

LA REVISTA DE LA CARRERA DE ESTADISTICA
VARIANZA



Universidad Mayor de San Andrés
Facultad de Ciencias Puras y Naturales



AÑO 3

Nº 3

Diciembre 2003

REVISTA VARIANZA

Número 3, Año 3
Noviembre del 2003

Dirección:

José Aníbal Angulo A.
Director IETA

Colaboradores

Rubén Belmonte

David Barrera

Dindo Valdez

Anibal Angulo

Nicolás Chávez

Raúl Delgado

Fernando Rivero

Jaime Pinto

Jorge Troche

Jaime Chumacero

Teresa Reinaga

Impresión:

IDEAGRAF

Calle Bueno # 370
Esq. Obispo Cárdenas
Telf. 2202115

PRESENTACIÓN

Leer lo que Bolivia produce es conocer lo que Bolivia es. Una oración que leí hace algún tiempo y respresenta una afirmación acertada en el caso que nos ocupa. La Revista Varianza representa, en esta oportunidad, lo que es la Carrera de Estadística. Los aportes que usted amable lector encontrará en estas páginas, reflejan una inquietud personal e institucional que se hace en una gestión, de parte de los autores.

Escribir la presentación de la Revista Varianza No. 3 es un compromiso y una satisfacción grata. En ella está inserta el rol que tiene la estadística en el contexto de las aplicaciones y el potencial como ciencia.

Espero que cada artículo y contribución sirvan para reflexionar en sus contenidos y sirvan de aliento y motivación para los colegas y otra personas interesadas en hacer llegar nuevos aportes para el próximo número.

Lic. Luis Zapata Escobar

*Carrera de Estadística
Instituto de Estadística Teórica y Aplicada
Facultad de Ciencias Puras y Naturales
Universidad Mayor de San Andrés*

**La Paz - Bolivia
Edificio Antiguo Planta Baja pb05
Tel: 2442100**

INDICE

LA ESPADA DE DAMOCLES DE LA ESTADÍSTICA_____	1
EL ÍNDICE KAPPA DE CONCORDANCIA_____	7
Rubén Belmonte Coloma	
ECONOMETRÍA - BOOTSTRAP - ECONOMETRÍA ESPECIAL_____	12
David Barrera Ojeda	
LA DESCOMPOSICIÓN ORTOGONAL DEL TEST χ^2 DE PEARSON APLICADO AL TEST DE FRIEDMAN_____	22
Dindo Valdez Blanco	
PANORAMA GENERAL DE LOS MÉTODOS MULTIVARIADOS_____	29
Anibal Angulo Angulo	
SISTEMAS DE INVENTARIOS MRP_____	36
Nicolás Chávez Quisbert	
ALTERNATIVAS_____	39
Raúl Delgado Álvarez	
UNA ALTERNATIVA METODOLÓGICA PARA EL CENSO NACIONAL AGROPECUARIO EN BOLIVIA	45
Fernando Rivero Sugiura	
ESTADÍSTICAS AGRÍCOLAS PLANES DE TIPO DE MUESTREO_____	48
Jaime Pinto Ajhuacho	
ALGUNOS PROBLEMAS DE PROBABILIDAD_____	54
SOLUCIONES A LOS PROBLEMAS PLANTEADOS EN LA VARIANZA 2_____	55
Anibal Angulo Angulo	



LA ESPADA DE DAMOCLES DE LA ESTADÍSTICA

¿Existe una espada de Damocles en la Estadística?. Expondremos distintas razones, y siempre que se opte por potenciar el cálculo numérico sin acompañarlo con el razonamiento estadístico, la respuesta no puede dejar de ser afirmativa. La decisión de optimizar la rapidez de cálculo y el ahorro de tiempo derivado, junto a facilitar una herramienta técnica y favorecer su difusión, conlleva que la espada de Damocles esté también sobre nuestras cabezas.

En última instancia es una decisión humana aventurarse en el uso de una tecnología sin los conocimientos necesarios. Quizá también aquellos que participan en la formación estadística, no hemos sido muy eficaces a la hora de explicar que una cosa es la instrucción estadística, necesaria para todos, y otra, muy distinta, que de ésta se derive la capacitación para realizar análisis estadísticos. También es posible que algunos docentes hayan actuado más como mercaderes de la Estadística que como difusores del razonamiento estadístico. Eso sería cierto si, durante el proceso de enseñanza, se ha puesto más énfasis en cómo se pulsan botones que en por qué se realizan las acciones en el entorno de una aplicación informática, asemejándose más a una venta de estadística que a una formación sobre la misma.

Pero el problema es mucho mayor y con muchos más agentes incidiendo en el futuro de la Estadística. Las revistas científicas requieren análisis estadístico en los artículos y, por descontado, la mayoría de revistas no son de contenido estadístico ni sus autores tiene como campo de trabajo la Estadística; pero se nota un apreciable incremento en la cantidad de análisis estadísticos que se mencionan en la publicaciones científicas (Smith, 1996).

Una buena parte de docentes universitarios transmiten a sus colegas la necesidad de utilizar técnicas estadísticas, aunque la mayoría de los alumnos dedican a su actividad profesional a cuestiones escasamente relacionadas con la Estadística. Su función profesional no será llevar a cabo investigaciones o estudios ni para el desarrollo científico ni para obtener nuevo conocimiento.

También se exige con mayor énfasis a las administraciones públicas que justifiquen sus acciones y eso se traduce en una extensa cantidad de resultados estadísticos. Las empresas privadas destinadas al análisis de datos insisten en la



importancia de extraer información y las empresas cuantifican sus procesos para maximizar beneficio y reducir costos. No olvidemos que algunas empresas han formado a sus operarios en el control estadístico de procesos, implicando a los mismos en la mejora del sistema productivo.

Todas estas fuerzas, entre otras, colaboran en transmitir la idea según la cual la estadística debe ser conocida por todos. Y, como parece potenciarse el saber cómo se hacen las cosas más que el comprender aquello que se realiza, resulta difícil exigir a las personas que se limiten sólo a un conocimiento teórico e intelectualmente enriquecedor. Por tanto, conocer es aplicar. En consecuencia, y como la cantidad de datos disponibles crece sin cesar, la mayoría de personas que tengan posibilidad y necesidad de utilizar técnicas estadísticas recurrirán a las aplicaciones informáticas. Pero, y para un uso riguroso y responsable del análisis estadístico en la investigación y la docencia, la utilización de los ordenadores y las aplicaciones informáticas no dejan de constituir una espada de Damocles sobre nuestras cabezas. Al igual que en la metáfora, ordenadores y las aplicaciones informáticas pueden o no caer, ello dependerá de que seamos conocedores de los posibles peligros que de su utilización pueden derivarse y de que sepamos evitarlos.

La primera razón a la cual haremos referencia entre aquellos problemas derivados de las aplicaciones informáticas es que favorecen la automatización del análisis estadístico. Podemos establecer un símil entre las actuales aplicaciones informáticas para el análisis de datos y las máquinas. No sería nada complejo instruir a una persona sin conocimientos de estadística para que se responsabilizara del control estadístico de procesos indicándole que cuando ocurra uno de entre una serie de hechos avisará a un técnico. No se advierte una diferencia entre el manejo de una máquina reducido a una serie de rutinas y el análisis estadístico mediante estas aplicaciones.

Una serie de acciones, se comprenda o no lo que implica, lleva a un resultado. ¿cuál es el problema? simplemente que no se requiere un especialista. Mientras un experto sigue un proceso de formación continuamente, además de los conocimientos que ya posee, en la mayoría de los casos no podemos esperar que aquellas personas, que operan sobre estas aplicaciones de forma rutinaria tengan un interés especial por mejorar sus conocimientos estadísticos. Por mucho que a quienes estamos en el campo de la estadística nos apasione esa temática, no podemos olvidar que otras personas contemplan esos análisis como un trabajo rutinario. En consecuencia, la automatización



no es una garantía para extraer toda información relevante de los datos con? la correcta utilización de las técnicas.

No abundaremos en la idea, pues se ha tratado ya, pero el uso masivo de las aplicaciones informáticas no permite poseer un mínimo de garantías sobre la calidad de los análisis realizados, una opinión con la cual no creemos que los diferenciamos en exceso de otros autores (Abraira, Cadalso, Gómez, Martín y Pita 2001). La tecnología estadística está, y su facilidad de manejo también, en una sociedad fuertemente competitiva la ventaja diferencial es fundamental y, por esa razón, quien pueda demostrar un dominio de la tecnología que otros no poseen tiene un valor añadido en su perfil profesional. La estrategia óptima consiste en disminuir la ventaja del competidor y, si es posible, obtener otra que no posea. En esta dinámica está la Estadística y sus aplicaciones informáticas, pues son dos aspectos del conocimiento que pueden permitir alcanzar un perfil profesional.

Consecuencia de la automatización y la masiva disusión y utilización, se detecta una creciente escasez de rigor estadístico. Se advierte la carencia de conocimientos teóricos y flagrantes errores. De nuevo la razón es la facilidad de acceso a ese tipo de tecnología en ausencia de firmes conocimientos estadísticos. En parte la responsabilidad de las empresas productoras y distribuidoras de estos programas, pues el diseño del producto se favorece que cualquier persona no experta pueda operar ese sistema, mientras los distribuidores enfatizan la facilidad con que se pueden obtener interesantes resultados estadísticos. Pero también existe un error en el enfoque docente, si no se consigue transmitir que conocer algo de Estadística no es suficiente para realizar análisis Estadísticos. Por supuesto utilizar esas aplicaciones e interpretar los resultados es una decisión libre, pero no sería desaconsejable una dosis de autocrítica.

Un problema muy importante que se deriva de las aplicaciones informáticas es que el usuario, en general, se ciñe a la oferta de técnicas estadísticas que proporciona un programa. Eso se debe seguramente a dos motivos. Primero, es muy probable que sólo disponga de una única aplicación informática para el análisis de datos. Segundo, muchos usuarios considerarán que un programa de renombre debe contener todas las técnicas estadísticas que se corresponden con los distintos problemas de investigación. La segunda razón es especialmente preocupante porque, otra vez, aparece la pertinaz idea de la existencia de una única vía entre el problema de investigación y la técnica estadística. Ninguna aplicación informática para el análisis de datos es una Biblioteca de



Alejaría en la cual el usuario hallará la totalidad de técnicas que puedan ser necesarias para realizar un análisis estadístico, y eso sin entrar en los diferentes coeficientes e índices que pudieran ser precisos en esa investigación.

No es muy habitual que en los resultados que proporcionan los programas se advierta al usuario del incumplimiento de los supuestos requeridos por las técnicas estadísticas, aunque su violación resulta en general, crítica (Hartmann,1974). Estas advertencias son más la excepción que la regla. Si el usuario posee conocimientos estadísticos suficientes, el problema queda mitigado. No afirmamos que se elimina el problema debido a que ningún estadístico es un profundo conocedor de la totalidad de técnicas estadísticas. Cuando quien realiza el análisis posee escasos conocimientos de estadística, las consecuencias en la estimación y la decisión estadística pueden resultar catastróficas.

Pero el punto crítico que queremos destacar en este momento es que resulta sorprendente esta aparente falta de interés por los supuestos en las aplicaciones informáticas. ¿Por qué estos programas no proporcionan mensajes escritos sobre los supuestos en los listados de resultados? ¿Cuál es la razón por la cual no se advierte al usuario sobre la violación de los supuestos?. No es tan complejo ni supone constreñir al analista de datos; al contrario, se agradecerían los mensajes de aviso. La explicación reside en los inicios de estas aplicaciones informáticas. Al principio eran programas cuyo mercado era única y exclusivamente el ámbito universitario, donde la mayoría de investigadores que los utilizaban tenían conocimientos suficientes para realizar los análisis estadísticos. Para acabar de complicar la situación, esas iniciales aplicaciones requerían conocer un complejo conjunto de comandos, siendo necesario elaborar algo parecido a un programa informático para obtener un análisis estadístico, siempre bajo entornos no gráficos, como eran por ejemplo, los sistemas operativos UNIX y DOS de la época. Incluso podemos remontarnos unos años antes y recordar que los análisis estadísticos se obtenían mediante una tarjeta perforada. De hecho eran unas pocas personas las que realizaban sus propios análisis estadísticos y los de los demás. En, salvo algunas técnicas muy simples y de fácil cálculo, el resto de investigadores no utilizaban complejas técnicas multivariantes.

Retomando la pregunta, la mayoría de aplicaciones que hoy en día se comercializan para el análisis de datos en entornos o interfaces gráficas de usuario arrancan de aquellos tiempos que parecen remotos. Un notable número de sus actuales



procedimientos se remontan a aquellos días. Ciertamente han sido retocadas las salidas de resultados, aprovechando las mejoradas posibilidades que ofrecen los entornos gráficos. Eso implica modificar el código informático, ya sea cambiando valores de variables de salida o mejorando la salida gráfica con nuevo código. Este es el tipo de modificación que ha primado, junto a desarrollar un código nuevo para dotar a las aplicaciones de procedimientos totalmente nuevos. Ahora bien, completar el viejo código o dotar al nuevo código informático de líneas de instrucciones para advertir al usuario sobre la violación de los supuestos no ha sido una prioridad. Es más, en muchos casos los resultados para poder realizar el análisis de los supuestos no se halla en las opciones por defecto de los procedimientos. Mientras el análisis estadístico sea realizado por expertos el problema es relativamente poco importante. Pero no olvidemos que para estas empresas informáticas los consumidores ya no son sólo expertos estadísticos y, como era de esperar, el escaso énfasis en los supuestos que arrastra desde los inicios tienen efectos descontrolados sobre la decisión estadística. En general, el programa proporciona el resultado se cumplan o no los supuestos.

