



Directora del Departamento
MARÍA EULALIA ARTETA MANRIQUE

Subdirector del Departamento
JAIME OBREGÓN PUYANA

ESTADÍSTICAS AL DÍA

Transversal 45 No. 26-70. Interior 1- CAN - Apartado Aéreo No. 80043.
Conmutador 5978300, Fax 5978399
Sitio Web: www.dane.gov.co
E-mail: dane@dane.gov.co
Bogotá, D.C.

ISSN 1657-625X

Director Editorial
JAIME OBREGÓN PUYANA

Subdirector Editorial
MAURICIO CANAL R.

Consejo Editorial
María Eulalia Arteta Manrique, Jaime Obregón Puyana,
María Eugenia Villamizar, Mauricio Canal R.,
Andrés Traslaviña, Martha Lucía Pabón

Producción e impresión
DIRECCIÓN DE MERCADEO Y EDICIONES

El Departamento Administrativo Nacional de
Estadística - DANE, se reserva los derechos de autor,
Ley 23 de 1982 y Ley 44 de 1993

*Los autores agradecen la revisión técnica de los artículos, por
parte de los colegas Hilda Patricia Ramírez, Eduardo Granados
y Hernando Chacón.*

No. 3, marzo de 2002

EDITORIAL DE LA DIRECCIÓN

La estadística entendida como soporte de la investigación científica para tratamiento de fenómenos cuantitativos y cualitativos, tiene mucho que ofrecer a través de diversas teorías, métodos y técnicas. El desarrollo de procesos que garanticen el mejoramiento de la calidad de dicha investigación, es un reto permanente que se debe abordar con pasión y método.

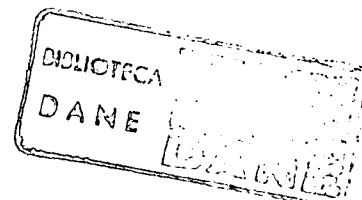
Este número de nuestra revista incluye tres artículos que tratan sobre la aplicación de métodos y técnicas para el mejoramiento de la calidad de la información estadística. Uno de ellos, recoge la revisión bibliográfica de métodos de imputación como una de las alternativas para el tratamiento de la no respuesta, inconsistencias o casos atípicos. En el segundo artículo se presenta un caso real y particular de aplicación práctica de una metodología de imputación.

El tercer artículo muestra el control preventivo de la calidad mediante los rangos de variación, es decir, el análisis del cambio esperado de un fenómeno a partir de su comportamiento histórico y cómo determinar los casos potenciales o efectivamente anormales.

Hemos dado espacio en la revista a jóvenes profesionales, quienes han realizado pasantías y trabajos en el DANE, dos de ellos en el marco del convenio con la Universidad Nacional, como un ejemplo que nos permite mostrar la riqueza obtenida al unir conocimientos innovadores, con la experiencia del DANE.

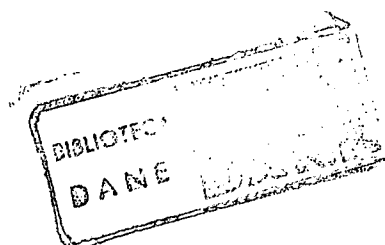
Es nuestro propósito fortalecer estas alianzas estratégicas con la Academia, que potencian las fortalezas del DANE como escuela de investigadores científicos en diversas áreas del conocimiento y, a la vez, contribuyen a mejorar la cultura estadística de nuestro país.

