. (1976) A. Method for Simulating Stable Random Variables. Journal of the American Statistical Association, Vol. 71, Pag. 340-344.

Du Mouchel, W. H. (1971). On the Asymptotic Normality of Maximum Likelihood Estimates when Sampling from a Stable Distribution. Am. Statist. 1, 948-957.

Fama, E. y Roll, R. (1971). Some Properties of Symmetric Stable Distributions, J. Amer. Statist. Assoc. 63, 817-836.

Feller, W. (1971). An introduction to Probability Theory and its Applications, Vol. II, Second Edition., Wiley, New York.

Gracia-Medrano, L. (1984). Aplicaciones de Técnicas de Regresión Robusta. Tesis, UNAM.

Gómez-Arias, J. (1988). Tópicos de Inferencia Estadística para Leyes Estables. Tesis, UNAM.

Gómez-Arias, J. y Pérez-Abreu, V. (1987). Construcción de Intervalos de Confianza para el Parámetro de Localización de una Distribución Estable Simétrica. Aportaciones Matemáticas, Comunicaciones 4, Pag. 99-117.

Heiler, S. (1981). Robust Estimates in Linear Regression. A simulation Approach. En Computational Statistics, Ed. Buning, H. y Naeva, P. W., de Gruyter, Berlin, 115-136.

METODOLOGIA DEL ANALISIS DE VARIANZA (UNO Y DOS CRITERIOS DE CLASIFICACION) CONSIDERANDO MUESTRAS CENSURADAS DEL TIPO II.

María Cristina Ortiz León
Facultad de Estadística (LINAE)
Av. Xalapa esq. A. Camacho
Xalapa, Ver., México.

RESUMEN:

En este trabajo, se presenta de manera general el censuramiento de tipo II, una prueba para decidir si la muestra debe ser censurada o no en cada celda de un ANVA con uno y dos criterios de clasificación. Se presenta la metodología basada en Estimadores de Máxima Verosimilitud Modificados para muestras censuradas del tipo II, en los casos del ANVA mencionados, asumiendo modelos de efectos fijos.

Este trabajo es de caracter divulgatorio y presenta, por tanto, ejemplos ilustrativos.

INTRODUCCION:

El Análisis de Varianza es una de las técnicas que con mayor frecuencia se utiliza en la investigación experimental, ya que determina, el grado de variabilidad atribuida a un conjunto de factores, llamados fuentes de variación. Esta técnica de Inferencia Estadística en el enfoque clásico esta basada en alguna distribución paramétrica y considera además un conjunto de supuestos tales como: Correcta relación funcional, Aditividad, Homogeneidad de Varianza, Normalidad e Independencia.

El incumplimiento en los supuestos se puede deber a varias razones, siendo muy frecuentes las fallas por la presencia de observaciones extremas o aberrantes (outliers). Existen diversos enfoques para resolver este problema. Uno de estos considera el censurar o quitar aquellas observaciones consideradas aberrantes o extremas y después aplicar una técnica estadistica modificada que considere el efecto del censuramiento.

I.-CENSURAMIENTO.

Se habla de censuramiento cuando se eliminan observaciones extremas en una muestra. Se considera una observación como censurada cuando se dice que solo contiene la información parcial acerca de la variable aleatoria de interés.

I.1 CENSURAMIENTO DEL TIPO II.

Sean $X_1, X_2, \ldots, X_n$ una muestra de tamaño n de una cierta población.

Arreglamos estas n observaciones en orden ascendente de acuerdo a la magnitud y censuramos la r_1 más pequeña y la r_2 más grandes de las observaciones.

Las observaciones restantes:

$$X_{(r_1+1)}, X_{(r_1+2)}, \ldots, X_{(n-r_2)}$$

constituyen una muestra censurada de tipo II de tamaño $n-r_1-r_2$. Nótese, que estas muestras censuradas se presentan naturalmente en ciertas situaciones experimentales; lo cual ilustramos con el siguiente ejemplo típico:

Los siguientes datos muestran los días en los cuales los 7 primeros de una muestra de 10 ratones mueren después de ser inoculados con un cultivo uniforme de tuberculosis humana (Tiku y otros 1986, pág. 22): 41 44 46 54 55 58 60.

Aquí n=10, r_1=0 y r_2=3, y las observaciones restantes constituyen una muestra censurada del tipo II de tamaño 7. El censuramiento ocurre en este caso solo del lado derecho.

II.- PRUEBA PARA DETECTAR SI UNA MUESTRA DEBE SER CENSURADA O NO

Para poder decidir, si una muestra debe ser censurada o no en cada celda, se presenta la siguiente prueba:

Prueba de $T(r_1, r_2)$

En primer lugar se conjetura a cerca de r_1 y r_2 en base al análisis exploratorio de nuestros datos.

Las hipótesis planteadas desde un principio son:

Ho: La muestra contiene outliers.
vs
Ha: La muestra no contiene outliers.

El estadístico calculado es:

$$T(r_1, r_2) = \frac{\left(1 - \frac{1}{n}\right)\hat{\sigma}_c}{\left(1 - \frac{1}{A}\right)\hat{\sigma}}$$

donde:

$$A = n - r_1 - r_2 \qquad m = n - r_1 - r_2 + r_1\beta_1 + r_2\beta_2$$

$$q_1 = \frac{r_1}{n} \quad , \quad q_2 = \frac{r_2}{n}$$

$$\hat{\sigma}_c = \frac{B + \sqrt{B^2 + 4AC}}{2A}$$

$$B = r_2\alpha_2(X_{(n-r_2)} - K) - r_1\alpha_1(X_{(r_1+1)} - K)$$

$$K = \frac{\sum_{i=r_1+1}^{n-r_2} X_{(i)} + r_1\beta_1 X_{(r_1+1)} + r_2\beta_2 X_{(n-r_2)}}{m}$$

$$C = \sum_{i=r+1}^{n-r_2} X^2_{(i)} + r_1\beta_1 X^2_{(r_1+1)} + r_2\beta_2 X^2_{(n-r_2)} - mK^2$$

Los valores de (α_1,β_1) y (α_2,β_2), están dados en la tabla de Tiku y otros (1986) para n=10 y 20.

$\hat{\sigma}$ es la varianza de la muestra sin censurar.

Si $r_1 = r_2$, entonces se habla de censuramiento simétrico y nuestras formulas se reducen a:

$$\alpha_1 = \alpha_2 = \alpha \ , \ \beta_1 = \beta_2 = \beta \text{ y } D=0,$$

$$m = n - 2r + 2r\beta$$

$$K = \frac{\sum_{i=r+1}^{n-r} X_{(i)} + r\beta(X_{(r+1)} + X_{(n-r)})}{m}$$

$$A = n - 2r$$

$$B = r\alpha(X_{(n-r_2)} - X_{(r_1+1)})$$

$$C = \sum_{i=r+1}^{n-r} X^2_{(i)} + r\beta(X^2_{(r+1)} + X^2_{(n-r)}) - m\hat{\mu}^2$$

$\hat{\sigma}_c$ se calcula de la misma manera.

Este estadístico se compara con uno de tablas calculado de la siguiente manera:

$$E_c = \frac{n-1}{n-r_1-r_2-1}\, u_\alpha^* + \frac{1}{5n}\left(1 + \frac{1}{n-2r_2+1}\right)$$

donde u_α^* es el 100_α percentil de la distribución beta con $(n-r_1-r_2-1,\ r_1+r_2)$ grados de libertad.

Regla de decisión:

$$\text{Si } T(r_1,r_2) < E_c \text{ se acepta Ho.}$$

Ejemplo:

Consideremos las siguientes 10 observaciones (Tiku y otros 1986 pag. 274) 568 570 570 570 572 572 572 578 592 596 y queremos saber si realmente la muestra debe ser censurada o no.

Primero planteamos la siguiente hipótesis:

Ho. La muestra debe ser censurada.

vs.

Ha. La muestra no debe ser censurada.