Hemos definido que deben entenderse las aplicaciones informáticas para el análisis de datos como productos. Desde esta perspectiva, las empresas que desarrollan estos programas están obligadas a mantener procedimientos y algoritmos durante un tiempo para asegurar su rentabilidad no siendo evidente que estas compañías respondan con rapidez a los avances teóricos e incorporen con prontitud estos desarrollos en sus programas. Pero un problema más importante es que se mantienen algunos procedimientos anticuados o cuestionados, que seguramente se utilizan por la seguridad que transmite a muchos usuarios el aval que supone su inclusión en una aplicación informática de renombre punto aparte?

(Bartholomew, ¿1997). Dionisio el Tirano (S.IV a. de J.C.) quiso que uno de sus cortesanos, Damocles, comprendiera cual es la felicidad de un Rey pleno de grandeza. Ordenó a sus servidores que tratasen al cortesano como a él mismo e invitó a Damocles a un fastuoso festín, donde ocupó el lugar del Rey. Se encontraba el cortesano embriagado por tanta felicidad cuando, al alzar su vista, vio una pesada espada colgada sobre su cabeza y sujeta a la crin de un caballo. Todavía llena, cayó la copa de las manos de Damocles, entendiendo el cortesano cual era la felicidad de un tirano. La espada de Damocles es una alegoría que se refiere al peligro que amenaza a un hombre en medio de una aparente prosperidad. En este sentido tomamos la alegoría para referirnos a los



inconvenientes, desventajas y peligros de una utilización indiscriminada de las aplicaciones informáticas destinadas al análisis de datos.

Autor Joan Manuel Batista- Foguet

Análisis Estadístico con Aplicaciones Informáticas: ¿Ayuda o sustituye el Razonamiento?

----- O -----

El Mundo Estadístico

Normalmente se piensa que los aviones con cuatro motores son más seguros que los que sólo tienen dos. Esto es totalmente falso, como se indica en la pagina 14 de Air & Space, agosto y septiembre 1993: “cuantos menos motores, menor probabilidad de que alguno de ellos se estropee”. Por tanto, los aviones más seguros son los que tienen un solo motor e INCLUSO NINGUNO.

En realidad, volar en avión es muy seguro. La práctica totalidad de los fallecidos en accidentes aéreos han muerto al llegar al suelo.

Un hombre tenía miedo de tomar un avión por aquello de los secuestros aéreos. Mirando unas estadísticas, encontró que la probabilidad de que hubiese una bomba en su vuelo era de una entre mil, mientras que la probabilidad de que hubiese dos era de una entre 100000 por lo tanto, tomó el avión llevando él mismo una bomba.

Durante la Segunda Guerra Mundial, a alguien se le ocurrió la idea de mirar dónde habían sido tocados los aviones al volver de sus misiones y-reforzar esos puntos. Así que se empezaron a hacer estadísticas acerca de qué zonas del avión estaban más expuestas. Al analizar los resultados, se dieron cuenta de un pequeño detalle: lo que había que reforzar eran las zonas que recibían más balazos de los aviones que NO volvían de sus misiones.





EL ÍNDICE KAPPA DE CONCORDANCIA

Rubén Belmonte Coloma

Muchos estudios de las áreas biológicas, psicológicas, sociológicas o politológicas requieren un análisis de concordancia entre variables ordinales, en general este análisis suele quedarse en una simple relación no direccional, mediante la prueba Chi-cuadrado ó la de verosimilitud.

Los avances del análisis de datos en los últimos tiempos hacen énfasis en algunos estadísticos de asociación con dirección, entre ellos, los de concordancia tienen un uso cada vez más frecuente.

En algunos casos, como es el caso en salud pública, se desea observar la concordancia entre varios métodos o evaluadores que clasifican al paciente (o el resultado de una observación) según una serie de posibilidades (categorías) mutuamente excluyentes. Partiremos de un análisis del caso dicotómico donde se desea comparar dos métodos de clasificación, tal como se presenta en el siguiente cuadro:

		Método B		Total
		Positivo	Negativo	
Método A	Positivo	a	c	f1
	Negativo	b	d	f2
Total		c1	c2	n

Intuitivamente la expresion:

$$(a + d) / n$$

Conocida como proporción de coincidencias frente al total de sujetos: puede ser una medida de concordancia, sin embargo, es previsible que encontremos concordancia entre ellos por puro azar, por ejemplo arrojando dos monedas que simulen los métodos es previsible encontrar en promedio del orden de un 50% de coincidencias.

Bajo el supuesto de que el sistema A sea un método establecido con todo el rigor científico y el método B es la opinión probablemente subjetiva de un individuo cualquiera; es previsible encontrar un cierto grado de concordancia debido en parte al azar.

Para salvar el efecto del azar se define el índice de concordancia kappa dado por Cohén (1960) de la siguiente manera:



$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Donde P_o es la proporción de concordancia observada y P_e es la proporción de concordancia esperada por puro azar.

En caso de acuerdo perfecto la concordancia será 1, por lo que $1 - P_e$ representa el margen de acuerdo posible no atribuible al azar. De ese margen nosotros observamos probablemente sólo una parte $P_o - P_e$, salvo que haya acuerdo perfecto $P_o=1$.

Así pues, en caso de concordancia perfecta el valor de kappa es 1; si la concordancia observada es igual a la esperada kappa vale 0; y en el caso de que el acuerdo observado sea inferior al esperado el índice kappa es menor que cero.

Para calcular P_e , la concordancia esperada, el razonamiento es el siguiente: de acuerdo con la tabla anterior la probabilidad de que el método A clasifique a un sujeto como positivo podemos estimarla como f_1/n ; mientras que la correspondiente probabilidad del método B la estimaremos como c_1/n .

Si consideramos que existe independencia entre ambos métodos de clasificación, la probabilidad de que coincidan clasificando al mismo sujeto como positivo será entonces el producto de las dos probabilidades (sucesos independientes). De la misma manera que se calcula los valores esperados para la construcción del estadígrafo Chi-cuadrado.

Aplicando el mismo razonamiento calculamos la probabilidad de que se produzca acuerdo entre los métodos al clasificar a un sujeto como negativo, y entonces la probabilidad de acuerdo para cualquiera de las dos clasificaciones será la suma de ambos valores, esto es:

$$P_e = \frac{f_1 \cdot c_1 \cdot f_2 \cdot c_2}{n^2}$$

La distribución exacta del estadígrafo kappa requiere de un tratamiento que no se presenta en este artículo, sin embargo para muestras multinomiales, el estimador de k es asintóticamente normal y su varianza puede encontrarse en Agresti (1990).

Según Cohén y Swerdlik (1998), el coeficiente Kappa fue propuesto originalmente por Cohén (1960) para el caso de dos evaluadores o dos métodos, por lo que a menudo



se le conoce como Kappa de Cohén, y fue generalizado para el caso de más de dos evaluadores por Fleiss, por lo que a veces también se habla del índice Kappa de Fleiss.

Desde el punto de vista descriptivo Landis y Koch propusieron unos márgenes para valorar el grado de acuerdo en función del índice Kappa que se presentan en la siguiente tabla:

Kappa	Grado de acuerdo
<0	Sin acuerdo
0 - 0,2	Insignificante
0,2 - 0,4	Bajo
0,4 - 0,6	Moderado
0,6 - 0,8	Bueno
0,8 - 1	Muy Bueno

Este índice se puede generalizar para clasificaciones multinomiales (más de dos categorías) y para más de dos evaluadores, siendo similar su interpretación.

En el caso de más de dos categorías, además del índice de concordancia global puede ser interesante determinar el grado de concordancia específico en alguna de las categorías (o en todas), lo que equivale a convertir el resultado posible en dos únicas respuestas. De esta manera para cada una de las categorías vamos convirtiendo la tabla original en tablas 2x2 y podemos entonces calcular el valor del correspondiente índice kappa como si se tratara de una variable dicotómica.

La gran utilización del índice de concordancia Kappa se debe probablemente tanto a la facilidad de cálculo, como a su clara interpretación; no obstante, tiene sus problemas y limitaciones.

El principal problema de esta medida de concordancia radica en que está pensada para clasificaciones nominales, en las que no existe un orden de graduación entre las diferentes categorías. El índice Kappa hasta ahora descrito únicamente tiene en consideración si hay o no acuerdo, esto es si se “clasifica o no al sujeto en la misma categoría.

Si deseamos tener en cuenta el hecho de que estamos manejando variables ordinales para calcular una medida de concordancia, existen diferentes posibilidades. La más sencilla es calcular individualmente la concordancia en cada categoría, pero de esta



forma seguimos sin ponderar el nivel de desacuerdo global según esa clasificación ordinal.

Otro enfoque más global consiste en asignar un “peso” a las diferentes posibilidades de desacuerdo, de tal manera que se considere como más importante un desacuerdo entre categorías alejadas que entre las próximas. Este peso variará entre 0 (acuerdo, misma categoría) y 1 (desacuerdo con categorías extremas).

El problema surge a la hora de determinar esos “pesos”, ya que el valor de concordancia obtenido será diferente según los pesos utilizados. Existen los pesos más utilizados (lineales o bicuadrados) y que suelen proporcionar por defecto los programas de computación y que son detallados a continuación.

Índice Kappa con “pesos”

Una solución que puede verse como intermedia entre las (un único kappa global o K kappas individuales para cada categoría), pero que, en general, sólo tiene sentido para variables ordinales, es el denominado Kappa ponderado, también propuesto por Cohén, en el cual se asignan unos pesos para cuantificar la importancia relativa entre los desacuerdos.

La idea de este índice ponderado es asignar a cada celda de la tabla un peso w_{ij} comprendido entre 0 y 1 que represente la importancia del desacuerdo. Dando el máximo peso al acuerdo perfecto, y pesos proporcionalmente menores según la importancia del desacuerdo

$$w_{ii} = 1 \quad 0 \leq w_{ij} < 1 \quad \forall i \neq j$$

Además, obviamente $w_{ij} = w_{ji}$.

Las proporciones ponderadas de acuerdos observados y esperados se definen:

$$P_{o(w)} = \frac{\sum_{i=1}^K \sum_{j=1}^K w_{ij} X_{ij}}{N} \quad P_{e(w)} = \frac{\sum_{i=1}^K \sum_{j=1}^K w_{ij} X_i \cdot X_j}{N^2}$$

y a partir de aquí, el Kappa ponderado

$$\kappa_w = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}}$$



Nótese que, en el caso extremo, si se definen los pesos como:

$$w_{ij} = 0 \quad \forall i \neq j; \quad w_{ii} = 1$$

el Kappa ponderado coincide con el Kappa global.

La principal ventaja del Kappa ponderado reside en la posibilidad de cuantificar diferentes grados de desacuerdo. Los valores de los pesos dependerán, en cada caso, de la importancia que se conceda a cada desacuerdo, hay que tener presente, sin embargo, que ello añade cierta dificultad a su interpretación: si en dos estudios diferentes, se calcula el Kappa ponderado con dos sistemas de pesos distintos, es difícil realizar comparaciones entre ellos. Los más usados en este sentido son, por su sencillez, los denominados pesos lineales, propuestos inicialmente por Cohén

$$w_{ij} = 1 - \frac{|i - j|}{k - 1}$$

y los denominados pesos bicuadrados, propuestos posteriormente por Fleiss y Cohen

$$w_{ij} = 1 - \frac{(i - j)^2}{(k - 1)^2}$$

Pesos lineales y bicuadrados para 3 categorías:

Categoría	Lineal (wl)			Bicuadrado (wb)		
	1	2	3	1	2	3
1	1	0,5	0	1	0,75	0
2	0,5	1	0,5	0,75	1	0,75
3	0	0,5	1	0	0,75	1

Bibliografía:

Agresti Alan. 1990. Categórica! Data Analysis. John Wiley and Sons New York.

Cohén J. Swerdlik M. 2000 Pruebas y Evaluación Psicológicas Mc Graw Hill México.

Cohén J. 1968 Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or parcial credit Psychol Bull.





ECONOMETRÍA - BOOTSTRAP - ECONOMETRÍA ESPECIAL

David Barrera Ojeda

Objetivo

La idea del presente trabajo, es un intento de mostrar los pasos a seguir antes de llegar a aplicar econometría espacial, cuando las variables son georeferenciadas, es importante señalar que se debe agotar las otras que nos brinda la econometría clásica, es importante señalar que la mayoría de las pruebas de la econometría clásica no se replican en econometría espacial, porque en el fondo es la aplicación de MCP (mínimos cuadrados ponderados).

En algunos trabajos es inevitable, su aplicación, si así fuese existen varias técnicas para determinar la matriz de contigüidad, en el presente trabajo se emplea la más simple.