M^a Eulalia Arteta M
MARÍA EULALIA ARTETA MANRIQUE
Directora del Departamento



CONTENIDO

	Pág.
Cómo manejar la no respuesta en encuestas estadísticas	3
No respuesta	3
Causas de no respuesta	3
Técnicas para mantener bajo el nivel de no respuesta	3
Clases de no respuesta	4
Patrón de valores faltantes	4
Estrategias para el manejo de datos faltantes	4
Crítica e imputación	5
Bibliografía	9
Imputación de fuentes en estado de deuda	10
I Marco conceptual	10
II Metodología general de imputación de datos faltantes	10
III Ejemplo	11
Bibliografía	14
Metodología de los rangos de variación para el control de calidad de información de las variables que conforman la muestra de comercio al por menor	15
I Marco conceptual	15
II Metodología general de rangos de variación	15
1. Estacionariedad	15
2. Estabilización de la varianza	16
3. Prueba de normalidad	17
4. Ajuste estacional	17
III Consideraciones adicionales para los rangos de variación	18
1. Ceros en toda la serie	18
2. Datos constantes a través de toda la serie	18
3. Ceros en último trimestre de la serie	18
4. Datos faltantes o ceros dentro de la serie	18
IV Ejemplo	18
1. Rangos de variación para la variable venta de farmacéuticos	19
2. Rangos de variación para la variable venta de artículos de aseo personal	19
3. Rangos de variación para la variable total de ventas	20
Bibliografía	22



CÓMO MANEJAR LA NO RESPUESTA EN ENCUESTAS ESTADÍSTICAS¹

Claudia Murcia Linares*
Alexis Maluendas Pardo*

Uno de los tantos inconvenientes a los que se ve enfrentado el proceso de desarrollo de una encuesta con fines estadísticos, es la presencia de datos faltantes, inconsistentes o atípicos, lo que, en muchos casos, dificulta el proceso de análisis y puede aumentar el sesgo en las estimaciones.

El Departamento Administrativo Nacional de Estadísticas -DANE, no es ajeno a esta problemática, y aunque durante la planeación y desarrollo de cada una de las encuestas socioeconómicas se esfuerce por contar con información de calidad, en muchos casos, es inevitable encontrarse con este tipo de dificultades.

Por ello se hace necesario investigar y actualizar métodos de imputación con el objeto de contar con bases de datos más completas, con criterios de imputación estandarizados que garanticen para todos los usuarios un alto grado de calidad en los análisis y resultados presentados por la Institución. Este aspecto aumenta su importancia al advertirse que el cálculo de algunos índices puede arrojar resultados diferentes, debido en gran parte, al tratamiento dado a los valores no observados.

En este artículo se presenta una visión general de los aspectos fundamentales, que deben tenerse en cuenta para el tratamiento de la no respuesta.

No respuesta

Cuando algunas de las unidades seleccionadas para ser medidas no responde una o varias de las preguntas del cuestionario diseñado para la investigación, se presenta el caso de no respuesta, fenómeno que causa pérdida de representatividad de la población en las variables no contestadas o de los individuos en el caso de la no respuesta total.

Causas de no respuesta

Para el caso de poblaciones humanas, la no respuesta puede ser causada por diversos factores; Cochran (1977) expone algunos de éstos como:

- No cubrimiento: falla en la localización de algunas unidades seleccionadas, ya sea por problemas de acceso cuando en el caso en que las unidades de selección son por ejemplo manzanas, se da el no listamiento de todos los hogares que pertenecen a determinada manzana.
- No localización: se presenta cuando al momento de realizar la encuesta no se encuentra alguna persona en casa.
- Incapacidad: es el caso en el cual los entrevistados no son poseedores de la información o son poco idóneos en el momento de contestar, como las personas con problemas mentales, niños o ancianos.
- Falta de cultura estadística: tienen lugar en aquellas personas que se rehúsan a ser entrevistadas o que muestran poco interés en brindar la información requerida en la encuesta.

Técnicas para mantener bajo el nivel de no respuesta

En algunas investigaciones es muy usual recurrir a diversas estrategias que minimicen el porcentaje de información faltante. Las usadas con mayor frecuencia, son:

- Revisitas (Call-backs): en encuestas con entrevistas personales, el primer contacto con el entrevistado puede ser infructuoso por muchas razones. Por ejemplo cuando una persona se encuentra enferma, accidentada o baja de ánimo. Entonces si en el primer intento para realizar la encuesta no fue posible contactar al entrevistado, se puede realizar una o más visitas en tiempos convenientes de tal forma que se logre recolectar la información del individuo en cuestión.
- Recordatorio (Follow-ups): es utilizado en los casos de encuestas por correo o por internet. Si en un tiempo estipulado el encuestado no envía el cuestionario con la información requerida, se envía otro formulario o se le recuerda el diligenciamiento, aclarando la importancia de la información suministrada para el estudio.