De la figura 1 es claro que si censuramos tendría que ser por el lado derecho y serían dos observaciones. Así que: $r_1=0$, $r_2=2$ de la tabla 1 tenemos que $\alpha_1=0$ y $\beta_1=1$; $\alpha_2=0.7493$ $\beta_2=0.7914$, $A= 10 - 0 - 2 = 8$

$$\hat{\sigma}_c = \frac{8.131988 + \sqrt{(8.131988)^2 + 4(8)(117.9)}}{2(8)} = 4.3606$$

$$\hat{\sigma} = 8.2559$$

$$T(0,2) = \frac{(1 - \frac{1}{10})(4.3806)}{(1 - \frac{1}{8})(8.2559)} = 0.548$$

$$E = \frac{2}{7}(0.594) + \frac{1}{50}\left(1 + \frac{1}{10 - 2 + 1}\right) = 0.786$$

Como $0.548 < 0.786$ aceptamos Ho. y la muestra efectivamente debe ser censurada.

III.-ESTIMACION DE PARAMETROS EN MUESTRAS CENSURADAS DEL TIPO II:

Para la estimación en muestras censuradas del tipo II, se pueden utilizar dos métodos: el método de Mínimos Cuadrados y el de Máxima Verosimilitud. Usaremos el segundo siguiendo las recomendaciones de Tiku y otros (1986) pag. 33.

Sea la función de verosimilitud L basada en la muestra censurada del tipo II de $X_{(i)}$ $(i = r_1 + 1, r_1 + 2, \ldots, n - r_2)$ está dada por:

$$L = \frac{n!}{r_1!\, r_2!}\, \sigma^{-(n-r_1-r_2)} \left[\prod_{i=r_1+1}^{n-r_2} f(\xi_{(i)})\right] \left[F(\xi_{(r_1+1)})\right] \left[1 - F(\xi_{(n-r_2)})\right]^{r_2}$$

(1)

Los estimadores de Máxima Verosimilitud son las soluciones de las ecuaciones.

$$\frac{\partial \log L}{\partial \mu} = 0 \qquad\qquad \frac{\partial \log L}{\partial \sigma} = 0$$

Para varios tipos de distribuciones las ecuaciones (2) no admiten soluciones explícitas. Debido a esto, se requieren métodos iteractivos adecuados; por lo tanto números autores han estudiado el tema entre los que podríamos mencionar a Cohen (1957) (1961), Harter y Moore (1966). Los métodos que propusieron trabajan bien y las iteracciones convergen bastante rápido a las soluciones verdaderas. Pero debido a la naturaleza implícita de esas iteracciones, es difícil hacer algún estudio analítico del resultado de Máxima Verosimilitud, especialmente para muestras pequeñas. Por consiguiente se modificara (2) y definiremos estimadores de máxima verosimilitud modificados como funciones explícitas de observaciones muestrales (Tiku, (1967), Tiku y Stewart, (1977)). La siguiente metodología estudia Estimadores de Máxima Verosimilitud modificados, según Tiku y otros (1986).

IV. - ANALISIS DE VARIANZA EN MUESTRAS CENSURADAS PARA MODELOS DE EFECTOS FIJOS CON UN SOLO CRITERIO DE CLASIFICACION:

Cuando se trabaja con muestras censuradas del tipo II y se quiere realizar un análisis de varianza para un sólo criterio de clasificación; es decir, se han censurado las muestras para cada tratamiento, la metodología es la siguiente:

Consideremos el modelo para un diseño completamente al azar con un solo criterio de clasificación y de efectos fijos:

$$Y_{ijk} = \mu + \tau_i + E_{ijk} \qquad \begin{matrix} i=1,\ldots,k \\ j=1,\ldots,n_i - r_{1i} - r_{2i} \end{matrix}$$

donde:

μ media general.
τ_i Efecto del i-ésimo tratamiento.
E_{ij} Error Experimental.

La hipótesis a probar es:

Ho. $\tau_1=\tau_2=\ldots=\tau_k$

vs.

Ha. Al menos un τ_j distinto.

El procedimiento a seguir para obtener la suma de cuadrados consiste en considerar A_i, B_i, C_i, m_i y K_i de A, B, C, m, K (definidos en la prueba de $T(r_1,r_2)$) valores calculados respectivamente del i-ésimo tratamiento.

Los Estimadores de Máxima Verosimilitud Modificados obtenidos para $\hat{\mu}$, $\hat{\tau}_i$, y $\hat{\sigma}$ son:

$$\hat{\mu} = K_{..} \qquad \hat{\tau}_i = K_i - \overline{K}_{..} \qquad y \qquad \hat{\sigma} = \frac{B + \sqrt{B^2 + 4AC}}{2\sqrt{A(A-1)}}$$

donde:

$$K_{..} = \frac{\sum_{i=1}^{k} m_i K_i}{m..}, \qquad m.. = \sum_{i=1}^{k} m_i$$

$$A = \sum_{i=1}^{k} A_i \qquad B = \sum_{i=1}^{k} B_i \qquad C = \sum_{i=1}^{k} C_i$$

La suma de cuadrados de los componentes se considera como sigue:

$$SCTRAT = \sum_{i=1}^{k} m_i (K_i - K..)^2$$

$$SCE = (A - kc)\,\hat{\sigma}^2$$

$$SCTOTAL = SCTRAT + SCE$$

Con lo cual se construye la tabla 1.

Y rechazamos Ho. si $Fc \geq F\alpha_{(k-1,A-k)}$

Ejemplo:

Consideremos los datos de la tabla 2.

De aquí censuramos los 12 tratamientos con $r_1 = r_2 = r = 1$ para todos los tratamientos de q = = 0.25. En este caso tenemos que α =0.7834, β=0.8049, tomando los valores en tablas.

La tabla 3 nos muestra el Análisis de Varianza correspondiente:

ya que $F_{(11,12)0.05\%}$=2.79.

Como $F_c > F_{(11,12)0.05\%}$ rechazamos Ho.

V. -ANALISIS DE VARIANZA CON DOS CRITERIOS DE CLASIFICACION.

Para el caso de Análisis de Varianza en donde se quiere determinar la variabilidad atribuida a 2 factores o fuentes de variación para muestras censuradas del tipo II se tiene el siguiente modelo:

$$Y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijl}$$

$$i=1,\ldots,c$$
$$j=1,\ldots,k$$
$$l=1,\ldots,n_{ij}-r_{1ij}-r_{2ij}$$

donde:

μ	Media general
α_i	Efecto del factor A en el nivel i
β_j	Efecto del factor B en el nivel j
$(\alpha\beta)_{ij}$	Efecto de la iterracción en el nivel (i,j)
E_{ij}	Error experimental.

El juego de hipótesis a probar es:

1) Ho. $\alpha_i = 0 \ \forall\ i$ vs Ha. $\alpha_i \neq 0$ para algún i
2) Ho. $\beta_j = 0 \ \forall\ j$ vs Ha. $\beta_j \neq 0$ para algún j
3) Ho. $\alpha\beta_{ij}=0 \ \forall\ ij$ vs Ha $\alpha\beta_{ij}\neq 0$ para algún par (i,j)

Se calculan los A_{ij} , B_{ij} , C_{ij}, m_{ij} y K_{ij} correspondientes a los A, B, C, m, K definidos anteriormente calculados en la i,j-ésima celda.