Introducción

Uno de los importantes avances en la ciencia económica de la última década ha sido la reincorporación explícita del efecto del espacio geográfico en el análisis de los problemas económicos. A partir de los trabajos de Krugman (1991a y 1991b, et al 1998) sobre lo que se ha llamado la “nueva geografía económica”, resaltando el papel de las externalidades espaciales en los modelos de comercio internacional y crecimiento, se han multiplicado los modelos que estudian la influencia del espacio sobre la localización de empresas, desarrollo de complejos industriales, desarrollo de las inversiones sociales, difusión del conocimiento y la tecnología, etc.

Econometría espacial

Anselin (1988), es probablemente la referencia más citada en los trabajos de econometría espacial, la define como “la colección de técnicas que lidian con las peculiaridades causadas por el espacio en el análisis estadístico de los modelos de la ciencia regional”. Once años más tarde, Anselin (1999) extiende la definición diciendo que “la econometría espacial es una rama de la econometría que se preocupa del tratamiento



adecuado de la interacción espacial (autocorrelación espacial) y la estructura espacial (heterogeneidad espacial) en modelos de regresión con datos de corte transversal y de panel de datos”. Para más detalle, Paelinck y Klaassen (1979).

Ejemplo Práctico

Se estudia, 15 municipios de la ciudad de La Paz, ¿por qué solamente 15?, debido a las siguientes razones:

1. La mayoría de los municipios no poseen informes, de las cuatro variables del modelo. No hay lugar para los missing.
2. Se reduce a las observaciones colindantes, puesto que para aplicar econometría espacial, necesariamente los municipios deben ser colindantes.

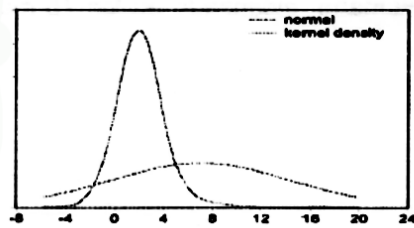
Las variables a estudiar son:

- INVSM: Inversión Social Municipal, son los recursos que el FIS desembolsa en los municipios (año 1997).
- RECOP: Recursos Propios, son los recursos propios con que cuenta cada municipio (año 1997).
- PAUS: Es el porcentaje de ausentismo electoral, en las elecciones municipales de 1995.
- IDH: Es el índice de Desarrollo Humano, alcanzado en los 15 municipios (año 1997).



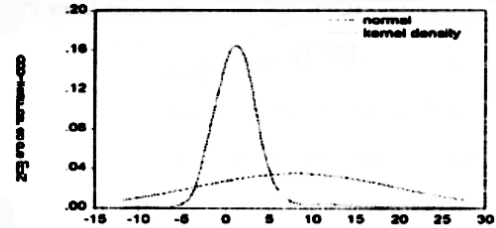
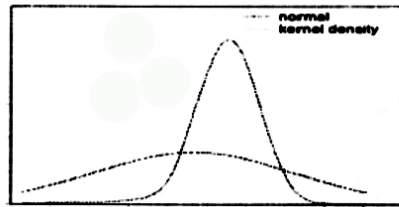
1. Econometría Clásica
1.a. Pruebas Bootstrap

Número de réplicas=1000				
Estadísticos de los T-student				
Coefficientes	cons	recop	paus	idh
Mean	2.142	1.694	0.96 2	1.16 4
Median	2.060	1.353	0.99 2	1.13 1
Skewness	0.749	7.654	0.23 1	0.68 6
Kurtosis	5.354	118.71 9	8.76 3	5.44 2
Test de Normalidad				
Shapiro-Fran- cia	8.102	9.217	7.42 9	7.18 2
Shapiro-Wilk	9.917	11.704	8.91 3	8.49 5
Sesgo	-0.248	0.510	0.04 9	0.35 3



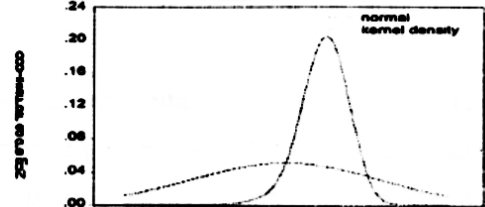
Gráfica de la función de densidad de la variable estandarizada

T-student beta3



gráfica de la función de densidad de la variable estandarizada

T-student beta4

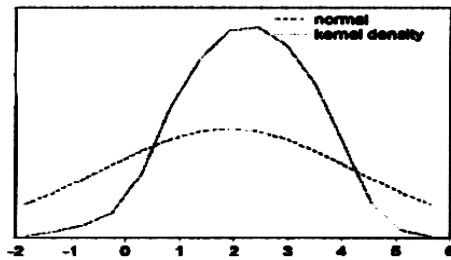




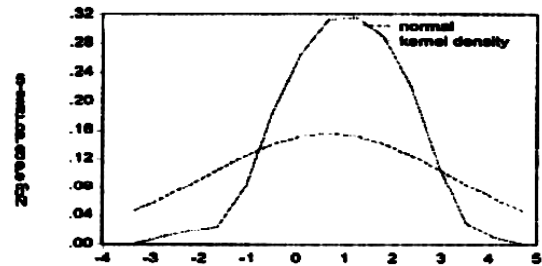
La presencia de los valores outliers, ocasiona problemas serios, como por ejemplo: Los T-stat son bastantes inestables, la significación recop (recursos propios) es la más afectada, en el sesgo se debe tener cuidado, puesto que se está tomando la media, puede ocasionar inconsistencias por el problema de la poca normalidad de los T-student.

Prueba Jackknife

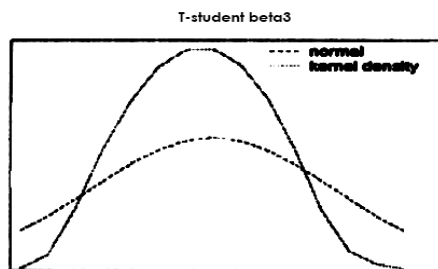
Pruebas Jackknife				
Número de obs.= 15				
Jackknife de los T-student				
Estadísticos	C	RECOP	PAUS	IDH
Media	2.237	1.016	-0.963	-1.384
Error.est	2.183	2.529	1.280	2.179
Sesgo	-2.632	-0.171	1.012	2.090



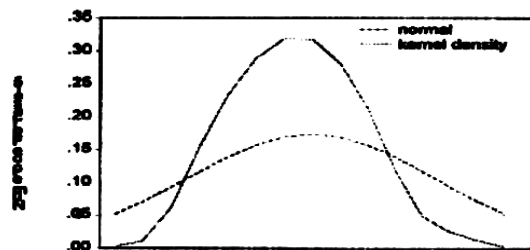
Gráfica de la función de densidad de la variable estandarizada



gráfica de la función de densidad de la variable estandarizada



T-student beta3



T-student beta4

Con la técnica de remuestreo jackknife, ocurre que el sesgo es mínimo, en la significación de recop, sin embargo es la variable que más varía.



1.b. Estimadores Robustos

Para subsanar ese problema: una alternativa es obtener estimadores robustos:

Regresión Robusta				
Método=Huber				
Variabes	Coficientes	Erorr.Est.	T-stat	Prob
C	21.45	12.979	1.653	0.127
RECOP	-0.095	0.127	-0.747	0.471
Gft	-11.524	11.139	-1.035	0.323
IDH	-8.976	26.548	-0.338	0.742
Suma de los residuos al cuadrado=404.318904				
Error Estandar de la regresión=6.062694451				
Valor de Máximo Verosimilitud= -45.99023099				

Regresión Robusta				
Método=Ramsay				
Variabes	Coficientes	Erorr.Est.	T-stat	Prob
C	11.293	6.552	1.724	0.113
RECOP	-0.183	0.062	-2.96	0.013
PAUS	-9.871	5.416	-1.822	0.096
IDH	12.46	13.162	0.947	0.364
Suma de los residuos al cuadrado=114.6177953				
Error Estandar de la regresión=3.227971433				
Valor de Máximo Verosimilitud=-36.53572455				

Regresión Robusta				
Método=Andrew				
Variabes	Coficientes	Erorr.Est.	T-stat	Prob
C	17.753	11.261	1.576	0.143
RECOP	-0.143	0.111	-1.29	0.223
PAUS	-10.839	9.541	-1.136	0.28
IDH	-0.77	23.085	-0.033	0.974
Suma de los residuos al cuadrado=320.7480712				
Error Estandar de la regresión=5.399899589				
Valor de Máximo Verosimilitud= -44.25362142				



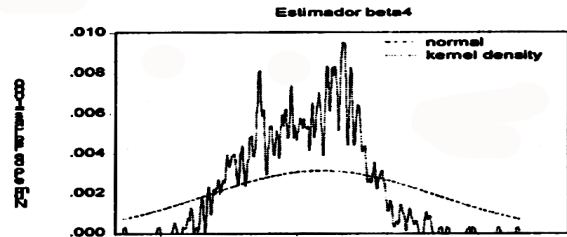
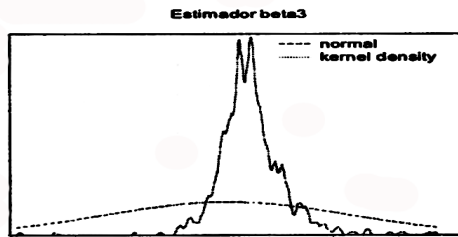
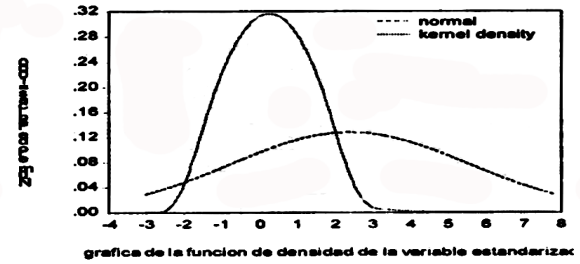
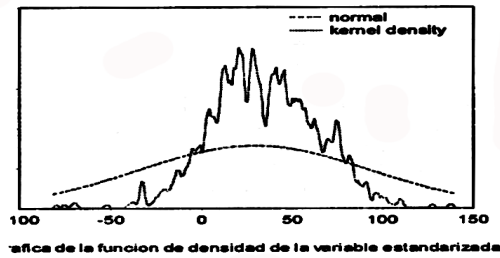
Regressión Robusta				
Método=Tuckey				
Variabes	Coefficientes	Erorr.Est.	T-stat	Prob
C	26.152	5.224	5.006	0.000
RECOP	1.697	0.208	8.172	0.000
PAUS	-20.442	3.424	-5.969	0.000
IDH	-22.763	9.794	-2.324	0.040
Suma de los residuos al cuadrado=40.11289226				
Error Estandar de la regresión=1.909614246				
Valor de Máximo Verosimilitud=-28.66143488				

El único método de Tukey arroja los mejores T-stat.

Nota.- Con este resultado hay que asumir con bastante cuidado, puesto que los valores outliers pueden ocasionar un efecto devastador, para comprobar, vamos a bootstroppear los coeficientes.

1.c Bootstrap de los Coeficientes

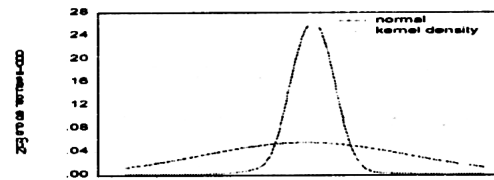
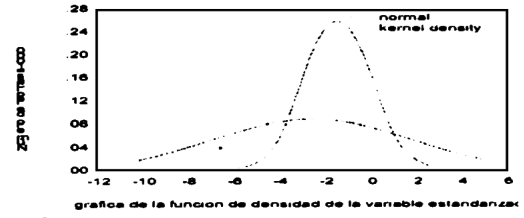
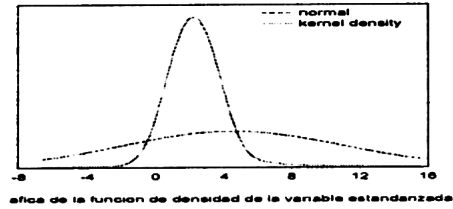
Resample Bootstrap				
Regresando=invs.				
Regresores=recop paus idh				
Número de réplicas=1000				
Estadísticos de los Coeficientes				
Coeficientes	cons	recop	paus	Idh
Mean	34.514	0.273	-11.334	-39.942
Median	33.157	0.257	-12.821	-36.502
Skewness	-0.056	2.850	0.404	0.015
Kurtosis	3.296	27.035	10.010	2.968
Test de Normalidad				
Shapiro-F	2.684	9.078	7.963	3.794
Shapiro-Wilk	2.588	11.46	9.679	4.002
Sesgo	-12.271	0.092	7.545	20.626



Definitivamente los valores outliers, generan inestabilidad de los estimadores de los coeficientes, ninguno es significativo, por esta razón vamos a excluir la variable recop, pero no podemos reducir los municipios, puesto que como son datos georeferenciales esto ocasionaría sesgos impredecibles, por esto excluimos recop.