¹Trabajo realizado como parte de una pasantía en la Dirección de Metodología y Producción Estadística del DANE.

*Estudiantes de Estadística de la Universidad Nacional de Colombia.

Las revisitas y los recordatorios son una buena estrategia para minimizar la tasa de no respuesta, pero puede implicar un aumento de los costos y pérdida de tiempo.

- Submuestreo de no respondientes: consiste en realizar un submuestreo de los no respondientes y hacer todos los esfuerzos posibles por obtener las respuestas de todos los elementos en esta submuestra. Este procedimiento reduce en un alto grado el sesgo de las estimaciones.

- "Aleatorización" de la respuesta: Algunas encuestas contienen preguntas sensibles, las cuales no tienen respuesta o su respuesta es falsa, ya que el encuestado se siente agredido en su intimidad. Una solución para ello, es el anonimato del encuestado, también llamada la técnica de la respuesta "aleatorizada". El supuesto es que la respuesta anónima garantizará la cooperación de todos los elementos seleccionados y la veracidad en sus respuestas. Este método es diseñado para encuestas autodilenciadas.

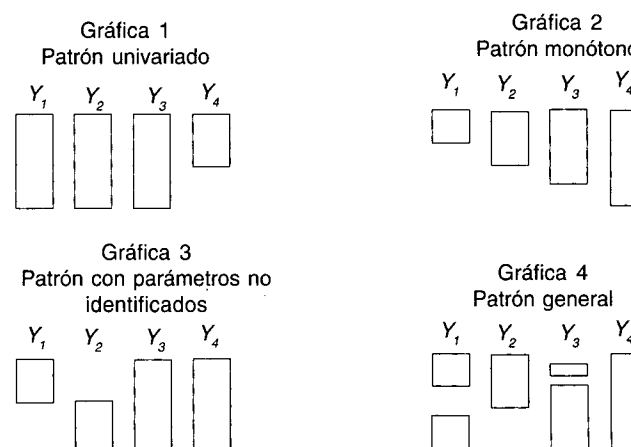
Clases de no respuesta

Droesbeke y Lavallée (1996) proponen una clasificación para la no respuesta, basada en dos características fundamentales, así:

- No respuesta total y parcial: la no respuesta total ocurre cuando no se ha podido medir ninguna de las variables de interés. La no respuesta parcial se tiene cuando sólo se ha podido medir un subgrupo de variables de interés.
- No respuesta ignorable y no ignorable: la no respuesta es ignorable cuando la probabilidad de respuesta no depende del valor de la variable medida, mientras que la no respuesta no ignorable sí depende de este valor.

Patrón de valores faltantes

Otro aspecto fundamental que debe tenerse en cuenta para el tratamiento de datos faltantes, es determinar en qué forma se presenta la no respuesta en las variables de investigación, aspecto que se conoce como patrón de datos faltantes. Cuando una sola de las variables presenta valores faltantes, nos encontramos ante el caso del patrón univariado el cual se puede observar en la gráfica 1. El patrón monótono se presenta cuando las variables pueden ser dispuestas de una manera escalonada como se muestra en la gráfica 2. En el caso que dos variables no sean observadas al mismo tiempo, se conoce como patrón con parámetros no identificados (gráfica 3) y cuando la pérdida de información no tiene un comportamiento particular se dice que los datos faltantes presentan un patrón general (gráfica 4).