Al igual que en la prueba anterior los Estimadores de Máxima Verosimilitud Modificados obtenidos para $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, $(\hat{\alpha},\hat{\beta})_{ij}$ y $\hat{\sigma}$ son:

$$\hat{\mu} = \bar{K}_{..} \qquad \hat{\alpha}_i = \bar{K}_{i.} - \bar{K}_{..} \qquad \hat{\beta}_j = \bar{K}_{.j} - \bar{K}_{..}$$

$$(\hat{\alpha\beta})_{ij} = K_{ij} - \bar{K}_{i.} - \bar{K}_{.j} + \bar{K}_{..}$$

$$\hat{\sigma} = \frac{B + \sqrt{B^2 + 4AC}}{2\sqrt{AC(A-1)}}$$

donde:

$$\bar{K}_{i.} = \frac{\sum_{t=1}^{c} \sum_{t=1}^{k} m_{ij} K_{ij}}{m_{..}}$$

$$\bar{K}_{i.} = \frac{\sum_{t=1}^{c} m_{ij} K_{ij}}{m_{.j}} = \frac{\sum_{j=1}^{k} m_{ij} K_{ij}}{m_{i.}}$$

$$m_{..} = \sum_{t=1}^{k} \sum_{j=1}^{c} m_{ij} \qquad m_{.i} = \sum_{j=1}^{c} m_{ij} \qquad m_{j.} = \sum_{t=1}^{k} m_{ij}$$

$$A = \sum_{t=1}^{k} \sum_{j=1}^{c} A_{ij} \qquad B = \sum_{t=1}^{k} \sum_{j=1}^{c} B_{ij} \qquad C = \sum_{t=1}^{k} \sum_{j=1}^{c} C_{ij}$$

La suma de cuadrados de cada uno de los componentes se considera como sigue:

$$SCA = \sum_{t=1}^{k} m_{i.} (\bar{K}_{i.} - \bar{K}_{..})^2 \qquad SCB = \sum_{j=1}^{c} m_{.j} (\bar{K}_{.j} - \bar{K}_{..})$$

$$SCAB = \sum_{t=1}^{k} \sum_{j=1}^{c} m_{ij} (K_{ij} - \bar{K}_{i.} - \bar{K}_{.j} + \bar{K}_{..})^2$$

$$SCE = (A - kc)\,\hat{\sigma}^2 \qquad SCTOTAL = SCA + SCB + SCAB + SCE$$

Con lo cual se construye tabla 4.

Las reglas de decisión en este caso son:

1) Rechazamos Ho. si $Fc_A \geq F\alpha(k-1, A-kc)$.

2) Rechazamos Ho. si $Fc_B \geq F\alpha(c-1, A-kc)$

3) Rechazamos Ho. si $Fc_{AB} \geq F\alpha((k-1)(c-1), A-kc)$

Ejemplo:

Consideremos los datos de la tabla 5.

En este caso se censuran las 12 celdas con $r_1=r_2=r=1$ para todos los tratamientos de $q_i=\frac{ri}{ni}=0.25$. Aquí tenemos que $\alpha_i=0.7834$, $\beta_i=0.8049$, tomando los valores en tablas.

La tabla 6 nos muestra el Análisis de Varianza correspondiente:

Reglas de decisión:

1) $F_A > F_{0.05,(2,12)}$ Rechazamos Ho.$_1$
2) $F_B > F_{0.05,(3,12)}$ Rechazamos Ho.$_2$
3) $F_{AB} > F_{0.05,(6,12)}$ Rechazamos Ho.$_3$

BIBLIOGRAFIA:

Box, G.E.P. and Cox D. R. (1964). An analysis of transformations (with discussion). J.R. Statist. Soc. B26, 211-252.

Cohen, A.C. (1957). On the solution of estimating equations from truncated and censored samples from normal populations. Biometrika 44, 225-236.

Cohen, A.C. (1961). Tables for maximun likelihood estimates; singly censored samples. Rechbometrics 3, 535-541.

Harter, H. L. and Moore, A. H. (1966). Locatel maximum likelihood estimation of the parameters of three parameter log-normal populations from complete and censored samples. J. Amer. Statist. Ass. 61,842-851.

Miller, R. G., Gong G and Muñoz A. (1981). Sirvival Analysis. Ed. John Wiley & Sons.

Ortiz M. C. L. (1988). Algunos Métodos Robustos en el Análisis de Varianza. Tesis de Licenciado en Estadística, U.V.

Tiku, M.L. (1967). Estimating the mean and standard desvition from a censored normal sample. Biometrika 54, 155-165.

Tiku, M.L. and Stewart D. E. (1977). Estimating and testing group effects from Type I censored normal samples in experimental design. Commun. Statist. A6(15), 1485-1501.

Tiku, M. L., Tan, W. Y. and Balakrishnan N. (1986). Robust Inference. Marcel Dekker, Inc.

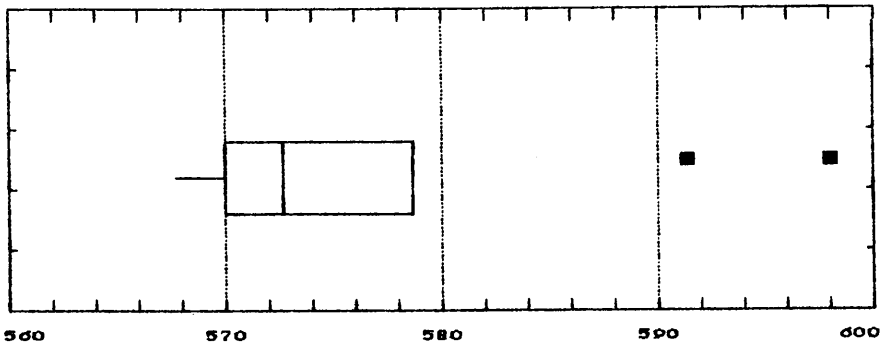


FIGURA 1. Gráfica de cajas y alambres.

Tabla 1. Análisis de Varianza para muestras censuradas del tipo II con un criterio de clasificación.

FUENTE	GL	SC	CM	FC
TRATAMIENTO	k-1	SCTRAT	CMTRAT	$\frac{CMTRAT}{CME}$
ERROR	A-k	SCE	CME	
TOTAL	A-1	SCTOTAL		

Tabla 2. Datos tomados de Ortiz (1988 pag. 94).

		REPLICAS			
		1	2	3	4
T	1	0.31	0.45	0.46	0.43
R	2	0.82	1.10	0.88	0.72
A	3	0.43	0.45	0.63	0.76
T	4	0.45	0.71	0.66	0.62
A	5	0.36	0.29	0.40	0.23
M	6	0.92	0.61	0.49	1.24
I	7	0.44	0.35	0.31	0.40
E	8	0.56	1.02	0.71	0.38
N	9	0.22	0.21	0.18	0.23
T	10	0.30	0.37	0.38	0.29
O	11	0.23	0.25	0.24	0.22
S	12	0.30	0.36	0.31	0.33

Tabla 3. Análisis de Varianza correspondiente a los datos de la tabla 2.

FUENTE	GL	SC	CM	FC
TRATAMIENTO	11	1.749	0.1590	8.125
ERROR	12	0.233482	0.0195683	
TOTAL	23	3.61		

Tabla 4. Análisis de Varianza para muestras censuradas del tipo II con dos criterios de clasificación.

FUENTE	GL	SC	CM	FC
A	k-1	SCA	CMA	$\frac{CMA}{CME}$
B	C-1	SCB	CMB	$\frac{CMB}{CME}$
AB	(k-1)(C-1)	SCAB	CMAB	$\frac{CMA}{CME}$
ERROR	A-kc	SCE	CME	
TOTAL	A-1	SCTOTAL		

Tabla5. Datos tomados de Box y Cox (1964 p 220).

	B			
A	1	2	3	4
UNO	3.23	1.20	3.33	2.22
	2.22	0.91	2.22	1.41
	2.17	1.14	1.59	1.52
	2.33	1.39	1.32	1.61
DOS	2.78	1.09	2.27	1.79
	3.45	1.64	2.86	0.98
	2.50	2.04	3.23	1.41
	4.35	0.81	2.50	2.63
TRES	4.55	3.33	4.35	3.33
	4.76	2.70	4.00	2.78
	5.56	2.63	4.17	3.23
	4.35	3.45	4.55	3.03

Tabla 6. Análisis de Varianza correspondiente a los datos de la tabla 5.

FUENTE	GL	SC	CM	FC
A	2	32.8874	16.4437	70.4287
B	3	16.0771	5.3590	22.9528
AB	6	1.1570	0.1928	0.8258
ERROR	12	2.8018	0.2325	
TOTAL	23	52.9233		

CONVERGENCIA DE SUMAS DE VARIABLES ALEATORIAS: EVIDENCIA EMPIRICA MEDIANTE SIMULACION.