1.d Bootstrap de los T-Student

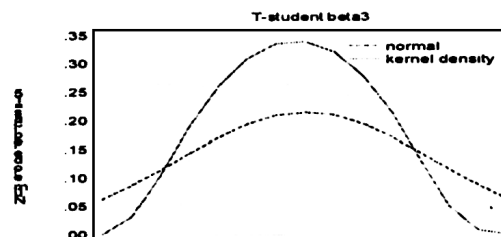
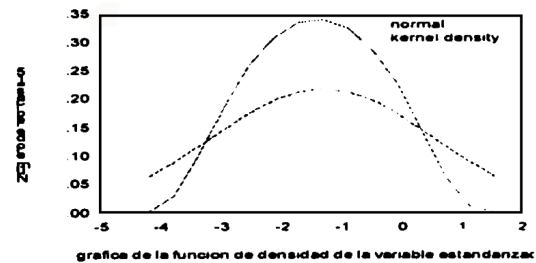
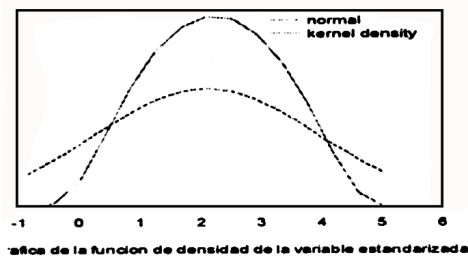
Regresando=invsm			
Regresores=recop paus			
Número de réplicas=1000			
Estadísticos de los T-Student			
Coeficientes	cons	paus	Idh
Mean	2.391	-1.531	-1.244
Median	2.250	-1.529	-1.241
Skewness	1.643	-0.510	-0.197
Kurtosis	12.976	6.406	13.763
Test de Normalidad			
Shapiro-Francia	8.7	7.063	8.629
Shapiro-Wilk	10.854	8.391	10.718
Sesgo	0.055	-0.045	-0.001



Los T-stat no mejoran, debido al coeficiente de curtosis, la constante es una excepción pero no interesa en este caso.

1.e Jackknife de los T-student

Pruebas Jackknife			
Número de obs.=15			
Jackknife de los T-student			
Estadísticos	C	PAUS	IDH
Media	2.236	-1.417	-1.190
Error.est	1.082	1.030	0.993
Sesgo	-1.408	0.971	0.741





Resulta que ninguno es significativo, otra alternativa es aplicar econometría espacial.

2. Econometría espacial

2.a. Matriz de Contigüidad

Uno de los elementos fundamentales de la econometría espacial es su forma de utilizar la información geográfica contenida en las observaciones de procesos que ocurren espacialmente. En este sentido, muchas de las técnicas desarrolladas en la geostatística y la estadística espacial han sido adaptadas para capturar los efectos espaciales en la estimación de modelos económicos.

Una de las formas más comunes de representar la ubicación geográfica de un conjunto de polígonos es a través de una Matriz de Conectividad o de Contigüidad. Esta es una matriz cuadrada que tiene el mismo número de filas (municipios) o columnas (municipios) que el número de polígonos independientes del mapa en estudio y que por convención se la denomina por W .

Los valores utilizados para representar vecindad son variados. La formulación más simple es una matriz de contigüidad binaria, es decir, los elementos de W serán igual a 1 si dos polígonos (municipios) son vecinos y cero en otro caso.

Esta matriz de contigüidad tiene ceros en la diagonal principal porque se asume que un polígono (municipio) no puede ser vecino consigo mismo. Adicionalmente, en la práctica esta matriz se estandariza por filas, es decir, se divide cada componente de la fila de la matriz por la suma de todos los elementos de esa fila de modo que la suma de cada fila es igual a uno, esta forma es muy útil para crear los rezagos espaciales.

2.b. Dependencia o autocorrelación espacial

Para poder aplicar, un modelo de econometría espacial, debe existir autocorrelación espacial, esto depende de la técnica que se aplique, entre ellas Moran, a la vez también depende como definamos la matriz de contigüidad (W), por simplicidad en el ejemplo usaremos, la contigüidad binaria luego estandarizada.

meth: 'moran'	istat: 1.3706
nobs: 15	imean: -0.0525
nvar: 3	ivar: 0.0185
morani: 0.1339	prob_0.1560



A un nivel de significación del 5%, no existe correlación espacial

Sólo con fines ilustrativos, vamos a suponer que efectivamente existe autocorrelación espacial, tendríamos los siguientes resultados:

Estimadores de los coeficientes:

Cte : 34.2688	Paus:-0.8285
Paus:-17.9821	Idh :2.0323
Idh >26.3695	R ² : 0.3198
Los T-student:	Rho:0.495
Cte:-1.3544	

Nota.- Observemos que los T-Student, mejora la variable IDH, si empleamos otras técnicas para determinar W, probablemente mejoren los T-Student, y esto es explicable, porque conociendo la ubicación geográfica de los 15 municipios, donde se aplicó no sea el más recomendable, pero el trabajo no es concluyente, sino más ilustrativo, mostrando las ventajas que posee frente a la econometría clásica.

Nota.- Para determinar la matriz W se debe contar con una información de primera, totalmente actualizada de los 15 municipios.

Nota.- Para dar validez a los estadísticos T-Student, es recomendable aplicar las pruebas de remuestreo a la econometría espacial, para esto se debe contar con un equipo de software de alta resolución.

Nota.- El coeficiente de autocorrelación espacial señala que existe una correlación de los residuos del 0.495 (autocorrelación de primer orden), es decir, si realmente se desea subsanar éste problema, se debe usar procesos ARMA en econometría espacial, esto complica ligeramente su tratamiento, es un poco menos simple.

Bibliografía:

- 1.- RYAN, D.L., VON HOHENBALKEN, B. y WEST, D.S. (1990), "An Econometric-Spatial Analysis of the Growth and Decline of Shopping Centers". Regional Science and Urban Economics, n°20 pp. 313-326.
- 2.- SPATIAL ECONOMETRICS USING MATLAB, JAMES P. LESAGE Department of Economics, University of Toledo, August, 1999
- 3.- METODO DE REMUESTREO EN SERIES TEMPORALES, DR. ANDRES ALONSO FERNÁNDEZ, Departamento de matemáticas, Universidad Autónoma de Madrid, 2002.

Software utilizado: MATLAB, SPACETATE y MACROEVIEWS





LA DESCOMPOSICIÓN ORTOGONAL DEL TEST χ^2 DE PEARSON APLICADO AL TEST DE FRIEDMAN

Dindo Valdez

Introducción

Anderson¹ (1959) analizó las tablas de contingencia de doble entrada con los totales marginales constantes desarrollando la descomposición del test χ^2 de Pearson en componentes ortogonales y Rayner²(1989) aplicó estas tablas y la prueba χ^2 a las pruebas de degustación de sabor de distintas variedades de un determinado producto. La descomposición del test χ^2 se puede aplicar como una prueba de homogeneidad entre filas (variedades), dicha descomposición ortogonal de este estadístico hace, esta prueba de hipótesis, más informativa y exhaustiva en este tipo de tablas donde su primer componente está relacionado con el estadístico Q_F del test de Friedman, y los subsecuentes componentes son extensiones a esta prueba.

Descomposición del test Chi cuadrado de Pearson

Primero, consideremos una tabla de contingencia de doble entrada de observaciones N_{ij} $i = 1,2,K, r$; $j = 1,2,K, c$, donde los totales de las filas $n_{i.}$ y los totales de las columnas $n_{.j}$ son valores conocidos (constantes), la clasificación de las columnas está ordenada y no así el de las filas. Para probar la hipótesis nula de homogeneidad entre las filas (tratamientos) es posible aplicar el conocido test χ^2 de Pearson, definido como:

$$\chi^2 = \sum_i \sum_j \frac{(N_{ij} - E_{ij})^2}{E_{ij}} \quad E_{ij} = n_{i.} n_{.j} / n, \text{ siendo } n \text{ el total de datos} \quad (1)$$

¹ Ver [1]

² Ver [2]



Sin embargo, es posible hacer mas potente a esta prueba descomponiéndola en $c-1$ componentes ortogonales para analizar efectos de escala, asimetría, curtosis, y momentos de mayor orden entre las filas de la tabla. Para definir tales componentes primero se otorga una puntuación a las categorías de las columnas $(1,2,K , c)$, luego se definen polinomios ortogonales denotados por $h_r(j)$, lo cuales son ortogonales respecto de p_j (la probabilidad marginal de la columna j).

$$\text{Tal que: } \sum h_r(j) \cdot h_s(j) \cdot p_j = \begin{cases} 0, & \text{si } r = s \\ 1, & \text{si } r \neq s \end{cases}$$

Emerson³ (1968) define estos polinomios $h_r(j)$ como:

$$h_r(j) = (jA_r + B_r)h_{r-1}(j) - C_r h_{r-2}(j), \quad j, r = 1, 2, K, c < n \quad (2)$$

donde A_r, B_r, C_r son constantes para un r dato tal que $h_r(j)$ llega a ser un polinomio de grado j , y los polinomios iniciales son: $h_{-1}(j)=0$ y $h_0(j)=1$. Tomando las probabilidades estimadas $p_j = n_j/n$, $j=1,2,K ,c$, proporcionales a las columnas marginales, se define al s -avo componente de la partición como:

$$V_{si} = \frac{1}{\sqrt{n_{i.}}} \sum_{j=1}^c N_{ij} \cdot h_s(j), \quad \text{para } s=1,K,c-1 \text{ e } i=1,K,r \quad (3)$$

De esta manera el estadístico de Pearson resulta particionado en $c-1$ componentes⁴,

$$\chi^2 = \sum_{s=1}^{c-1} \sum_{i=1}^r V_{si}^2 = \sum_{s=1}^{c-1} Q_s$$

Los componentes V_{si} satisfacen la condición lineal,

$$V_{s1}\sqrt{n_{1.}} + V_{s2}\sqrt{n_{2.}} + \dots + V_{sr}\sqrt{n_{r.}} = 0, \quad s = 1, K, c - 1$$

³ P.L. Emerson ha realizado varios trabajos comparando su método con el de Gram Schimdt, llegando a la conclusión que su método es más rápido y preciso.

⁴ Detalles de tal partición están dados en [3]



Bajo la hipótesis nula de homogeneidad entra las filas, cada componente V_{si} tiene distribución asintótica normal estándar, en consecuencia los componentes $Q_s = \sum V_{si}^2$ de la partición tienen distribución asintótica χ_{r-1}^2 . Tales componentes son llamados estadísticos score, y tiene similares propiedades a los estadísticos obtenidos por el método de la razón de verosimilitud⁵.

Aplicación de la descomposición al test de Friedman

Supongamos que r (>2) tratamientos son rankeados (clasificados) cada uno por n personas (jueces), entonces es posible realizar un cuadro de contingencia de doble entrada con r filas (tratamientos) y r columnas (rangos) como se muestra en la tabla siguiente:

Tratamiento	Rango				n_i
	1	2	\dots	r	
T1	N_{11}	N_{12}	\dots	N_{1r}	n
T2	N_{21}	N_{22}	\dots	N_{2r}	n
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
Tr	N_{r1}	N_{r2}	\dots	N_{rr}	n
n.j	n	n	\dots	n	$n \times r$

N_{ij} es el número de veces que el tratamiento i recibe el rango j

r es el número de muestras aleatorias

n es el tamaño de las muestras

Lo anterior queda claro ya que cada tratamiento y cada rango es asignado n veces. Aplicando la definición de la ecuación (1), el estadístico χ^2 de Pearson para este tipo de tablas resulta:

$$\chi^2 = \frac{r}{n} \sum_{i=1}^r \sum_{j=1}^r \left(N_{ij} - \frac{n}{r} \right)^2 = \sum_{s=1}^{r-1} \sum_{j=1}^r V_{sj}^2 = \sum_{s=1}^{r-1} Q_s$$

⁵ Ver [2], sección 3.4



Con
$$V_{si} = \frac{\sqrt{r-1}}{r} \left(\sum_{j=1}^r h_s(j) \frac{N_{ij} - n/r}{\sqrt{n/r}} \right), \quad s = 1, K, r-1; i = 1, K, r.$$

El primer componente de esta descomposición resulta ser el estadístico del test de Friedman, tal que:

$$Q_1 = \sum_{j=1}^r V_{1i}^2 = \frac{12}{rn(r+1)} \sum_{j=1}^r \left(R_i - \frac{n(r+1)}{2} \right)^2 = Q_F$$

Donde R_i es la suma de los rangos para el i -ésimo tratamiento $R_i = \sum_{j=1}^r N_{ij}$.

Ya que los componentes V_{si} envuelven momentos de orden s , los componentes Q_s detectan los efectos del s -avo momento bajo la hipótesis nula y por esta razón, se puede considerar al test de Friedman como un test de posición ya que envuelve polinomios de primer grado en su cálculo, de esta manera los subsecuentes componentes Q_s para $s=2, K, r-1$, se consideran las extensiones al test de Friedman, puesto que éstos miden adicionalmente los efectos en la escala, asimetría, curtosis, y momentos de mayor orden respectivamente. Para mostrar la utilidad de tal descomposición se realizó un estudio en un grupo de personas universitarias entre los 18 y 25 años de edad con el objeto de medir la preferencia sobre cinco sabores de yogurt bebible (coco, frutilla, uva, limón y chicle) donde cada persona clasificó a cada sabor del 1º al 5º en el orden de preferencia llegando a realizar 95 observaciones (secuencias de preferencia) de manera que cada sabor recibió un total de 95 valoraciones (entre 1s, 2s, etc.). Es así que con el conjunto de observaciones se elaboró una tabla de contingencia, donde las filas corresponden a los sabores (tratamientos) y las columnas a los rangos (bloques).