Estrategias para el manejo de datos faltantes

Una vez definida la estructura general de los datos faltantes en encuestas estadísticas, el paso a seguir es determinar cómo enfrentar el problema. Algunas de las opciones a las que se puede acudir, son:

- Análisis de casos completos: consiste en tomar los registros con algún valor faltante y descartarlos del análisis. La ventaja de este método está en su fácil implementación y en el hecho de que se obtienen inferencias válidas en el análisis, aspecto que no es compartido por otras aproximaciones (Little, R. 1992). Sin embargo, se debe tener cuidado ya que se puede perder mucha información en los casos donde la no respuesta se observa con gran frecuencia. Este método es usado frecuentemente como base de comparación con otros métodos.
- Análisis de casos disponibles: consiste en realizar el análisis con la base de datos incompleta, es decir, los individuos donde se presente no respuesta parcial son incluidos en el análisis y no se utiliza ningún tipo de reponderación para el caso de la no respuesta total. Presenta inconvenientes al momento del cálculo de la matriz de varianzas y covarianzas. Además, se ha encontrado que en simulaciones basadas en datos altamente correlacionados es mejor el análisis de casos completos, mientras que el análisis de datos disponibles es mejor en simulaciones basadas en datos débilmente correlacionados.
- Procedimientos ponderados: son una forma estándar para compensar la no respuesta total (Kalton y Kasprzyk, 1986). La esencia de todos los procedimientos ponderados es incrementar los pesos de respondientes especificados de modo que ellos representen los no respondientes.

- Procedimientos basados en imputación: consisten en llenar o imputar valores para los datos faltantes de las variables usadas en un análisis. Este método tiene como ventaja la maximización de la información disponible de los datos incompletos y permite el uso de los métodos estándar de análisis de los datos completos (López, A. 2000).

Crítica e imputación

El proceso de imputación estadística es una de las estrategias para el manejo de datos faltantes y está íntimamente ligado al concepto de crítica estadística. La crítica es el proceso de identificar valores erróneos o faltantes; la imputación es el proceso para corregirlos.

Los datos colectados en la encuesta pueden presentar tres tipos de problemas:

- Datos inconsistentes
- Datos faltantes
- Datos atípicos

Bautista, L. (1998), afirma que la evaluación de si los datos de una encuesta son o no aceptables, depende básicamente del uso que se va a hacer de ellos. Desde un punto de vista muy amplio, la calidad de la información producida está asociada a:

- La población objetivo
- El grado de varianza asociado a los resultados
- La magnitud del sesgo debido a errores operativos
- La pertinencia de los métodos de observación y medición, en particular de las preguntas planteadas en el cuestionario.

Desde una perspectiva más operativa, la aceptabilidad se determina por la integridad total de los datos en términos de:

- ¿Qué tan completos están?
- ¿Qué tan correctos son?
- ¿Qué grado de consistencia interna presentan?

Los procesos de crítica e imputación surgen de la necesidad de garantizar calidad en la información, es así como estos dos procesos tienen por objeto determinar el grado de integridad de los datos en referencia a estos tres últimos aspectos (completitud, corrección y consistencia interna).

La crítica estadística es un proceso que implica la aplicación de reglas predeterminadas al conjunto de datos. Si un registro no cumple con alguna de las reglas establecidas, debe ser señalado para ser corregido o cambiado; una vez definido el plan de crítica se deben

realizar las operaciones de verificación, validación y consistencia.

- La verificación es el proceso que generalmente acompaña a la grabación de los datos en un medio magnético y que consiste en garantizar que lo grabado es un fiel reflejo de lo contenido en los cuestionarios.
- La validación es el proceso por el cual se determina si los datos cumplen ciertas reglas preestablecidas de aceptabilidad. La verificación de la regla a puede realizarse al momento de grabar, determinando si la variable cumple o no la regla establecida. Por ejemplo, que los datos para una variable se encuentren entre unos valores mínimo y máximo.
- La consistencia es el proceso que determina incoherencias a partir de la información completa del formulario y se refiere a la preservación de las relaciones entre variables. Un caso particular podría ser: el número de empleados temporales no puede ser mayor al total de empleados por la empresa.