Espinosa, V.M.
Centro de Investigación y Docencia Económicas, A.C.
Km. 16.5 Carretera México-Toluca, Cuajimalpa 01210, D.F.

RESUMEN.

En este material se presenta el apoyo necesario para utilizar, explorar y aprovechar el paquete de cómputo CONVERGE que pretende dar evidencia empírica de la convergencia de sumas estandarizadas a la función de distribución normal estándar y de resultados teóricos relacionados con esta convergencia.

1. INTRODUCCION.

En la siguiente sección se introducen los conceptos de sumas de variables aleatorias, función de distribución normal estándar, función de distribución empírica, convergencia en distribución, la cota de Berry-Esseen, y la prueba de bondad de ajuste de Anderson-Darling.

En la sección número 3, que constituye la parte central de este documento, aparecen algunos ejercicios sugeridos para utilizar el paquete de simulación, resaltando en cada uno de ellos algún resultado específico de la correspondiente parte teórica. El objetivo de esta sección es optimizar el uso de esta herramienta de cómputo, y servir de apoyo al docente para relacionar el contenido de su curso con prácticas que reafirmen los conceptos adquiridos por el estudiante.

2. CONCEPTOS BASICOS MANEJADOS POR EL PAQUETE DE SIMULACION.

El paquete se maneja mediante sencillos menús interactivos que permiten seleccionar la o las distribuciones a simular y los valores de los parámetros involucrados.

Inicialmente aparece un cuadro que identifica la distribución a simular, la gráfica de su función de densidad o de probabilidades, y remarca el hecho de que las variables simuladas son idénticamente distribuidas.

El cuadro central donde se presentan los resultados de la simulación se compone de la graficación de ejes estandarizados, la impresión de títulos que identifican el caso que se está tratando, el dibujo de la distribución normal estándar, la simulación y graficación de la función de distribución empírica, y el cálculo y presentación de la conclusión de la prueba de bondad de ajuste de Anderson-Darling.

Para una mejor comprensión de estos elementos, se presenta a

continuación una descripción de cada uno de los conceptos involucrados.

Distribucion Normal Estandar.

Se dice que una variable aleatoria X se distribuye como una normal estándar si su función de densidad está dada por

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \; e^{-x^2/2} \qquad -\infty < x < \infty$$

mientras que la función de distribución por no tener una expresión elemental, adquiere la forma

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

La función de densidad tiene la particularidad de ser simétrica respecto al eje de las ordenadas, por lo que su media es precisamente igual a 0, mientras que su varianza es igual a 1. Esto se denota escribiendo

$$X \sim N(0,1)$$

Precisamente esta distribución juega un papel preponderante en la Teoría de la Probabilidad, debido en mucho a los resultados que aquí se abordarán.

Sumas de Variables Aleatorias.

Puesto que los resultados que se obtendrán serán referidos a sumas de variables aleatorias, es conveniente definir lo siguiente:

Definición. Sea una sucesión $X_1,..,X_n$ de variables aleatorias; entonces una suma estandarizada de la sucesión está dada por

$$\frac{1}{B_n} \left[\sum_{i=1}^{n} X_i - A_n \right]$$

donde A_n y B_n son constantes que sólo dependen de n. Usualmente, y cuando existen, estas constantes se escogen como la media y la desviación estándar de la suma de las variables. La forma de calcularlas, bajo condiciones de independencia está dada por el siguiente teorema:

Teorema. Sea una sucesión $X_1,..,X_n$ de variables aleatorias independientes. Entonces

$$E \left[\sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E\,(X_i)$$

y

$$\text{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \text{Var}\,(X_i)$$

Funcion de Distribucion Empirica.

En el proceso de simulación la función de distribución empírica juega un papel importante por lo que se anota aquí su definición formal:

Definición. Sea $X_1,..,X_n$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con función de distribución F(X). La función de distribución empírica está dada por

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x)}(X_i)$$

donde $I_{(a,b)}(x)$ es la función indicadora definida por

$$I_{(a,b)}(x) = \begin{cases} 1 & \text{si } x \in (a,b) \\ 0 & \text{si } x \notin (a,b) \end{cases}$$

entonces, la función empírica es monótona no decreciente, de forma escalonada y cumple con que $F_n(-\infty) = 0$ y $F_n(\infty) = 1$. Un resultado importante es el teorema de Glivenko-Cantelli (ver Tucker, 1967) que asegura que $F_n(x)$ converge con probabilidad uno a la función de distribución de las variables $X_1,..,X_n$.

Convergencia en Distribucion.

La aproximación de la distribución normal a la distribución de sumas estandarizas se da en el sentido de la convergencia en distribución, que se define como

Definición. Sea $X_1,..,X_n$ una sucesión de variables aleatorias con funciones de distribución $F_n(x)$ y X otra variable aleatoria cualquiera con función de distribución F(x). Se dice que la sucesión converge en distribución a X si

$$\lim_{n\to\infty} F_n(x) = F(x) \qquad \forall\ x \text{ punto de continuidad de } F(x)$$

Por comodidad, se referirá la convergencia en distribución simplemente como la convergencia de la sucesión a la variable X.

Cota de Berry-Esseen.

Asociada con el concepto de rapidez de convergencia, la cota de Berry-Esseen establece la máxima diferencia que puede presentarse entre la función de distribución empírica y la

distribución normal con parámetros $\mu = 0$ y $\sigma^2 = 1$, en todo el recorrido de la variable aleatoria x; la cota queda en función del tercer momento absoluto de x. A continuación se enuncia este concepto formalmente:

Teorema (Berry, 1941 y Esseen, 1945). Sea $x_1, x_2, x_3, \ldots$ una sucesión de variables aleatorias independientes e idénticamente distribuidas con $E(x_k) = 0$, $Var(x_k) = \sigma^2$ y $E(|x_k|^3) < \infty$. Sea $F_n(x)$ la distribución de $S_n/\sigma\sqrt{n}$, $S_n = \frac{1}{n}\sum x_i$ y $\Phi(x)$ la distribución normal estándar. Entonces

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{C\, E(|x_k|^3)}{\sigma^3\sqrt{n}}$$

donde C es una constante tal que $1/\sqrt{2\pi} \leq C < .8$.

Por lo tanto, para casos específicos de variables aleatorias, a medida que la cota es menor, se dice que su suma estandarizada converge más rápidamente a la distribución normal estándar (para la demostración consultar Shiryayev, 1979).

Prueba de Bondad de Ajuste de Anderson-Darling.

Para evaluar si la hipótesis de convergencia de una suma estandarizada a la distribución normal se cumple, se utiliza la prueba de Anderson-Darling que está dada por (consultar D'Agostino, 1986):

$$A^2 = n \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 / [\{F(x)\}\{1-F(x)\}]\, dF(x)$$

donde $F_n(x)$ es la función de distribución empírica, y $F(x)$ es la distribución bajo la hipótesis.

Dada una muestra aleatoria, $x_1, \ldots, x_n$, una forma alternativa para A^2 que se utiliza para su cálculo es la siguiente:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n} [\,(2i-1)\ln Z_{(i)} + (2n+1-2i)\ln(1-Z_{(i)})\,]$$

donde $Z_i = F(x_i)$ y $Z_{(i)}$ son las estadísticas de orden de las Z_i.

La hipótesis de nulidad se rechaza si $A^2 > A_{(1-\alpha)}$, el α-ésimo percentil superior de la distribución (la cual está tabulada).

3.GUIA PARA EL USO DEL PAQUETE DE SIMULACION.

Existen diferentes aspectos sobre la convergencia de sumas de variables aleatorias que pueden ser estudiados mediante la simulación tales como la rapidez de convergencia, afectación por la variación de parámetros, diferencias causadas por los distintos

tipos de variables aleatorias involucradas, contraejemplos, y otros. Aspectos todos analizados y sustentados teóricamente, pero que pueden ser confirmados heurísticamente con la ventaja de recurrir a elementos visuales y experimentales que coadyuven en la comprensión del concepto.