Tratamiento (sabor)	Escala de Preferencia					Total
	1	2	3	4	5	
Coco	12	18	20	19	26	95
Frutilla	8	17	24	22	24	95
Uva	13	16	18	21	27	95
Limón	15	31	21	16	13	95
Chicle	47	13	12	17	6	95
Total	95	95	95	95	95	475

La escala 1 es la menos gustada y 5 es la mayor preferencia



La hipótesis a docimar es que cada variedad de yogurt tiene el mismo nivel de aceptación (homogeneidad), al aplicar la prueba χ^2 de homogeneidad entre filas se obtuvo un valor de 87,47 con 16 grados de libertad y una significancia de 7.2E-12; por tanto podemos concluir que la distribución en las preferencias de los 5 sabores de yogurt no es homogénea al 5% de significación. En cambio si utilizamos los datos brutos y aplicamos el test de Friedman para probar la hipótesis que los 5 sabores tienen similares resultados se obtiene $Q_F = 41.364$ con una significancia igual a 0. Consecuentemente rechazamos la hipótesis nula y concluimos que los cinco sabores no produjeron idénticos resultados respecto a sus rangos asignados.

Puesto que existen diferencias entre los tratamientos, es importante analizar la influencia de los componentes ortogonales del test χ^2 para aclarar dónde radican estas diferencias exactamente, para lo cual calculamos los cuatro componentes. Aplicando la ecuación (2) los polinomios ortogonales de Emerson resultan:

J	1	2	3	4	5
$h_1(j)$	1,00	1,00	1,00	1,00	1,00
$h_2(j)$	-1,41	-0,70	0,00	0,70	1,41
$h_3(j)$	1,19	-0,59	-1,19	-0,59	1,19
$h_4(j)$	-0,70	1,41	0,00	-1,41	0,70
$h_5(j)$	0,26	-1,06	1,60	-1,06	0,26

Tales polinomios son ortonormales bajo la H_0 con $p_j = 0.2$ y el valor esperado de la variable N_{ij} es 19 en todos los casos.

Efectos de posición y escala. Basados en los anteriores resultados, la media o efecto lineal para el i -ésimo sabor está relacionado con los elementos V_{1i} quedando definida como:

$$V_{1i} = \frac{2}{5} \sum_{j=1}^5 h_1(j) \frac{N_{ij} - 19}{\sqrt{19}}, \quad i = 1, \dots, 5$$



De igual manera la variación llamada también efecto de dispersión o escala queda definido por:

$$V_{2i} = \frac{2}{5} \sum_{j=1}^5 h_2(j) \frac{N_{ij} - 19}{\sqrt{19}}, \quad i = 1, \dots, 5$$

Los subsecuentes componentes son calculados aplicando la ecuación (3) y se presentan en la tabla:

Variedad	Efecto			
	Lineal	Cuadrático	Asimetría	Curtosis
	V_{1i}	V_{2i}	V_{3i}	V_{4i}
Coco	1,88	-0,05	0,77	0,24
Frutilla	2,40	-1,26	0,38	0,49
Uva	2,14	0,38	0,25	0,00
Limón	-1,36	-1,91	1,75	-0,85
Chicle	-5,06	2,85	-3,17	0,12

Fuente	Grados de Libertad	Suma de Cuadrados	Significancia
Posición Q_1	4,00	41,36	0,00
Escala Q_2	4,00	13,55	0,01
Residuo $Q_3 + Q_4$	8,00	15,05	0,06
Total χ^2	16,00	69,96	

Esta tabla tiene un estilo parecido al clásico análisis de varianza, basado en los componentes ortogonales del test χ^2 de Pearson. La tabla muestra que la suma de cuadrados de los efectos de ubicación es el estadístico Q_F del test de Friedman para datos ordenados por rangos. El componente Q_2 se considera como un nuevo estadístico, el cuál mide los efectos de escala. Esta tabla confirma una vez más que existen diferencias en cuanto a la posición entre las distintas variables de yogurt también muestra algo adicional y revelador, la existencia de diferencias significativas en la escala, tal diferencia no podría haber sido revelada si no se utilizaba la descomposición del test



Chi cuadrado. El residuo mide los efectos de asimetría y curtosis simultáneamente y éste no es significativo al 5%.

En estos estudios de comparación, tales datos usualmente son analizados mediante el tradicional análisis de varianza, pero de una manera distinta al análisis de rangos, ya que esta prueba puede verse afectada por valores atípicos que llevarían a tener una heterogeneidad en la varianza y por reasumir la normalidad de los datos. El análisis realizado es más robusto que el ANOVA, en el sentido que ninguno de los dos supuestos: homogeneidad en la varianza y normalidad de los datos, han sido asumidos. También este estudio ha demostrado ser más exhaustivo e informativo que la simple aplicación del test de Friedman.

Bibliografía

[1] P. Anderson, Revista Electrónica Biometrics.

[2] J.C. Rayner, Smooth Test of Goodness of Fit.

[3] Dindo Valdez, Pruebas Suavizadas de Bondad de Ajuste con Aplicación en Pruebas no Paramétricas.

----- O -----

Si amas algo

Según un Estadístico: “Si amas a alguien, déjalo ir; si el te quiere, las probabilidades que vuelva son de un 86.5 por ciento; si no te quiere, tus relaciones con él caen en el campo de lo improbable, con un margen de error de un 3 por ciento”.





PANORAMA GENERAL DE LOS MÉTODOS MULTIVARIADOS

Aníbal Angulo

Los métodos multivariados son extraordinariamente útiles puesto que ayudan a los investigadores a dar sentido a grandes conjuntos de datos, complicados y complejos, que constan de una cantidad de variables medidas en grandes números de unidades experimentales. La importancia y la utilidad de los métodos multivariados aumentan cuando se incrementa el número de variables que se están midiendo y el número de unidades experimentales que se están evaluando.

A menudo, el objetivo primario de los análisis multivariados es resumir grandes cantidades de datos en relativamente pocos parámetros. El tema subyacente de muchas técnicas multivariadas es la simplificación.

El interés de los análisis multivariados es encontrar relaciones entre 1) las variables respuesta, 2) las unidades experimentales, y 3) tanto las variables respuesta como las unidades experimentales. Se podría decir que existen relaciones entre las variables respuesta cuando, en realidad, algunas de las variables están midiendo una cantidad común.

Podrían existir relaciones entre las unidades experimentales si algunas de ellas son semejantes entre sí. Muchas técnicas multivariadas tienden a ser de naturaleza exploratoria en lugar de confirmatoria. Es decir, muchos métodos multivariados tienden a motivar hipótesis y no probarlas. Considere una situación en la cual un investigador tiene 50 variables medidas sobre más de 3000 unidades experimentales. Los métodos estadísticos tradicionales suelen exigir que un investigador establezca algunas hipótesis, reúna algunos datos y, a continuación, use esos datos para comprobar o rechazar esas hipótesis. Una situación alternativa que se da frecuentemente es un caso en el cual un investigador dispone de una gran cantidad de datos y se pregunta si pudiera haber una información valiosa en ellos. Las técnicas multivariadas suelen ser útiles para examinar los datos en un intento por saber si hay información que valga la pena y sea valiosa en esos datos.



Una distinción fundamental entre los métodos multivariados es que algunos se clasifican como “técnicas dirigidas por las variables”, en tanto que otros se clasifican como “técnicas dirigidas por los individuos”.

Las técnicas dirigidas por las variables son aquellas que se enfocan primordialmente en las relaciones que podrían existir entre las variables respuesta que se están midiendo.

Las técnicas dirigidas por los individuos son la que se interesan principalmente en las relaciones que podrían existir entre las unidades experimentales o individuos que se están midiendo, o en ambos. Algunos ejemplos de este tipo de técnica se encuentran en el análisis discriminante, el análisis de agrupaciones y el análisis multivariado de la varianza (manova): (Multivariate Analysis of Variance)

Muchos métodos multivariados ayudan a los investigadores a crear nuevas variables que tengan propiedades deseables, entre éstas indicamos las siguientes técnicas.

Análisis de Componentes Principales (PCA).

Cuando un investigador está empezando a pensar acerca del análisis de un nuevo conjunto de datos, debe considerar varias preguntas acerca de ellos. Las preguntas importantes incluyen: 1) ¿Existen algunos aspectos en los datos que resultan extraños? 2) ¿Se puede suponer que los datos están distribuidos normalmente? 3) ¿Hay algunas anomalías en los datos? 4) ¿Existen datos outliers⁶(atípicos)?

La razón más importante para realizar un Análisis de Componentes Principales es usarlo como herramienta para cribar los datos de variables múltiples. Se pueden crear nuevas variables, llamadas calificaciones de componentes principales, que se pueden usar como entrada para programas de trazado de gráficas y situación de datos y, con frecuencia, un examen de las presentaciones gráficas resultantes revelarán las anomalías en los datos que se está planeando analizar. Además, se pueden analizar por separado las mediciones de los componentes principales para ver si se cumplen las

⁶ Las unidades experimentales cuyas variables medidas parecen incoherentes o son atípicas con relación a las mediciones realizadas en las otras unidades experimentales suelen llamarse datos outliers.



hipótesis relativas a la distribución, como la normalidad de las variables y la independencia de las unidades experimentales. A menudo, se requieren esas hipótesis para que sean válidos ciertos tipos de análisis estadísticos.

El Análisis de Componentes Principales suele ser bastante útil para los investigadores que desean realizar la división en subgrupos de las unidades experimentales, de modo que unidades experimentales similares pertenezcan al mismo subgrupo. En este caso, se pueden usar las calificaciones de los componentes principales como entrada para los programas de agrupación, lo que suele incrementar la eficacia de estos programas, reduciendo al mismo tiempo el costo de su uso. Además, se recomienda usar las mediciones de los componentes principales para ayudar a validar los resultados de los programas de agrupación.

Análisis por Factores

El análisis por factores (FA: Factor Analysis) es una técnica que se emplea frecuentemente para crear nuevas variables que resuman toda la información de la que podría disponerse en las variables originales. El análisis por factores también se usa para estudiar las relaciones que podrían existir entre las variables medidas en un conjunto de datos. Un objetivo básico de esta técnica es determinar si las variables respuesta exhiben patrones de relaciones entre sí, tales que esas variables se pueden dividir en subconjuntos de modo que las variables en un subconjunto estén fuertemente correlacionadas con cada una de las otras y que las variables en subconjuntos diferentes tengan bajas correlaciones entre sí. Por tanto, el FA se usa con frecuencia para estudiar la estructura de correlación de las variables en un conjunto de datos. Una semejanza entre el FA y el PCA es que este último también se puede usar para crear nuevas variables que no están correlacionadas entre sí. Esas variables se llaman clasificación de factores.

Una ventaja que parece tener el FA sobre PCA, es que cuando se están creando nuevas variables, generalmente, las nuevas variables creadas por el FA son mucho más fáciles de interpretar que las creadas por el PCA. Si un investigador desea crear un conjunto más pequeño de nuevas variables que se puedan interpretar y que resuman la mayoría de la información existente en las variables medidas entonces se recomienda usar el Análisis por Factores.



Análisis Discriminante

El análisis discriminante (DA Discriminant Analysis) se usa principalmente para clasificar individuos o unidades experimentales en dos o más poblaciones definidas de manera única. Para desarrollar una regla discriminante que clasifique las unidades experimentales en una de varias categorías posibles, el investigador debe tener una muestra aleatoria de unidades experimentales de cada grupo posible de clasificación. Entonces, el DA proporciona los métodos que permitirán a los investigadores establecer reglas que se puedan emplear para clasificar otras unidades experimentales en uno de los grupos de clasificación.

Análisis Discriminante Canónico

El análisis discriminante canónico (CDA: Canonical Discriminant Analysis) es un procedimiento con el que se crean nuevas variables que contienen toda la información útil para la discriminación de la que se dispone en las variables originales. A menudo, estas nuevas variables conducen a reglas más sencillas para clasificar las unidades experimentales en los diferentes grupos.

Regresión Logística

Con frecuencia se usa la regresión logística para modelar la probabilidad de que una unidad experimental caiga en un grupo particular, con base en la información medida en la propia unidad. Estos modelos se pueden usar con fines de discriminación. En el caso de las tarjetas de crédito Bancario, se puede modelar la probabilidad de que un individuo con ciertas características demográficas sea un buen riesgo de crédito; después de desarrollar este modelo, se puede usar para predecir la probabilidad de que un nuevo solicitante caiga en un cierto grupo de riesgo. Los individuos cuya probabilidad pronosticada para el grupo de “buen riesgo” sea mayor que 0.5 se determinan como buenos riesgos de crédito.

Análisis por agrupación

Suponga que un arqueólogo descubre un gran escondite de fragmentos de alfarería y toma pequeñas muestras representativas de cada fragmento, y que cada muestra se puede analizar y que se pueden averiguar las cantidades relativas de



diferentes elementos químicos, como zinc, magnesio, hierro, etc. El Arqueólogo quiere separa los fragmentos en montones distintos, de modo que los que queden en cada uno de esos montones provenga de la misma pieza de alfarería. Resulta claro que esto puede ser una tarea difícil porque el arqueólogo no sabe cuántos montones de fragmentos resultarán, cuantos fragmentos quedaran en cada montón o si algunos fragmentos en realidad pertenezcan al mismo montón. El análisis por agrupación es un método multivariado que puede ayudar a resolver este problema.