Después de detectado el dato erróneo, faltante, inconsistente o atípico, el método más simple para su corrección parece ser el retornar a la fuente primaria de información, lo cual puede ser impráctico por razones de costos y falta de colaboración del respondiente en algunos casos; pero aún después de este tratamiento puede seguirse presentando el mismo problema pese a los esfuerzos hechos en seguimiento y control.

La escogencia de la estrategia que se utilizará depende no solamente de la estructura general descrita anteriormente, sino también del tipo de investigación y de la perspectiva del investigador. Cuando el investigador decide tratar el dato faltante, reemplazándolo por algún valor posible, debe considerar aspectos como el tipo de variable, el porcentaje de no respuesta que presenta y escoger las variables más correlacionadas con la variable a imputar.

En la tabla 1 presenta un esquema de los métodos de imputación más comunes, identificando sus ventajas y sus desventajas. Estos métodos tienen la ventaja de no ser excluyentes, es decir, la aplicación de un determinado método para una variable no implica la no aplicabilidad de otro método a otra u otras variables.

En todos los casos se deben especificar aspectos referentes a la calidad de los datos; en la publicación de los resultados se debe orientar al lector acerca de las tasas de no respuesta total y parcial y tasas de imputación por cada campo o variable imputada, de tal manera que pueda valorarse el impacto sobre los resultados de la encuesta y sirva de guía en las decisiones que se tomarán con base en la información suministrada. En la tabla 2, se incluye un modelo de ficha técnica referente a la calidad estadística de los datos.

Tabla 1
Comparación de los métodos de imputación

Métodos de imputación	Definición	Procedimiento	Ventajas	Desventajas
DETERMINÍSTICA	El dato faltante recibe un único valor posible, a fin de que se cumplan las reglas de crítica planteadas.	Se realiza únicamente con información del mismo registro, en las especificaciones de crítica, de una vez se asume la acción de reemplazar el dato por el valor "correcto".	* Ideal, ya que el dato se deduce de forma acertada, usando la información obtenida	
SUBSTITUCIÓN	Se hace uso de una base de datos auxiliar o de un archivo administrativo, el cual contiene información referente a las mismas variables del mismo individuo en un tiempo anterior.	Se utiliza especialmente para encuestas continuas, se debe tener cuidado con la definición de la variable a imputar, ya que para diferentes estudios la definición de la variable puede no ser la misma.	* Ideal, ya que usa información auxiliar de una forma plausible	
EN FRÍO O DE EXPERTOS	El valor faltante es imputado a criterio de un experto en la temática o con un dato histórico no necesariamente del mismo individuo.	Se encuentra el mejor valor "adivinado"	* Útil cuando los datos a imputar son pocos	* No tiene en cuenta la distribución de los datos y las relaciones entre variables * Existe incertidumbre en el sesgo * Es un valor subjetivo
ENCALIENTE	Consiste en la sustitución del dato faltante por uno perteneciente a la misma encuesta.	Se divide la base de datos en dos grupos: los receptores y los donantes. En el grupo de los receptores se incluyen los registros con al menos un dato faltante, inconsistente o atípico y el grupo de los donantes contiene los registros con todos sus datos completos o "limpios". Cada registro de los receptores encuentra un donante muy similar.		
A. Imputación secuencial	Se crea una matriz de imputación en la cual se construyen todas las posibles combinaciones de las variables auxiliares. A medida que se van leyendo los datos, la matriz se actualiza y cuando encuentra un dato a ser imputado, se toma el correspondiente de la matriz de imputación.	El receptor tiene el mismo valor que el donante para todos los campos de similitud e imputación. Cada matriz de imputación debe reiniciarse utilizando valores ad-hoc a fin de evitar que los primeros registros imputados sean imputados con blanco.	* Se generan imputaciones coherentes dentro de cada clase * Se dispone de software para su aplicación (solas) * No requiere supuestos distribucionales	* Algunas veces es complicado encontrar las variables auxiliares adecuadas * La reiterada utilización de un mismo donante puede generar fuertes sesgos en los estimativos del estudio
B. Selección aleatoria	Consiste en crear un subgrupo de donantes, los cuales tienen los mismos valores en las variables de similitud, y seleccionar uno al azar como donante.	Se codifican las variables numéricas previamente a su utilización como variable de similitud, para evitar el crecimiento de la matriz de imputación. Se establece un criterio de jerarquía entre estas variables.		