A continuación se presenta un conjunto de prácticas conducentes a experimentar mediante simulación los tópicos arriba mencionados, abordándolos en una secuencia que coincide con el avance teórico que usualmente se da al tema. Ello permite que este material pueda alternar con sesiones en que se presenten y demuestren formalmente los resultados con las prácticas que reafirmen los conocimientos adquiridos y familiaricen al estudiante con su significado.

Se recomienda seleccionar la práctica que mejor corresponda al avance del curso, de acuerdo al objetivo que se establece, al principio de cada una. Para su realización se proponen una o varias actividades que implican la ejecución de algunas simulaciones, la observación de las condiciones en que se efectúan, recabación de características acerca de las distribuciones simuladas, y la contestación a interrogantes que conducen al estudiante a la confirmación del objetivo planteado. Se incluyen además conclusiones generales que señalan los posibles resultados e interpretaciones que el estudiante debió obtener, así como referencias a los teoremas que apoyan tales conclusiones.

Resultaría provechoso que el usuario del paquete leyera todas las instrucciones de la actividad que realizará antes de comenzar, pues con ello podría planear mejor su trabajo, estaría atento a los detalles que debe observar en cada simulación y podría recabar en el momento oportuno toda la información que se le requiere. Se deja al criterio del profesor o conductor del curso la decisión de dar a conocer simultáneamente las conclusiones a que debe llegar el estudiante, o si se prefiere proporcionárselas una vez concluida la práctica.

Para utilizar el paquete en una microcomputadora personal deberá encenderse la máquina con el sistema operativo MS-DOS instalado . Inmediatamente después de dar la fecha y hora es necesario seguir los siguientes pasos:

1) Teclear

>a: [Enter] (cambia a la unidad de diskette)

2) Instalar el diskette del paquete en la ranura de la unidad de disco flexible.

3) Teclear

>CONVERGE [Enter] (entra al programa)

4) Teclear

OLD TCL.TRC [Enter] (lee el archivo)

5) Teclear

[F9] (corre el programa)

En este momento pueden realizarse ya las prácticas que se deseen mediante la selección apropiada en los menús que aparecen.

→ Práctica Número 1. Teorema de deMoivre-Laplace y Rapidez de Convergencia.

Objetivo:

Experimentar la convergencia a la distribución normal de la suma estandarizada de ensayos Bernoulli. Observar el efecto que causa la variación del parámetro de la distribución.

* Actividad (a):

Simular: "1.Control Completo de Parámetros.", "1.bernoulli"
"Probabilidad de éxito? " .5
"Valor de N (5-1000)? " 50
"Simulaciones (5-1000)? " 50

- ¿Qué tipo de distribución es la Bernoulli? ¿Es continua o discreta? con p=1/2, ¿Es simétrica o no? ¿tiene dominio acotado?
- ¿Podría decirse que guarda algún parecido con la distribución normal?
- ¿Parecería entonces 'lógico' aproximar las probabilidades de la suma de variables aleatorias con distribución Bernoulli y p = 1/2 por medio de la distribución normal?
- Reconociendo la curva suave al ejecutar la simulación propuesta como la función de distribución normal y la función escalonada como la distribución empírica de la suma de las variables Bernoulli, ¿Se puede concluir que por este argumento gráfico que en efecto se puede aproximar la distribución de la suma por medio de la normal?
- Una vez graficada la distribución teórica de la normal y la función de distribución empírica, y antes de que aparezca el resultado de la prueba de Anderson-Darling, pruebe visualmente: ¿Se puede considerar que se aproxima a la distribución normal? ¿Es muy alta la aproximación?, ¿satisfactoria o decididamente deficiente?
- Acerca de las preguntas anteriores, y habiendo anotado ya el resultado de la prueba de bondad de ajuste, ¿Cuál fue la conclusión proporcionada? ¿A que nivel de significancia? ¿Coincide con la apreciación hecha *a priori*? En caso negativo, ¿A que puede atribuirse?

Conclusión:

Desde 1730, Abraham deMoivre (De Moivre, 1738) estableció que cuando el número de sumandos en una suma de variables aleatorias Bernoulli con parámetro p = 1/2 -modelo asociado desde los inicios de la probabilidad con el fenómeno del número de 'águilas' observadas en N lanzamientos de una moneda legal- su función de distribución tendía a la distribución normal. En la práctica significa que es posible calcular de manera bastante aproximada la probabilidad de que el número de éxitos sea igual o inferior a un número dado haciendo uso de la función de

distribución normal. La simulación efectuada debió confirmar de manera empírica tan histórico resultado.

* Actividad (b):
Simular: "1.Automática: Variación de N", "1.bernoulli"

- ¿Cuál es el valor de p?, ¿Cuál es el número de simulaciones?
- ¿Qué valores va tomando N?,
- Conforme aumenta N, ¿Cuántos "escalones" de la función de distribución empírica aparecen?, ¿A que se debe?
- ¿Son equidistantes estos "escalones"?, ¿Por qué?
- Nuevamente, antes de observar el resultado de la prueba de Anderson-Darling, ¿Cuál sería la apreciación visual acerca de la convergencia a la distribución normal?
- Ahora, ¿Coincide la apreciación subjetiva con la prueba estadística? ¿Debió corregirse?
- ¿Cómo se va dando la aproximación a la normal al variar N?

Conclusión:

Generalizando, de acuerdo al Teorema de deMoivre-Laplace (ver Gnedenko, 1976), para una suma estandarizada de variables Bernoulli con parámetro p fijo, al tender el número de sumandos N a infinito, la probabilidad de que la suma sea menor o igual a cierto valor puede evaluarse con la función de distribución normal calculada en ese mismo punto; esto debe reflejarse en las simulaciones observando que la distribución empírica y la distribución normal acumulativa estándar tienden a aproximarse a medida que crece N, ¿Ha ocurrido eso?

* Actividad (c):
Simular: "3.Automática: Varían Parámetros", "1.bernoulli"

- ¿Cuál es el valor de N?, ¿Cuántas simulaciones se hacen para construir la distribución empírica?
- ¿Qué valores va tomando el parámetro p?, ¿Hacia que valor tiende?
- A medida que cambia el valor de p, ¿Varía la conclusión acerca de la convergencia a la distribución normal? Reflexionando un poco, ¿Qué fenómeno podría ejemplificar este caso en que el parámetro p tiende a cero? ¿Podría llamarse de 'los eventos raros ?
- Recuerde: trate de establecer su propia conclusión antes de conocer el resultado de la prueba de bondad de ajuste. Si en algún caso no se ha dado la aproximación a la normal, ¿Cómo puede explicarse la diferencia? ¿Está presente una discrepancia en los extremos? ¿Puede considerarse determinante una diferencia en la parte central?
- ¿Qué comportamiento general podría deducirse?

Conclusión:

Si bien el teorema de deMoivre-Laplace asegura la convergencia a la distribución normal de sumas estandarizadas de

variables aleatorias Bernoulli con p fijo, la manera en que se da la aproximación no es uniforme, como lo asegura el Teorema de Berry-Esseen, por lo que con una N en particular, esta aproximación puede ser satisfactoria para algunos valores de p y no serlo para otros. Como comportamiento general, para valores de p cercanos a cero (o a uno, por simetría) la convergencia a la distribución normal es cada vez más lenta. Por ello, en el caso de que la máxima discrepancia esté dentro de límites aceptables signifique un tamaño muy grande de N, se cuenta como alternativa con la aproximación de la distribución Poisson a la binomial que pide que $np \longrightarrow \lambda$ si n tiende a infinito, por lo que en la práctica, si se tiene una p pequeña, con una N grande, o incluso moderada, se puede aproximar la distribución de la suma (no estandarizada) por medio de la ley Poisson (consultar Gnedenko y Kolmogorov, 1954). Es este otro ejemplo de convergencia en un ámbito más amplio, que corresponde a las leyes infinitamente divisibles.

→ Práctica Número 2. Teorema Central del Límite para variables aleatorias independientes con Varianza Finita.