El análisis por agrupación (CA: Cluster Analysis) es semejante al discriminante en el sentido de que se usan para clasificar individuos o unidades experimentales en subgrupos definidos de manera única. Este análisis se puede emplear cuando el investigador cuenta con una muestra aleatoria previamente obtenida de cada uno de los subgrupos definidos de manera única. El análisis por agrupación trata los problemas de clasificación cuando no se sabe de antemano de cuáles subgrupos se originan las observaciones.

Análisis Multivariado de la varianza

El análisis multivariado de la varianza (MANOVA) es una generalización del análisis univariado de la varianza (ANOVA: Analysis of Variance), una técnica usada para comparar las medias de varias poblaciones en una sola variable medida.

Cuando se miden varias variables en cada unidad experimental, podría producirse un ANOVA sobre cada variable medida, usando una variable a la vez; por ejemplo si se miden 30 variables, un investigador podría producir 30 análisis separados, uno para cada variable. Sin embargo, esto no es inteligente, por desgracia la mayoría de los experimentos se están analizando con la aplicación de análisis de una variable a la vez.

Los estadísticos promueven dos objeciones principales para los análisis separados de cada variable medida. Una objeción es que las poblaciones que se están comparando pueden ser diferentes en alguna variable, pero no en otras. A menudo el investigador se encuentra confuso en cuanto a cuales poblaciones son diferentes en realidad y cuáles son semejantes. Los análisis multivariados de la varianza pueden ayudarlos a comparar varias poblaciones al considerar, simultáneamente todas las variables medidas.

Una segunda objeción es que se tiene protección inadecuada contra errores tipo I cuando se realizan análisis de una variable a la vez. Recuerde lo visto en los cursos



introdutorios de estadística, lo que ocurre en error tipo I siempre que se rechaza una hipótesis verdadera. Cuanto más variables analice un investigador, mayor es la probabilidad de que por lo menos una de las variables que se están analizando y de hallar por lo menos uno de estos análisis estadísticamente significativos, es decir, tiende a uno.

Es evidente que el gran riesgo de cometer errores de tipo I deba inquietar a los experimentadores. El investigador debe estar confiado cuando afirma que dos o más poblaciones tienen medias diferentes con respecto a una variable medida y debe confiar que conduzcan a análisis similares sobre conjuntos semejantes de datos.

Debe realizarse un MANOVA siempre que se están comparando entre sí dos o más poblaciones diferentes sobre un número grande de variables respuesta.

Si un MANOVA muestra diferencias significativas entre las medias de las poblaciones, entonces el investigador puede confiar en que verdaderamente existen diferencias reales. En este caso, resulta razonable considerar el análisis de una variable a la vez para detectar dónde ocurren en realidad las diferencias.

Si el ANOVA no revela diferencias significativas entre las medias de las poblaciones, entonces el investigador debe tener precaución extrema al interpretar los análisis de una variable a la vez. Esos análisis pueden identificar como positivos falsos.

Análisis de Variables Canónicas

El análisis de variables canónicas (CVA: Canonical Variates Analysis) es un método en el que se crean nuevas variables con conjunción con los análisis multivariados de la varianza. Estas nuevas variables son útiles porque ayudan a los investigadores a determinar dónde ocurren las diferencias importantes entre las medias de las poblaciones, cuando se están comparando poblaciones sobre muchas variables diferentes mediante el uso simultáneo de todas las variables medidas. En ocasiones, las variables canónicas pueden surgir diferencias importantes que, del contrario podrían pasarse por alto.

Análisis de Correlación Canónica.

El análisis de correlación canónica es una generalización de la correlación múltiple en los problemas de regresión. Se requiere que las variables respuesta se dividan en dos grupos. La asignación de las variables en estos dos grupos siempre debe motivarse por la naturaleza de las variables respuesta y nunca por una inspección de los datos. Por



ejemplo, una asignación legítima sería aquella en la que las variables en uno de los grupos sean fáciles de obtener y no caras para medirse, mientras que las que se encuentren en el otro grupo sean difíciles de obtener y caras para medirse. Una cuestión básica que se espera responder con el análisis de correlación canónica es si se pueden usar las variables que se encuentran en uno de los grupos para predecir las variables en el otro. Cuando se puede, entonces este análisis intenta resumir las relaciones entre los dos conjuntos de variables, mediante la creación de nuevas variables a partir de cada uno de los dos grupos de variables originales.

Bibliografía.

Alvarez Cáceres, R (1995) Estadística Multivariante y no Paramétrica Aplicada a las Ciencias de Salud.

Anderberg M. J. Clusters Analysis for Applications N. 4. Academic Press 1973

Dallas E. Jhonson Métodos Multivariados Aplicados al Análisis de Datos.

M. Ángeles Cea, D Ancona Análisis Multivariante Teoría y Práctica en el Investigación Social

----- O -----

Estadístico

Es alguien que es bueno con los números pero carece de personalidad suficiente como para ser Contador.

Un Estadístico podría meter su cabeza en un horno y sus pies en hielo, y decir que en promedio se encuentra bien.

Tres Estadísticos salen de cacería. En esto ven a un ciervo pastando, tranquilamente. El primer estadístico apunta, dispara, y la bala pasa dos metros y diez centímetros a la derecha del ciervo. El segundo estadístico apunta, dispara y la bala pasa dos metros y diez centímetros a la izquierda del ¡Bien, le hemos dado!





SISTEMAS DE INVENTARIOS MRP

Nicolás Chávez Quisbert

Introducción

Las técnicas MRP (Materials Requirement Planning - Planificación de las Necesidades de Materiales) son una solución relativamente nueva a un problema clásico en producción, el controlar y coordinar los materiales para que se hallen a punto cuando son precisos y al propio tiempo sin necesidad de tener un excesivo inventario.

La gran cantidad de datos que hay que manejar y la enorme complejidad de las interrelaciones entre los distintos componentes trajeron consigo que, antes de los años sesenta, no existiera forma satisfactoria de resolver el problema mencionado, lo que propició que las empresas siguiesen, utilizando los stocks de seguridad y las técnicas clásicas, así como métodos informales, con el objeto de intentar evitar en lo posible problemas en el cumplimiento de la programación debido a falta de stocks, por desgracia, no siempre conseguían sus objetivos, aunque casi siempre incurrían en elevados costos de posesión. Se tuvo que esperar a los años sesenta para que la aparición del ordenador abra las puertas al MRP, siendo ésta más que una simple técnica de gestión de inventarios.

Cabe señalar que los sistemas MRP no constituyen un cuerpo de conocimientos cerrado, sino que han estado evolucionando en forma continua. Inicialmente se usaba el MRP para programar inventarios y producción (Sistemas MRP I) luego se fue incluyendo la planificación de capacidad de recursos (Sistemas MRP II), y por último una vez desarrollados estos sistemas, se amplía el sistema a la planificación y control de otros departamentos de la empresa (Sistemas MRP III).

Estructura y soporte informático de los sistemas MRP

Los sistemas MRP han sido desarrollados mediante diferentes formas de estructura modular. Este carácter modular se refiere tanto a la realización de procesos como al software que soporta el funcionamiento de estos sistemas.



MRP I

El Sistema de MRP I o Planificación de Necesidades de Materiales , trata de determinar óptimamente, qué se debe aprovisionar y/o fabricar, en qué cantidad, y en qué momento para cumplir con los compromisos adquiridos.

El Objetivo del MRP I es brindar un enfoque más efectivo, sensible y disciplinado a determinar los requerimientos de materiales de la empresa.

	Técnicas Clásicas	M.R.P
- Tipo de demanda	Independiente (aleatoria).	Dependencia (predeterminada).
- Determinación de la demanda.	Previsión estadística en base a la demanda histórica.	Explosión de las necesidades en base al Plan Maestro de Producción.
- Tipo de artículos	Finales y piezas de repuesto.	Partes y componentes.
- Base de los pedidos	Reposición	Necesidades
- Stocks de seguridad	Necesario para paliar la aleatoriedad de la demanda.	Tiende a desaparecer salvo en los productos finales.
- Objetivos directos	Satisfacción del cliente.	Satisfacción de las necesidades de producción.

El sistema MRP comprende la información obtenida de al menos tres fuentes o ficheros de información principales que a su vez suelen ser generados por otros subsistemas específicos, pudiendo concebirse como un proceso cuyas entradas son:

- El plan maestro de producción.
- El estado del inventario.
- La lista de materiales.

Así pues, la explosión de las necesidades de fabricación no es más que el proceso por el que las demandas externas correspondientes a los productos finales son traducidas en órdenes concretas de fabricación y aprovisionamiento para cada uno de los ítems que intervienen en el proceso productivo.

El proceso de explosión de necesidades consiste en realizar para cada ítem, empezando por los de nivel superior, los siguientes pasos:

1. Cálculo de la demanda dependiente debida a órdenes de fabricación de todos aquellos ítems que requieren directamente dicho componente.



2. Determinación de las necesidades brutas por adicción de la demanda independiente (Plan Maestro de Producción) a la demanda dependiente (calculada en el paso anterior).
3. Cálculo de las necesidades netas por sustracción del inventario disponible (y recepciones programadas de órdenes de fabricación previas) de las necesidades brutas.
4. Cálculo del tamaño de las órdenes de fabricación (aprovisionamiento) necesarias y de calado de las mismas un número de períodos igual al plazo de fabricación o aprovisionamiento.

En la programación de un sistema MRP, se debe tener en cuenta la información relevante de cada ítem y que básicamente es:

- Referencia o código de identificación.
- Nivel más bajo en que se encuentra el ítem en la estructura de fabricación.
- Política de determinación del tamaño del lote.
- Plazo de suministro (fabricación o aprovisionamiento).
- Inventario disponible inicial.
- Unidades comprometidas (a descontar del inventario inicial).
- Stock de seguridad (si existe).

Asimismo, se deberán realizar como mínimo los siguientes cálculos durante la explosión de necesidades:

NECESIDADES BRUTAS = DEMANDA INDEPENDIENTE + DEMANDA DEPENDIENTE

Recepciones programadas correspondientes a las órdenes en curso cuya recepción es conocida en el instante inicial.

INVENTARIO DISPONIBLE = INVENTARIO DISPONIBLE AL FINAL DEL PERÍODO ANTERIOR + RECEPCIONES PROGRAMADAS + RECEPCIONES DE ÓRDENES DE PRODUCCIÓN-NECESIDADES BRUTAS.

NECESIDADES NETAS = STOCK DE SEGURIDAD + NECESIDADES BRUTAS - INVENTARIO ANTERIOR - RECEPCIONES PROGRAMADAS.



De las aplicaciones realizadas se deduce que dicho sistema es el más óptimo ya que reduce costos, tiempos y todo tipo de recursos. Además, permite adoptar estrategias de mejor toma de decisiones.

Bibliografía

BERTOGLIO OSCAR (1992), INTRODUCCIÓN A LA TEORIA GENERAL DE SISTEMAS, LIMUSA, MEXICO

COHEN DANIEL (1993), SISTEMAS DE INFORMACIÓN PARA LA TOMA DE DECISIONES, McGRAW HILL, MEXICO

FLORES F. LUIS H. (2001), SISTEMA ESTRATÉGICO DE PLANEACIÓN DE REQUERIMIENTO DE MATERIALES PARA UN PROCESO PRODUCTIVO “CASO FABRICA ALBUS”, EMI, LA PAZ

HEIZER JAY (1997), DIRECCIÓN DE LA PRODUCCIÓN DECISIONES TÉCNICAS, PRENTICE HALL, ESPAÑA

HODSON WILLIAM, (1996), MANUAL DEL INGENIERO INDUSTRIAL, McGRAW HILL, MEXICO

MATHUR K. SOLOW D. (1996) , INVESTIGACIÓN DE OPERACIONES, PRENTICE HALL, MÉXICO

MOSKOWITZ H. WRIGHT G. (1982), INVESTIGACION DE OPERACIONES, PRENTICE HALL, MEXICO

TORREY MARCO ANTONIO, (1976), TÉCNICAS INDUSTRIALES DE PRODUCCIÓN, DOLMENZA, PERU

----- 0 -----

La Parábola



Jesús a sus discípulos: En verdad os digo, $Y = X^2$

Los discípulos comentan entre si, y dice Pedro:

Maestro, no entendemos

Es una Parábola Bruto!



ALTERNATIVAS

Raúl Delgado Álvarez

El objetivo del presente artículo, es el de reforzar el aspecto Psicopedagógico de la solución de ejercicios, la búsqueda de otra manera de resolver un ejercicio o la demostración de un Teorema afianza la destreza en un estudiante, por ello se propone un problema y se exponen dos soluciones y se invita a los lectores hacer llegar otra u otras soluciones.

Problema

Hallar el momento k-ésimo de la distribución T-Student, a partir de dicha expresión hallar el coeficiente de deformación vertical o Curtosis, además evaluar el mismo si “n” (grados de libertad) crece indefinidamente. ¿Que conclusión se puede obtener?