Tabla 1 (continuación)
Comparación de los métodos de imputación

Continuación

Métodos de imputación	Definición	Procedimiento	Ventajas	Desventajas
C. El vecino más próximo	Bajo la misma filosofía del método anterior se busca el donante que más se aproxime en las variables de similitud, sin que sea exactamente el mismo y que además pertenezca a una clase similar en otra variable.	Busca el donante en forma sucesiva hasta llegar a un punto en que la imputación no infringe las reglas de crítica.		
SIMPLE	Se utilizan los datos "correctos" de la misma encuesta, se procede a estimar un parámetro como: la media y la moda, cuyo valor sirva para imputar la variable con datos faltantes, inconsistentes extraños.	Para variables cuantitativas continuas se reemplaza el valor faltante, inconsistente extraño por el promedio (media general o media de clase*) de la variable a imputar. Para variables cualitativas y cuantitativas discretas, el valor a imputar se reemplaza por la moda (moda general o de clase*).	* Se utiliza estadística descriptiva Alto grado de estabilidad, ya que se imputa un valor esperado	* Subestima la varianza * Las distribuciones marginales y la matriz de correlaciones se distorsionan * Los errores estándar de los valores imputados son demasiado pequeños
REGRESIÓN	Se crea un modelo de regresión lineal donde las variables explicativas son las variables más altamente correlacionadas con la variable a ser imputada.	Se emplea información auxiliar para imputar valores faltantes, inconsistentes y extraños.	* Mejor que imputación de la media	* Errores estándar demasiado pequeños * Existe incertidumbre en la estimación * Se subestima la varianza
MÚLTIPLE	Se reemplaza el valor faltante, inconsistente o extraño por dos o más posibles valores, creando así dos o más conjuntos de datos completos; con cada conjunto se hacen los análisis y finalmente se hace un análisis conjunto.	Se generan dos o más bases de datos imputadas, ya sea por métodos de regresión o bayesianos, para realizar en cada base la estimación de parámetros, los cuales se promedian, a fin de obtener un estimador final proveniente de varias bases imputadas.	* Se obtienen inferencias válidas para fracciones modestas de valores faltantes * Se generan imputaciones de los datos faltantes de su distribución predictiva posterior * Es sencillo analizar los datos con los paquetes estándar de análisis estadístico * Se encuentra disponible en varios paquetes estadísticos (SAS, SOLAS) * Está motivado bayesianamente con propiedades frecuentistas buenas	* La creación de una base de datos con imputación múltiple es en general difícil * Requiere algunos supuestos acerca de la distribución de los datos

Tabla 1 (conclusión)
Comparación de los métodos de imputación

Métodos de imputación	Definición	Procedimiento	Ventajas	Desventajas
MÁXIMA VEROSIMILITUD	Hace uso del algoritmo EM (estimación-maximización), el cual es iterativo y lo que se busca es imputar con los posibles valores que maximicen la función de distribución de los datos	Partiendo de un valor inicial de la media y de la matriz de varianzas y covarianzas, el algoritmo crea una matriz completa con datos imputados en cada iteración, renovando así los parámetros, hasta que éstos no cambien significativamente, teniendo finalmente una base de datos imputada de acuerdo a los valores finales de los parámetros. Para aplicar este método, las variables deben ser de tipo continuo.	* Los estimadores obtenidos están en el espacio de parámetros * Estos estimadores son consistentes y eficientes si el modelo de base es correcto (MAR***, MCAR**) * Los errores estándar, pruebas de hipótesis e intervalos de confianza basados en el teorema del límite central, son válidos	* Requiere fuertes supuestos distribucionales (Normalidad)

* Las clases están dadas por variables de clasificación.