Objetivo:

Reconocer que por el Teorema Central del Límite Clásico la convergencia de sumas de variables aleatorias a la distribución normal se aplica a sucesiones de variables aleatorias independientes idénticamente distribuidas con varianza finita.

Reafirmar el concepto de rapidez de convergencia y constatar que varía de acuerdo al tipo de variable que se trata.

* Actividad (a):
Simular: "2.Automática: Varía Distribución",
"1.Discretas con Varianza Finita"

- ¿Cuáles son las variables aleatorias consideradas en esta opción?
- ¿Qué características distinguen a las distribuciones simuladas en este grupo?
- ¿Se puede considerar que la función de distribución empírica se aproxima a la distribución normal estándar? ¿Visual y estadísticamente? ¿En todos los casos?
- ¿Es esta actividad una generalización de la Práctica No. 1? ¿En qué sentido?
- ¿Qué se puede concluir?

Conclusión:

Todas las variables aleatorias consideradas en esta actividad son discretas con varianza finita y las simulaciones realizadas debieron dar como resultado una aproximación satisfactoria a la normal estándar, por lo que la convergencia a esta distribución puede extenderse, hasta ahora, de la distribución Bernoulli a la familia de las distribuciones discretas con segundo momento central finito (ver, por ejemplo, Harris, 1966).

* Actividad (b):
Simular: "2.Automática: Varía Distribución",
"2.Continuas con Varianza Finita"

- ¿Qué variables fueron simuladas?
- Respecto a la actividad anterior, ¿Qué característica de las variables aleatorias cambió?
- ¿Cuáles son las conclusiones acerca de la aproximación a la distribución normal estándar? ¿Se sigue cumpliendo la convergencia? ¿Coinciden dichas conclusiones según los criterios visual y estadístico? ¿Se ha dado el peso correcto a las discrepancias en la parte central de las distribuciones y en los extremos?
- ¿Siguen siendo equidistantes los escalones de la distribución empírica? ¿Cuál es la explicación?
- ¿Cuánto vale la varianza de las distribuciones simuladas? ¿Es alguna de ellas infinita?

Conclusión:

La experimentación sugerida en esta actividad conduce a extender la convergencia de sumas de variables aleatorias independientes idénticamente distribuidas a la normal estándar, al caso en que las variables involucradas sean continuas (ver Harris, 1966 nuevamente). Por ello debe esperarse que en todas las simulaciones realizadas, la aproximación de la función de distribución empírica a la distribución normal se halla cumplido.

Las conclusiones alcanzadas en las actividades (a) y (b) se deben a que el Teorema Central del Límite asegura que cuando se toman sumas estandarizadas de variables aleatorias independientes e idénticamente distribuidas, basta con que tengan varianza finita para que su distribución converja a la normal estándar, sin importar que otra característica compartan con la variable aleatoria normalmente distribuida, como puede ser la continuidad, la simetría o el dominio no acotado.

→ Práctica Número 3. Sumas de variables aleatorias independientes con Colas Pesadas. Leyes Estables.

Objetivo:

Verificar que existen casos en que la distribución de sumas estandarizadas de variables aleatorias independientes e idénticamente distribuidas no se aproximan a la distribución normal y resaltar que se debe a la carencia de varianza finita.

Mostrar que la distribución normal pertenece a su propio dominio de atracción, considerada como una ley estable.

* Actividad (a):

Simular: "2.Automática: Varia Distribución", "3.Con Colas Pesadas".

- En las dos distribuciones simuladas, ¿Se aproxima la función de distribución empírica a la distribución normal?
- ¿En que falla?, ¿Qué comportamiento presentan las simulaciones que sea distinto al esperado?
- ¿Que características tienen en común con la distribución normal?, ¿Son continuas, simétricas o de dominio infinito?
- ¿Cuánto vale la varianza en cada caso?

- ¿A que se puede atribuir la no convergencia a la distribución normal?

Conclusión:

En efecto, existen sumas estandarizadas de variables aleatorias independientes e idénticamente distribuidas que no convergen a la distribución normal. Ello se debe a que la existencia de una varianza finita no sólo es una condición suficiente, sino que también es necesaria. Cuando este atributo no está presente, la función de distribución empírica muestra valores más grandes en términos absolutos que los correspondientes a la distribución normal, esto es, los "escalones" de la distribución empírica, corregida por factores de localización y escala, empiezan antes y terminan después que el intervalo seleccionado de ± 4 desviaciones estándar. Este fenómeno se presenta cuando se trata de variables aleatorias con "colas pesadas" y da lugar, como generalización, al estudio de las distribuciones estables, que es la única familia a donde pueden converger sumas estandarizadas de este tipo de variables, que no consideran necesariamente la condición de varianza finita (para mayor referencia consultar Gnedenko y Kolmogorov, 1954).

* Actividad (b):

Simular: "1.Automática: Variación de N", "8.normal"

- ¿Que valores de N se consideran? ¿Influyen en la convergencia?
- ¿Se cumple la convergencia a la distribución normal? Visualmente, ¿Fue la misma apreciación?
- ¿Cómo explica que no coincida la distribución empírica exactamente con la curva suave si se trata de la distribución normal en ambos casos?
- ¿Que tan buena es la aproximación observada?
- Intuitivamente, ¿Es lógico este resultado?

Conclusión:

Es sabido que la suma finita de variables aleatorias normalmente distribuidas es también normal, por lo que, con la adecuada estandarización, su distribución es la normal con media cero varianza uno, lo cual es mucho más que sólo afirmar que converge a ella (ver Harris, 1966). Sin embargo, existen otros argumentos, pues cabe señalar que por contar la misma distribución normal con varianza finita , el Teorema Central del Límite (consultar Harris, 1966) asegura que la suma estandarizada de una sucesión de variables aleatorias independientes distribuidas de acuerdo a la ley normal, convergen asimismo a la normal estándar. Refleja además, el hecho de que toda distribución estable pertenece a su propio dominio de atracción (ver Gnedenko y Kolmogorov, 1954), puesto que la misma distribución normal es un ejemplo de ley estable, y por lo tanto, se asegura de nueva cuenta que a ella converge una suma estandarizada de variables aleatorias normales.

→ Práctica Número 4. Sumas Aleatorias de Variables Aleatorias

Objetivo:

Verificar que la convergencia a la distribución normal de sumas estandarizadas se sigue cumpliendo aún cuando el número de sumandos también sea aleatorio.

* Actividad (a):
Simular: "4.Automática: Sumas Aleatorias", "3.geométrica"
"4.Automática: Sumas Aleatorias", "6.exponencial"

- Observe que aparece un número de variable aleatoria;
- ¿Coincide ese número con el valor de N?
- ¿Cómo es el comportamiento de N? ¿Estrictamente creciente?
- ¿Termina cumpliéndose la convergencia a la distribución normal estándar? (Continúe ejercitando su capacidad de observación: ¿Llega por su cuent a la misma conclusión que la prueba de Anderson-Darling? ¿Coincide el grado de su conclusión con el nivel de significancia expuesto por la prueba?
- ¿Podría decirse que ante la presencia de un número aleatorio de sumandos persiste la convergencia a la normal estándar? ¿Debería imponerse alguna restricción en esa aleatoriedad?

Conclusión:

Suponga que se toman sumas estandarizadas de variables aleatorias independientes idénticamente distribuidas con varianza finita, donde el número de sumandos es escogido aleatoriamente de acuerdo a la distribución de una sucesión de variables que sólo toman valores positivos, y que la probabilidad de tomar valores mayores a un número fijo, tiende a uno si el número de variables consideradas tiende a infinito. Entonces, bajo estas condiciones se puede asegurar (Rényi, 1976) que la convergencia de las sumas así formadas se realiza a la distribución normal estándar. Esto puede explicarse porque si bien el comportamiento del número de componentes en la suma es errático, finalmente termina por tender a infinito para lograr la aproximación deseada. ¿Las simulaciones presentaron evidencia en ese sentido? ¿Fueron sus conclusiones correctas? ¿Intuyó la necesidad de la condición impuesta al patrón aleatorio del número de sumandos (las variables aleatorias con valores positivos en el paquete son uniformes discretas con valores de 0 a T, con T aumentando en cada simulación)?.