Primera Solución

$E[T^k] = 0$, para k impar, dado que la distribución t-Student es simétrica respecto al eje vertical, luego sólo existe para k par, es decir, $E[T^{2k}]$

Dada la variable $T = \frac{Z}{\sqrt{V/n}}$, como $Z \sim N(0,1)$ y $V \sim \chi_n^2$ siendo además variables aleatorias independientes, entonces:

$$E[T^{2k}] = E\left[\frac{\sqrt{n}Z}{\sqrt{V}}\right]^{2k} = E[n^k Z^{2k} V^{-k}]$$

$$E[T^{2k}] = n^k E[Z^{2k}]E[V^{-k}] \quad (1)$$

Hallando; $E[Z^{2k}]$:

$$E[Z^{2k}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} Z^{2k} e^{-\frac{1}{2}Z^2} dz, \text{ la expresión subintegral es función par, luego:}$$

$$2 \int_0^{\infty} \frac{1}{\sqrt{2\pi}} Z^{2k} e^{-\frac{1}{2}Z^2} dz, \text{ haciendo el cambio de variable } u = \frac{Z^2}{2} \quad du = zdz$$



$$\frac{2}{\sqrt{2\pi}} \int_0^{\infty} (\sqrt{2u})^{2k-1} e^{-u} du = \frac{2^k}{\sqrt{\pi}} \int_0^{\infty} u^{(k+\frac{1}{2})-1} e^{-u} du$$

$$E[Z^{2k}] = \frac{2^k \Gamma(k+\frac{1}{2})}{\sqrt{\pi}} \quad (2)$$

Hallando: $E[V^{-k}]$:

$$E[V^{-k}] = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \int_0^{\infty} v^{-k} v^{\frac{n}{2}-1} e^{-\frac{v}{2}} dv, \text{ Sea el cambio de variable; } h = \frac{v}{2} V, dh = \frac{1}{2} dv$$

los límites de integración no se modifican, luego:

$$= \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} 2^{\frac{n}{2}-k} \int_0^{\infty} h^{(\frac{n}{2}-k)-1} e^{-h} dh$$

$$E[V^{-k}] = \frac{2^{-k}}{\Gamma(\frac{n}{2})} \Gamma(\frac{n}{2}-k) \quad (3)$$

De (2) y (3) en (1), se tiene: $E[T^{2k}] = n^k E[Z^{2k}]E[V^{-k}]$

$$= n^k \frac{2^k \Gamma(k+\frac{1}{2})}{\sqrt{\pi}} \frac{2^{-k}}{\Gamma(\frac{n}{2})} \Gamma(\frac{n}{2}-k)$$

$$E[T^{2k}] = n^k \frac{\Gamma(k+\frac{1}{2})\Gamma(\frac{n}{2}-k)}{\sqrt{\pi}\Gamma(\frac{n}{2})} \quad (4)$$

Segunda Solución:

Hallando directamente el momento k-ésimo:

$E[T^k] = 0$, cuando k es impar por simetría, luego sólo existe para k par, es decir:



$$E [T^{2k}] = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \int_{-\infty}^{\infty} t^{2k} (1 + \frac{t^2}{n}) dt, \text{ siendo la expresión subintegral función par:}$$

$$= 2 \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \int_0^{\infty} t^{2k} (1 + \frac{t^2}{n})^{-\frac{n+1}{2}} dt, \text{ Haciendo el cambio de variable: } Z = \frac{t}{\sqrt{n}}$$

$$= 2 \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \int_0^{\infty} (\sqrt{n} z)^{2k} (1 + z^2)^{-\frac{n+1}{2}} \sqrt{n} dz$$

$$= 2 \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} n^{k+\frac{1}{2}} \int_0^{\infty} z^{2k} (1 + z^2)^{-\frac{n+1}{2}} dz$$

Como la expresión subintegral, contiene suma de cuadrados, haciendo: $z = \text{tag } \theta$, se tiene: Si $z = 0$, entonces, $\text{tag } \theta = 0$, con lo cual $\theta = 0$. Cuando $z \rightarrow \infty$, $\text{tag } \theta \rightarrow \infty$, $\theta = \frac{\pi}{2}$

También: $dz = \sec^2 \theta d\theta$, con lo cuál:

$$= 2 \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} n^{k+\frac{1}{2}} \int_0^{\frac{\pi}{2}} (\text{tag } \theta)^{2k} (1 + \text{tag}^2 \theta)^{-\frac{n+1}{2}} \sec^2 \theta d\theta, \text{ utilizando identidades}$$

trigonométricas:

$$= 2 \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} n^{k+\frac{1}{2}} \int_0^{\frac{\pi}{2}} (\text{tag } \theta)^{2k} (\sec^2 \theta)^{\frac{1-n}{2}} d\theta,$$

$$= 2 \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} n^{k+\frac{1}{2}} \int_0^{\frac{\pi}{2}} (\text{sen } \theta)^{2k} (\cos \theta)^{n-1-2k} d\theta, \text{ escribiendo los exponentes}$$

como: $2k = 2(\frac{2k+1}{2}) - 1$, $n-1-2k = 2(\frac{n-2k}{2}) - 1$



$$= 2 \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} n^k \int_0^{\frac{\pi}{2}} (\text{sen } \theta)^{2\frac{(2k+1)}{2}-1} (\text{cos } \theta)^{2(\frac{n-2k}{2})-1} d\theta, \text{ considerando una de las}$$

expresiones de la función Beta: $\beta(p, q) = 2 \int_0^{\frac{\pi}{2}} (\text{sen } \theta)^{2p-1} (\text{cos } \theta)^{2q-1} d\theta$ además:

$$\beta(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

$$= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} n^k \beta\left(\frac{2k+1}{2}; \frac{n-2k}{2}\right)$$

$$= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} n^k \frac{\Gamma(k+\frac{1}{2})\Gamma(\frac{n}{2}-k)}{\Gamma(\frac{n+1}{2})} \text{ Simplificando:}$$

$$E[T^{2k}] = n^k \frac{\Gamma(k+\frac{1}{2})\Gamma(\frac{n}{2}-k)}{\sqrt{\pi}\Gamma(\frac{n}{2})} \text{ como la ecuación (4)}$$

El momento de k-ésimo orden es

$$E[T^{2k}] = n^k \frac{\Gamma(k+\frac{1}{2})\Gamma(\frac{n}{2}-k)}{\sqrt{\pi}\Gamma(\frac{n}{2})}, k=1,2,\dots$$

$$E[T^2] = \frac{n\Gamma(1+\frac{1}{2})\Gamma(\frac{n}{2}-1)}{\sqrt{\pi}\Gamma(\frac{n}{2})} = \frac{n(\frac{1}{2})\Gamma(\frac{1}{2})\Gamma(\frac{n}{2}-1)}{\sqrt{\pi}(\frac{n}{2}-1)\Gamma(\frac{n}{2}-1)} = \frac{n}{n-2}, \text{ cuando } k=1$$

$$E[T^4] = \frac{n\Gamma(2+\frac{1}{2})\Gamma(\frac{n}{2}-2)}{\sqrt{\pi}\Gamma(\frac{n}{2})} = \frac{n(1+\frac{1}{2})(\frac{1}{2})\Gamma(\frac{1}{2})\Gamma(\frac{n}{2}-2)}{\sqrt{\pi}(\frac{n}{2}-1)(\frac{n}{2}-2)\Gamma(\frac{n}{2}-2)} = \frac{3n^2}{(n-2)(n-4)}, \text{ cuando } k=2$$

Luego la Curtosis será: $b = \frac{\mu_4}{\sigma^4} = \frac{E(T^4)}{(E(T^2))^2}$ porque $E(T) = 0$ (momento impar)



$$b = \frac{3n^2/(n-2)(n-4)}{n^2/(n-2)^2} = 3 \left(\frac{n-2}{n-4} \right)$$

Si n es grande, $\lim_{n \rightarrow \infty} 3 \frac{(n-2)}{(n-4)} = 3 \lim_{n \rightarrow \infty} \left(1 + \frac{2}{n-4} \right) = 3, b = 3$

Como era de esperarse, ya que cuando n crece, la distribución T-Student converge en distribución a la Normal y en la Normal el coeficiente de Curtosis es $b = 3$.

¿Quién debe ser el jefe?

Cuando se creó el ser humano, todos sus partes querían ser él jefe

El cerebro dijo! Ya que yo controlo esto y pienso por todos, yo debería ser el jefe.

Los pies dijeron: Ya que nosotros transportamos el cuerpo a donde desea el cerebro y le permitimos así hacer lo que el cerebro quiere, nosotros deberíamos ser los jefes.

Las manos dijeron: Ya que nosotros hacemos todo el trabajo y ganamos dinero para mantener todo el cuerpo, nosotros deberíamos ser los jefes.

Y así siguieron el corazón, los ojos, las orejas, y los pulmones.

Por fin habló el ojo del culo, pidió ser el jefe. Las otras partes del cuerpo se echaron a reír ante la idea de que un ojo del culo pudiera ser el jefe.

El ojo del culo montó en cólera, se cerró y se negó a funcionar. Rápidamente el cerebro enfebreció; los ojos se pusieron bizcos y vidriosos; los pies, demasiado débiles para andar; las manos colgaban sin fuerza, el corazón y los pulmones luchaban por sobrevivir.

Entonces todos suplicaron al cerebro que cediera y permitiese al ojo del culo ser jefe. ¡Así se hizo! Todos las otras partes del cuerpo hacían el, trabajo mientras que el ojo del culo las dirigía y se ocupaba, principalmente, de la mierda, como todo jefe digno de este título.

Moraleja: No hace falta ser cerebro para llegar a jefe; un ojo del culo tiene, claramente, más posibilidades

¡Mire a su alrededor para convencerse!





UNA ALTERNATIVA METODOLÓGICA PARA EL CENSO NACIONAL AGROPECUARIO EN BOLIVIA

Fernando Rivero Sugüiura

El segundo y último Censo Nacional Agropecuario en Bolivia se realizó en septiembre de 1984, después de una serie de postergaciones desde el año 1977 en que se inició la planificación, hasta el año 1979 donde debería realizarse el empadronamiento. Debido a los problemas políticos sociales reinantes en el país, debió ser postergado hasta el año 1980, sin embargo, los acontecimientos de la dictadura militar de esa época, impidieron su realización y obligó a una nueva suspensión pues la Comunidad Mundial canceló los fondos financieros a Bolivia para este proyecto.

El retorno del sistema democrático en el país, permitió retomar la ejecución del proyecto censal en 1983, pero los problemas naturales de sequías e inundaciones en el sector occidental y oriental provocaron la migración campesina y de esta forma impidieron nuevamente su realización.

Un año después, 1984 se ejecutaba el operativo censal a cargo del Instituto Nacional de Estadística, el apoyo de la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO), el Ministerio de Asuntos Campesinos y Agropecuarios (MACA), la Comunidad Europea y el trabajo esforzado de los profesores rurales que debían apoyar con la labor de empadronadores.

La ejecución del trabajo de campo tuvo éxito en algunas regiones del país, pero en otras no alcanzó la cobertura deseada por los problemas político sindicales que se vivían en el momento y la oposición campesina al Censo Agropecuario. En el departamento de La Paz, prácticamente se empadronó sólo dos provincias, lo mismo que en el Chapare cochabambino donde no se pudo realizar el trabajo censal por los problemas ya mencionados.

En conclusión, esta información que fue completada parcialmente en los próximos años, no sirvió más allá de resultados parciales de información estadística agropecuaria y su postergación al desarrollo del sector.

Este antecedente, preocupa y preocupó siempre la realización futura de las encuestas agropecuarias. El marco muestral único con que contaba el sector, era este



censo limitante de cobertura y a consecuencia de ello, las encuestas provenientes de este marco debían afrontar una serie de problemas en la planificación, el trabajo de campo y el incremento de costos.

Es claro que no es el único problema de cobertura que sufre un marco muestral tipo listado de unidades muestrales. Aunque éste no sufriese de falta de cobertura, los problemas se van dando paulatinamente va pasando el tiempo, siendo alguno de ellos la sobrecobertura, la duplicación de unidades muestrales, etc., lo que no sucede con un marco muestral combinado, es decir, de área y lista.

Después del último censo, habían transcurrido once años, hasta que en 1995 se planificó el III Censo Nacional Agropecuario y un año después (1996) se realizó una prueba piloto con este objetivo, pensando de antemano los problemas que afectaba la calidad del marco muestral anterior.

El gobierno de aquel entonces, tenía la intención de contar con información censal del sector, mejorando la metodología de captación de información, bajo el sistema de interpretación de imágenes satelitales que en ese entonces era una novedad respaldada por la FAO y los países de Latinoamérica.

Han pasado siete años desde aquel lejano 1996 y en la actualidad, ha quedado en el olvido aquello que podía haber sido la solución que se buscaba en materia de información agropecuaria bajo los criterios de un método de captación de datos estadísticos precisos y oportunos.

Las experiencias acumuladas en los países de la región, sobre la utilización de metodologías de encuestas agropecuarias empleando marcos de muestreo múltiples, donde se caracterizan las imágenes satelitales, fotografías aéreas y mapas, han llegado a la conclusión que éste es el método más adecuado por la estabilidad que significa conformar un marco muestral de áreas geográficas en base a estos elementos cartográficos y su permanencia fija no desactualizada en al menos 5 años.