** Se presentan valores faltantes de forma completamente aleatoria MCAR, si el patrón de valores faltantes no depende de los valores observados ni de los faltantes.

*** Se presentan valores faltantes de forma aleatoria MAR, si el patrón de valores faltantes depende únicamente de los valores observados.

Tabla 2
Ficha técnica de imputación

INVESTIGACIÓN								
PERÍODO DE RECOLECCIÓN								
Nombre de la variable	Descripción	Tipo de variable	Valores válidos	Porcentaje de información faltante	Porcentaje de información atípica	Porcentaje de información imputada	Método de imputación	Códigos

BIBLIOGRAFÍA

BAUTISTA, L . Diseños de muestreo estadístico. Universidad Nacional de Colombia. Bogotá. 1998

COCHRAN, W. G Sampling Techniques. Tercera edición. Wiley. Nueva York. 1977.

DROESBEKE, J.J. y LAVALLÉE, P. La no respuesta en las encuestas. En: Methodologica. 1996. Nº 4.

KALTON, G. y KASPRZYK, D. The treatment of missing survey data. In: *Survey methodology*. 1986. Vol.12. Pag. 1-16.

LITTLE, R. Regression With Missing X's: A Review. In: Journal of the American Statistical Association. 1992. Vol. 87.

LÓPEZ, A. Cuantificación de la severidad fisiológica en pacientes admitidos a cuidado intensivo usando métodos de Imputación Simple y Hot-Deck. 2000. Universidad Nacional de Colombia. Bogotá, D. C.

IMPUTACIÓN DE FUENTES EN ESTADO DE DEUDA

I- Marco conceptual

José Elías Mendivelso (Estadístico)
Andrés René Lozano (Estadístico)

Una gran cantidad de información acerca de las características económicas tanto de individuos como de establecimientos industriales o países, es recopilada con fines de análisis para planear y tomar decisiones.

El registro sistemático de mediciones u observaciones numéricas, efectuado a intervalos fijos de tiempo, se conoce como serie de tiempo y como esta se compone de datos numéricos, es común usar la estadística para analizarla ya sea de forma descriptiva o de manera inferencial. El procedimiento de esta última, es la utilización de muestras que representen a la población de estudio, con el fin de producir conclusiones válidas para toda la población.

Uno de los problemas que se presentan en el análisis estadístico inferencial es la falta de algunos registros en la serie, lo que conlleva aumentar el error en la varianza de las estimaciones de los parámetros poblacionales. Para hacer menos grave el error, se presentan dos métodos de estimación con datos faltantes, que son la reponderación y la imputación.

La imputación es un método muy usado, con el cual se debe hacer el esfuerzo de afectar a lo sumo el 2% de los datos. Si el porcentaje de datos imputados es muy alto, se crea un error sistemático o sesgo en la varianza del estimador puntual. Pero aun si un método de imputación no produce un apreciable error, no se debe ignorar el efecto que la imputación tiene en la precisión de la varianza del estimador puntual.

La imputación es útil porque hace más viable el análisis de un conjunto de datos, lo que asegura consistencia entre los resultados de diferentes análisis y reduce el sesgo de no respuesta.

Varias técnicas estadísticas requieren de conjuntos de datos rectangulares o en forma de matriz y que con presencia de datos faltantes, los registros se pueden restringir a un conjunto de datos completos. Esta restricción sacrifica información parcial en aquellas encuestas que no han sido diligenciadas totalmente y que se pueden utilizar o aprovechar si se hace imputación.

En la literatura estadística hay una variedad de métodos que se han propuesto para imputar datos; estos métodos son clasificados según si se genera una sola imputación para cada valor faltante (imputación simple) o se generan bajo simulaciones, m imputaciones para cada valor faltante el cual genera m conjuntos de datos completos (imputación múltiple).

Algunos métodos de imputación usan un modelo explícito como el de una regresión ajustada, una razón o la imputación por la media. En otros métodos el modelo es implícito como el de la imputación en paquete caliente (hot deck) y la imputación por donadores vecinos.