→ Práctica Número 5. Cadenas de Markov.

Objetivo:

Mostrar que en algunos casos, la convergencia a la distribución normal sigue cumpliéndose cuando las variables aleatorias no son ya independientes. Ejemplificar cuando se trata de Cadenas de Markov de dos Estados.

* Actividad (a):
Simular: "1.Automática: Variación de N",
"15.cadenas de Markov"

- En una cadena de Markov de dos estados, ¿Cuántos valores puede tomar la variable aleatoria? ¿Qué probabilidades define la matriz de transición?
- Para cada valor de N, según su apreciación visual y la prueba estadística, ¿Se cumple la convergencia a la distribución normal? En terminos de rapidez,

¿Podría considerarse satisfactoria?

- ¿Puede considerarse cierto el postulado de que la distribución normal puede aproximar a la distribución de la suma estandarizada de variables aleatorias, aunque sean dependientes?

Conclusión:

La convergencia de sumas estandarizadas de variables aleatorias que forman una cadena de Markov a la distribución normal se cumple en el caso general, cuando se tienen n estados. En el caso particular de dos estados, por supuesto se sigue verificando la convergencia (ver Rényi, 1976), incluso con una alta rapidez. Si las probabilidades de transición de éxito a fracaso y viceversa suman uno, se rompe la dependencia y se reduce al caso de las sucesiones de ensayos Bernoulli. Por ello, se puede considerar de alguna manera este ejemplo como una generalización del teorema de deMoivre-Laplace, que ilustra además la convergencia a la distribución normal en el caso de variables aleatorias dependientes.

BIBLIOGRAFIA.

D'Agostino, S. (1986) « *Godness of fit Technics* ». M. Dekker, New York.

DeMoivre, A. (1738) « *The Doctrine of Chances* ». 2nd. ed.

Gnedenko, B.V. (1976) « *The Theory of Probability* ». MIR, Moscow.

Gnedenko, B.V., y Kolmogorov, A.N. (1954) «*Limit Distributions for Sums of Independent Random Variables* ». Addison-Wesley, Massachusetts.

Hamming, R.W. (1962) « *Numerical Methods for Scientists and Engineers* ». McGraw-Hill, New York.

Harris, B. (1966) « *Theory of Probability* ». Addison-Wesley, Massachusetts.

Kemeny, J.G., y Kurtz, T.E. (1985) « *True Basic Reference Manual* ». Addison-Wesley, Massachusetts.

Kemeny, J.G., y Kurtz, T.E. (1985) « *True Basic User's Guide* ». Addison-Wesley, Massachusetts.

Kennedy, W.y Gentle, J. (1980) « *Statistical Computing* ». M. Dekker, New York.

Rényi, A. (1976) « *Calculo de Probabilidades* » Reverté, S.A., Barcelona.

Shiryayev, A.N. (1979) « *Probability* ». Springer-Verlag, New York.

Todhunter, M.A., F.R.S. (1865) « *A History of the Mathematical Theory of Probability* ». Chelsea, New York.

Tucker, H.G. (1967) « *A Graduate Course in Probability* ». Academic

PROGRAMA CORAN. PARA EL ANALISIS DE CORRESPONDENCIAS EN MICROCOMPUTADORAS IBM-PC Y COMPATIBLES.

MEJIA AVILA CARLOS.*
VARGAS CHANES DELFINO.**

RESUMEN

En este trabajo se presenta el programa CORAN el cual permite realizar el Análisis de Correspondencias, con opción para el cálculo de coordenadas de variables o individuos suplementarios y obtiene gráficas en dos dimensionnes. Este programa se ha adaptado a microcomputadoras IBM-PC o compatibles, esta basado en el programa publicado por Lebart, L. et. al. (1984). Se encuentra a la disposición de las personas interesadas en su aplicacion.

INTRODUCCION

El análisis de correspondencias es una técnica fundamentalmente descriptiva multidimensional, que permite analizar variables discretas, registradas mediante tablas de contingencia o de tablas, cuyos elementos sean números positivos. Al igual que otros métodos de análisis de datos multidimensionales, éste es un método exploratorio en el sentido de que impone a los datos un mínimo de estructura en cuanto a hipótesis y modelos probabilísticos.

Permite obtener representaciones geométricas que muestran las proximidades entre renglones y columnas de una tabla cruzada, como puntos en un solo espacio con dimensión menor. Los renglones y las columnas pueden convivir en este mismo espacio para obtener una gráfica conjunta.

Debido a la gran cantidad de datos que se maneja con el análisis de correspondencias es indispensable el auxilio de algun programa computacional para poder llevarlo a cabo. Ya que esta tecnica es de uso reciente no se encuentra incluida en los paquetes de análisis estadistico mas comunes como el SAS o SPSS y solo se puede ejecutar construyendo MACROS en SAS (SAS Institute Inc. 1984) o utilizando programas de FORTRAN que corren en computadores de gran tamaño.

---

* CENTRO DE INVESTIGACIONES FORESTALES Y AGROPECUARIAS DE GUANAJUATO . Apdo. Postal # 112 Celaya, Gto.
** CENTRO DE INVESTIGACIONES FORESTALES Y AGROPECUARIAS DE TABASCO . Apdo. Postal # 17 Huimanguillo, Tab.

Por lo anterior, y ante la necesidad que existía en el INSTITUTO NACIONAL DE INVESTIGACIONES FORESTALES Y AGROPECUARIAS de poder utilizar esta tecnica en sus propios centros de trabajo, en donde solo se cuenta con equipo computacional compatible con la PC de IBM, se adaptó el programa que aqui se presenta.

PRINCIPALES CARACTERISTICAS DEL PROGRAMA:

A) El programa está escrito en FORTRAN IV, puede correr en computadoras de tamaño medio o grande. La versión que aquí se presenta fue compilado en una microcomputadora compatible con la PC DE IBM, usando el compilador de MICROSOFT FORTRAN 77 V3.20 02/84.

B) Puede manejar grandes matrices de datos, sin limitación práctica en el número de renglones (la matríz de datos nunca se almacena en memoria central). Cuenta con una subrutina de diagonalización fuera de memoria, bajo la cual el número de columnas a manejar puede ser también bastante grande; por ejemplo una matríz de (5000,500) podria manejarse bajo esa opción utilizando 64100 posiciones de memoria. Mas adelante se dá una guia para determinar el monto de memoria requerido.

C) El programa es autocontenido y cuenta con los procedimientos gráficos y numéricos necesarios.

D) Puede procesar tablas binarias tambien como tablas de contingencia. Aunque su ejecución, en el caso de grandes tablas binarias, no puede competir con un programa de análisis de correspondencias múltiples, trabajando directamente sobre una matríz codificada reducida.

PARAMETROS del PROGRAMA:

Para utilizar CORAN es necesario crear un archivo de datos, ya sea usando un procesador de textos como WORD STAR o SIDE KICK, o el editor de linea del sistema operativo EDLIN y definir en él los siguientes parámetros.

REG. 1.- TITULO. Titulo del Estudio en 80 caracteres.

REG. 2.- SEIS PARAMETROS con formato 6I4.
1. IEXA Número de renglones de datos en el archivo.
2. NQEXA Número de columnas en el archivo. (El identificador de renglon no se cuenta como una columna)

3. NVIDI Longitud del identificador de renglón.
Cada unidad en NVIDI representa 4 caracteres; máximo NVIDI=15, correspondiendo a 60 caracteres (los primeros 4, 8 o 12 aparecen en la gráfica). Este identificador es necesario y debe estar al principio de cada renglón.

4. LFMT Número de registros de formato
(Si no se pone LFMT=1)

5. MODIG Modo de definir selección de renglones.
Hay dos formas de selección y definición de renglones para análisis.
Si MODIG=0 todos los renglones estan activos.
Si MODIG=1 hay registros de selección.
Es decir algunos renglones son suplementarios o ignorados.