La FAO ha invitado a personalidades especializadas en el tema de la construcción de marcos muestrales de área y lista, a brindar apoyo técnico permanente a los países de la región en la que participó Bolivia. En estos talleres se expusieron temas referidos a la metodología y sus condiciones y las frecuentes dificultades que significaba no entrar a éste nuevo sistema metodológico de captación de datos. En la actualidad, muchos de los países de América han tomado la experiencia, tal es el caso de Colombia, Perú, Ecuador,



Argentina y Chile que son países con aproximadamente siete años empleando tal metodología, algunos más que otros. Brasil es uno de los pioneros en usar la técnica descrita, por la amplitud de su territorio agrícola, Uruguay que por las características de su territorio no se ha interesado de dicho sistema. Paraguay y Venezuela que han entrado en esa necesidad y Bolivia que espera la oportunidad de contar con información estadística de encuestas agropecuarias, ojalá diseñando un marco muestral previo, de las características metodológicas discutidas.

Confiar que el sector agropecuario de nuestro país se beneficie en algún momento con esta metodología, ayudaría a implementar información agropecuaria de calidad que en la actualidad no la tenemos.

----- O -----

Examen de moneda

Este es un examen de estadística tipo test; a todas las preguntas hay que responder “cierto” o “falso”. Total, que el profesor ve que hay un estudiante sentado al final de la clase que esta tirando una moneda al aire, la coge en su mano, la mira, hace una marca en su hoja de respuestas, y al final este estudiante es el último que queda por entregar su examen. El profesor se le acerca y le pregunta:

Pero chico, si no siquiera has abierto el examen, como puedes tardar tanto en responder al azar?

Es que estoy comprobando las respuestas.





ESTADÍSTICAS AGRÍCOLAS

PLANES DE TIPO DE MUESTREO

Jaime Pinto

Los conceptos y definiciones de la estadística agrícola varían de un gran país a otro, un concepto que se maneja dentro de ésta, es la Superficie Total, y su significado dependerá de la fuente de información sobre superficies que se hayan consultado.

Las fuentes de las estadísticas sobre superficies, son los Censos agrícolas, las estadísticas agrícolas y los catastros.

Por lo que tiene una importancia, en las estadísticas agrícolas, las mediciones de superficies, realizando estimaciones mediante procedimientos o planes de muestreo.

Las construcciones de la muestra expuestas a continuación, pueden considerarse como “planes tipo de muestreo”, es decir, construcciones bastante sencillas que pueden servir de base a otras más complejas.

Supondremos, en todo lo que sigue, que las unidades de muestreo se han obtenido en la forma debida; sólo analizándose por consiguiente, las operaciones de selección y de estimación, para ello analizaremos en tres partes, que son: Procedimiento de selección, Procedimiento de estimación y Eficacia.

Muestreo Monofásico: Muestreo Simple Al Azar; Procedimiento Normal De Estimación

Procedimiento de selección

Consideremos una región del país, dividido en M unidades de muestreo, con sus cifras de superficie asociadas: $Y_1 ; Y_2 ; Y_3 ; \dots ; Y_M$

Donde la superficie total $Y = \sum_{i=1}^M Y_i$ e Y es el Parámetro que se estimará.

Luego elegimos una muestra de m unidades de muestreo, por el procedimiento de “muestreo simple al azar sin repetición”, es decir se introducen en una urna M papeletas



que se marcan 1,...,k,...,M, respectivamente, se elige una de esas M papeletas; la probabilidad de que resulte elegida cualquiera de ellas será 1/M.

Al ser la selección sin repetición, esta papeleta no se restituye a la urna. Se vuelve a elegir otra papeleta; la probabilidad de que cualquiera de las (M - 1) papeletas restantes resulte elegida será 1/(M-1), se continúa de esta forma hasta elegir m papeletas.

Puede verse que esta selección es equivalente al siguiente: se hace una lista completa de todas las combinaciones de m papeletas distintas tomadas del total M, designando el número de estas combinaciones por $C(M;m)$.

La probabilidad de que resulte elegida cualquiera de las combinaciones de esta lista sera $1/C(M;m)$.

Representando la muestra de m unidades de muestreo por $Y_1 ; Y_2 ; Y_3 ; \dots \dots \dots Y_m$, se puede establecer una clara distinción entre las M unidades que componen el universo estudio y las m unidades que componen la muestra.

Procedimiento de estimación

En la estimación del parámetro Y , se observará primeramente que este parámetro puede escribirse de la forma: $Y = M \cdot \bar{Y}$, donde el promedio \bar{Y} se define por $\bar{Y} = \frac{1}{M} \sum_{k=1}^M Y_k$

Estimar Y es calcular $\hat{y} = M \cdot \bar{y} = M \frac{1}{m} \sum_{i=1}^m Y_i$

Lo característico de este procedimiento de estimación es que no se emplean más observaciones que las indicadas por $Y_1 ; Y_2 ; Y_3 ; \dots \dots \dots Y_m$

Eficacia.

Lo indicado en el Procedimiento de selección y de estimación, una construcción de la muestra de carácter fundamental; forma parte de cualesquiera otras construcciones más complejas. En la práctica, se prefieren en la mayoría de los casos estas otras construcciones, por su mayor eficacia.



Muestreo Monofásico: Muestreo Estratificado; Procedimiento Normal De Estimación

Procedimiento de selección

La característica de esa construcción de la muestra es la estratificación; el universo de M unidades se divide en L subuniversos o estratos, de los cuales se eligen sendas muestras.

Realizando el mismo razonamiento que en el muestreo al azar, el de la urna, la construcción es la siguiente: M papeletas se dividen entre L urnas y se elige una muestra de papeletas de cada una de ellas.

Tomando en cuenta la notación similar a la anterior, tendremos que los L Estratos pueden presentarse en la forma siguiente:

Estrato Unidades de muestreo Parámetro

$$\begin{array}{l}
 1 \ Y_{11} ; \dots Y_{1k} ; \dots \dots \dots Y_{1M_1} \ Y_1 = \sum_{k=1}^{M_1} Y_{1k} \\
 \dots \dots \dots \\
 H \ Y_{H1} ; \dots Y_{Hk} ; \dots \dots \dots Y_{HM_H} \ Y_H = \sum_{k=1}^{M_H} Y_{Hk} \\
 \dots \dots \dots \\
 L \ Y_{L1} ; \dots Y_{Lk} ; \dots \dots \dots Y_{LM_L} \ Y_L = \sum_{k=1}^{M_L} Y_{Lk}
 \end{array}$$

Para el estrato H – ésimo, se tendrá

$$Y_H = M_H \cdot \bar{Y}_H = \sum_{k=1}^{M_H} Y_{Hk}$$

Análogamente al parámetro Y , \bar{Y} es la superficie media por unidad de muestreo del estrato H-ésimo.

Para el universo total, se tendrá.

$$Y = \sum_{H=1}^L M_H \cdot \bar{Y}_H = \sum_{H=1}^L Y_H = \sum_{H=1}^L \sum_{K=1}^{M_H} Y_{HK}$$



De cada uno de estos L estratos se elige una muestra por “muestreo simple al azar sin repetición” en la forma ya explicada anteriormente.

Del estrato H-ésimo, se obtendrá la siguiente muestra:

$$Y_{H1} ; \dots Y_{HK} ; \dots Y_{HmH}$$

Procedimiento de estimación

Con esta muestra se calculará la siguiente estimación de Y_H :

$$y'_H = M_H \cdot \bar{y}_H = M_H \frac{1}{m_H} \sum_{i=1}^{m_H} Y_{Hi}$$

Una vez calculadas las estimaciones de los parámetros Y_H , $H = 1, \dots, L$ se calculará la estimación del parámetro total Y :

$$y' = \sum_{K=1}^L y'_H$$

Eficacia

El empleo del muestreo por estratos, establece cuatro operaciones.

1. Elección de una variable de estratificación.
2. La elección del número L de estratos.
3. La determinación de la forma de estratificación del universo, es decir “la construcción de L estratos”.
4. La selección de la extensión de la muestra que ha de elegir de cada estrato.

Se presenta un ejercicio de estimación del total, para ello de un conjunto de 77 unidades agropecuarias, se extrajo una muestra de 24 unidades, captándose la Variable X, superficie total y teniendo en cuenta los procedimientos en los dos tipos de Muestreo, se desarrolla la estimación del total de superficie.



Muestreo Monofásico: Muestreo Simple Al Azar: Procedimiento Normal De Estimación Del Total

X:2,3,3,4,2,4,5,3,4,6,9,8,7,8,9,7,11,12,13,17,12,13,20,17

$$\text{Estimando } Y : \hat{y} = M \cdot \bar{y} = M \frac{1}{m} \sum_{i=1}^m Y_i = 77 * \frac{1}{24} * 199 = 638,45$$

$$\text{Coeficiente de Variación: } CV(x) = \frac{DE(x)}{\text{Media}} * 100 = \frac{5,02}{8,29} * 100 = 60,65 \%$$

Muestreo Monofásico: Muestreo Estratificado: Procedimiento Normal De Estimación Con Tres Estratos

	Estrato 1	Estrato 2	Estrato 3	Total
1	2	6	11	
2	3	9	12	
3	3	8	13	
4	4	7	17	
5	2	8	12	
6	4	9	13	
7	5	7	20	
8	3		17	
9	4			
M	35	20	22	
m .	9	7	8	24
Total Xi	30	54	115	
Media	3,33	7,71	14,37	8,29
Desv. Estandar	0,94	1,03	2,99	
Estimador Total	116,6	154,28	316,25	587,20
CV(x)	28,28%	13,35%	20,85%	

Comentando sobre la eficiencia relativa, se uede ver que el muestreo estratificado ofrece una variación menor, esto se debe a su estratificación.



Bibliografía:

- Chery G. R. 1967, Practicas de las encuestas estadísticas, Ariel. España
- Kish. 1965. Survey sampling. John Wiley & Sons, N.Y.
- Adela Abad de Servin, Introducción al muestreo, Editorial Limusa, México, 1978
- Willian G. Cochran, Técnicas de Muestreo, Compañía Editorial Continental, S.A. México, 1977 John Wiley & Sons N.Y. Segunda edición.

---- O ----

Profesión más antigüa

Un Médico, un Ingeniero y un Estadístico están charlando sobre cual de sus profesiones es la más antigua. Empieza el Médico:

Pues mira, la Biblia dice que Dios creo a Eva de una costilla de Adán, esto obviamente requiere cirugía, y por lo tanto la medicina es la profesión más antigua.

El ingeniero replica:

Si, bueno, pero antes de eso, la Biblia dice que Dios separó el orden del caos, ésta fue obviamente una obra de ingeniería.

El estadístico se echa para atrás en la silla y dice sonriendo tranquilamente porque sabe que ha ganado esa mano:

Si, pero como te crees que Dios creo el caos?

Dios es real, a menos que sea declarado entero.





ALGUNOS PROBLEMAS DE PROBABILIDAD

MAZO DE BARAJA COMPLETO.

Tengo un mazo de la baraja francesa, completo, de 52 caras. Las mezclo cuidadosamente y saco 11 cartas al azar. ¿Qué probabilidad tengo de que salga un comodín?. ¿Y, si saco 17? ¿Y, si saco 26?.

SOBRE LA SUPERFICIE DE UNA ESFERA.

Sobre la superficie de una esfera marcamos tres puntos al azar. ¿Cuál es la probabilidad de que los tres puntos queden en una misma semiesfera?

TRES BOLAS.

Para elegir a un muchacho entre tres se prepara una bolsa con dos bolas negras y una bola blanca. Los tres van sacando, por orden una bola que no devuelven. Quien saque la bola blanca gana. ¿Quién lleva más ventaja: el primero, el segundo o el tercero?.

EL PENTAGONO.

Se toma al azar un punto situado a varios kilómetros de Pentágono. ¿Qué probabilidad hay de que desde él puedan verse tres lados del Polígono?





SOLUCIONES A LOS PROBLEMAS PLANTEADOS EN LA VARIANZA 2

LOS INCONVENIENTES DE SER DESPISTADO.

Cuando el profesor aceptó la apuesta del estudiante había olvidado por completo que dos de los estudiantes, que siempre se sentaban juntos, eran gemelos.

EL MISMO NÚMERO DE PELOS

La probabilidad buscada es uno. Existe la certeza de que al menos dos habitantes de la nación tienen el mismo número de cabellos.

Dividiendo el número de habitantes por cinco tendremos la superficie en mm^2 que habría que tener el cuero cabelludo para que no se repitan dos cabelleras, en cuanto al número de pelos.

Como $40106/5=8106 \text{ mm}^2=8m^2$ es indudable que habrían de ser muy cabezotas los habitantes de la ciudad.

LAS HERMANAS DE LOS OJOS AZULES

Lo más probable es que las hermanas Flores sean cuatro en total, de las que tres tendrían los ojos azules. Efectivamente, si hay n hermanas de las cuales b tienen los ojos azules, la probabilidad de que elegidas al azar dos de ellas resulten ambas de ojos azules es $b(b-1)/n(n-1)$.

Como nos dice que esta probabilidad es $1/2$ el problema consiste en determinar los valores enteros de b y n que le dan a la expresión anterior el valor $1/2$. Las Soluciones mínimas son $n=4$ y $b=3$; las inmediatamente superiores son ya $n=21$ y

$b=15$. Siendo extremadamente improbables que en una familia haya 21 hermanas, la mejor conjetura es suponer que las hermanas sean cuatro, y que tres de ellas tengan ojos azules.