6. INDIM Indicador de forma de salida de resultados.
Si INDIM<>0 salida directa a impresora.
Si INDIM=0 salida al archivo CORRESPO.SAL

REG(S) 3. IDENTIFICADORES DE COLUMNAS.
Identificadores de cuatro caracteres son leídos en formato fijo (20A4) para las NQEXA columnas originales (antes de que las columnas sean seleccionadas).
Un registro por cada 20 nombres; por ejemplo si tengo 24 columnas, se requieren dos registros.

REG(S) 4. SELECCION DE COLUMNAS
Se leen indicadores, en formato 80I, para las NQEXA columnas de la matriz de datos, de la siguiente manera:
0 = columna eliminada del análisis.
1 = columna activa.
2 = columna ilustrativa.

REG(S) 5. SELECCION DE RENGLONES
Cuando MODIG=0 no hay registro de selección de renglones; todos estan activos.
Cuando MODIG=1 los indicadores se leen, con formato 80I, para cada uno de los IEXA renglones, con los siguientes codigos: 0 = renglón eliminado del análisis.
1 = renglón activo.
2 = renglón ilustrativo.
Si se tuvieran 350 renglones se requerirían cinco registros de indicadores para el caso MODIG=1.

REG(S) 6. FORMATO DE LECTURA DE DATOS.
El formato se escribe entre paréntesis en LFMT registros. Comienza con NVIDIxA4 (longitud de identificador de renglón). El resto se lee con formato real(F) aún para valores enteros.

REG(S) 7. DATOS

Los datos se leen de acuerdo al formato anterior. Deben existir IEXA renglones, cada uno con NQEXA+1 columnas. Los registros pueden ser de más de 80 caracteres y un renglón puede ocupar más de un registro.

REG. 8. SIETE PARAMETROS EN FORMATO 7I4

1.NFAC Número de coordenadas principales a calcular. Solo imprime seis.

2.LIST3 Parámetro para imprimir información de renglón:
0 = No se imprime.
1 = Se imprimen coordenadas y contribuciones de renglones. La información de columnas siempre se imprime.

3.NGRAF Número de gráficas (NGRAF<11).
En esta versión los planos principales son sucesivamente (1,2),(2,3),(3,4) etc..Todos los puntos -renglones y columnas activos o suplementarios - se muestran.

4.NPAGE Número de páginas para el ancho de cada gráfica. (NPAGE < 9).

5.NLING Número de líneas por gráfica. Se recomienda que NLING = 60 x NPAGE -2.
Si NLING=0,se usa la misma escala en ambos ejes.

6.JBASE Dimensión del espacio para aproximación en el caso de lectura directa:
Si JBASE=0, la diagonalización toma lugar en la memoria principal (selección usual).
Si JBASE=1 : se asume un valor de JBASE=NFAC+3.
En otro caso JBASE= Dimensión del espacio para aproximación.

7.NITER (Solamente si JBASE<>0), es el número de iteraciones en lectura directa. Si NITER=0 se usa NITER=8.

REQUERIMIENTOS DE ESPACIO PARA -CORAN-.

El programa principal contiene un vector Q, cuya longitud (Z)se define antes de compilar el programa, según las necesidades del usuario y la capacidad de la máquina en la que se correrá, que controla la asignación de memoria de todo el programa. La longitud del vector Q es función de las dimensiones de la matriz de datos a analizar y de la especificación de parámetros selecionada (en particular el modo de diagonalización).

La longitud de Q se puede calcular usando la siguiente ecuación:

Donde:
NVAR es el número total de columnas, activas e ilustrativas.
ITOT es el número total de renglones en el archivo de datos.
NACT es el número de columnas activas.
NFAC es el número de ejes principales pedido.
JBASE es el parámetro definido anteriormente.

$$N = \max[NAVR \times NFAC \; ; \; (ITOT + NVAR) \times 3]$$

Primer caso: JBASE = 0.

$$Z >= NACT^2 + 2 \times ITOT + 9 \times NVAR + 2 \times N$$

Segundo caso: JBASE<>0.

$$M = \max[N \; ; \; NACT \times JBASE]$$

$$Z >= JBASE^2 + 2 \times ITOT + 9 \times NVAR + 3 \times M$$

Ejemplo.- Si la matriz de datos es (1000,200), y si deseamos extraer seis factores, con todas las columnas y renglones activos:

NACT = NVAR = 200
NFAC = 6
ITOT = 1000

Entonces $N = \max(200*6 \; ; 3*(1000+200)) = \max(1200,3600) = 3600$
en el primer caso

$$Z >= (200)^2 + 2*1000 + 9*200 + 2*3600 = 51,000$$

en el segundo caso, asumiendo el valor de JBASE= NFAC+3= 9.

$$M = \max[3600 \; ; \; 200*9] = 3600$$

$$Z >= (9)^2 + 2*1000 + 9*200 + 3*3600 = 14,681$$

NOTA.- La versión que se presentó en el III Foro de Estadística y que esta a disposicion de quien la solicite, está compilada con una longitud del vector Q de 10,000. Lo que nos permitiría analizar una matriz de (800,50) bajo la opción JBASE=0; o una matriz de (700,120) con JBASE=1, pidiendo seis ejes principales en ambos casos.

RESULTADOS

El programa genera las siguientes ayudas para la interpretacion del análisis de los datos:

- Estudio de la tabla de valores propios. En esta aparecen los porcentajes de varianza acumulada por los factores. En la medida que estos se aproximen a 1 indica que la representación obtenida es de buena calidad.

Con base a estos valores se selecciona el número de factores a interpretar.

- Mediante el análisis de contribuciones absolutas podemos conocer, como una variable participa en la construcción de un eje. Si estas contribuciones absolutas son muy fuertes (del orden del 40 al 50% por ejemplo) son en general dudosas ya que expresan un cierto "desequilibrio de la síntesis".

- Las correlaciones elemento-factor al cuadrado o contribuciones relativas, son la parte tomada por un factor en la "explicación" de la dispersión de un elemento, al centro de gravedad.

- Estudio del plano factorial (gráficas) para interpretación

Dado que el Análisis de Correspondencias se realiza generalmente en dos partes; la primera es el tratamiento preliminar de los datos que consiste en inspeccionar el caracter de las variables, si estas son activas o suplementarias. La segunda consiste en realizar el análisis final con los elementos activos y ubicar los elementos suplementarios.

En este caso se correrá la primera vez el programa, utilizando la opciòn de grabar los resultados del análisis en un archivo, el cual se podrá accesar utilizando un procesador de palabras como SIDEKICK, para que en base a las ayudas para la interpretación que ofrece el programa, podamos decidir cuales son las variables activas y cuales suplementarias, así como el número de factores a interpretar, para posteriormente realizar la corrida definitiva imprimiendo los resultados.

Para verificar la exactitud del programa se compararon corridas de datos analizados con una MACRO de SAS (Vargas Ch. D. 1986) y una versiòn del programa que corre en una computadora IBM 370-135, usando FORTRAN con variables de doble presición. Los resultados obtenidos fueron muy satisfactorios pues generalmente el tiempo de proceso fue moderado (dos minutos) para matrices de 50 x 20 elementos, y la diferencia numérica en los resultados obtenidos no fue relevante.

CONCLUSION

Consideramos que el poder contar con este programa facilitará el uso del Análisis de Correspondencias, ya que nos permite efectuarlo sin necesidad de equipo computacional grande, situación que generalmente enfrentan las Universidades y Centros de Investigación en provincia.

BIBLIOGRAFIA

Lebart,Ludovic et.al.(1984). Multivariate Descriptive Statistical Analysis. Jhon Wiley & Sons.

SAS Institute Inc. (1984). A user guite to SAS (Statistical Analysis System). Cary, NC, USA.

Vargas Chanes Delfino (1986). Aplicación del Análisis de Correspondencias en el estudio de la interacción medio-ambiente*vegetación en el Valle de Apatzingán. (Tesina, ined) Instituto de Investigaciones de Matemáticas Aplicadas y Sistemas. UNAM.