II- Metodología general de imputación de datos faltantes

En esta metodología se utiliza la información de la muestra mensual manufacturera, de tal manera que los datos imputados se aproximen a los valores reales. La metodología supone que los datos de la muestra poseen autocorrelación temporal y homogeneidad en las diferentes etapas de agregación; esto significa que la imputación debe estar de acuerdo con el comportamiento de la serie histórica y de los niveles que contienen al dato faltante.

Para la imputación de registros en estado de deuda se utiliza la razón de crecimiento de los datos en la serie o variación de los datos, presentada en la metodología de imputación de Andrés Lozano, titulada "Estimación de novedades en estado de deuda", definida como:

$$\text{Variación} = \frac{X_t}{X_{t-1}}$$

Donde,

X_t = dato en el período t

X_{t-1} = dato en el período anterior $t-1$

Bajo estas consideraciones, se estimará primero la variación que tendrá el dato faltante con respecto al dato del período anterior, teniendo en cuenta el comportamiento histórico de la serie de variaciones en cada establecimiento industrial y el comportamiento histórico de las variaciones dentro de cada clase industrial según la CIIU REV 3. A.C.; a partir de esta estimación se generará el dato faltante.

El cuadro 1, presenta un ejemplo, realizado para el establecimiento con número de orden 509 y que pertenece a la clase industrial con código 2 729, de las variaciones calculadas para las variables personal permanente, y producción.

Cuadro 1
Variaciones de las variables total de empleados permanentes y de la producción
2000-2001 (mayo)

Año	Mes	Total empleados permanentes	Variación del total de empleados permanentes	Total de empleados permanentes por clase industrial	Variación del total de empleados permanentes por clase industrial	Producción (miles de pesos)	Variación de la producción	Producción por clase industrial (miles de pesos)	Variación de la producción por clase industrial
2000	Enero	124		425		3.033.550		15.765.529	
	Febrero	124	1,00	429	1,01	3.331.605	1,10	18.321.540	1,16
	Marzo	124	1,00	427	1,00	3.167.678	0,95	18.127.165	0,99
	Abril	118	0,95	418	0,98	2.368.393	0,75	16.004.662	0,88
	Mayo	117	0,99	417	1,00	3.099.148	1,31	17.312.161	1,08
	Junio	117	1,00	401	0,96	3.140.401	1,01	18.996.862	1,10
	Julio	114	0,97	401	1,00	4.265.501	1,36	19.957.135	1,05
	Agosto	111	0,97	437	1,09	4.396.068	1,03	20.473.469	1,03
	Septiembre	109	0,98	435	1,00	4.762.902	1,08	21.589.968	1,05
	Octubre	109	1,00	436	1,00	3.715.239	0,78	20.488.398	0,95
	Noviembre	112	1,03	440	1,01	5.818.543	1,57	23.159.622	1,13
	Diciembre	112	1,00	438	1,00	4.174.629	0,72	21.763.649	0,94
2001	Enero	114	1,02	393	0,90	4.314.837	1,03	22.192.023	1,02
	Febrero	112	0,98	398	1,01	5.219.292	1,21	23.572.223	1,06
	Marzo	112	1,00	406	1,02	6.038.856	1,16	23.805.178	1,01
	Abril	112	1,00	402	0,99	5.380.395	0,89	21.436.420	0,90
	Mayo	112	1,00	406	1,01	5.601.469	1,04	25.033.372	1,17

La variación del dato que se va a imputar se obtiene en términos de la variación histórica promedio en los últimos $t-k$ períodos y de la variación estacional anual en los períodos $t-11$ hasta $t-13$ en el establecimiento industrial y en la clase industrial, donde k representa el número de períodos retrasados y que se utilizan para observar la evolución histórica de la serie dentro del establecimiento o dentro de la clase industrial.

El modelo para imputar la variación, es:

$$\text{Variación}_t = \beta_1 \text{Vhe} (t-1, t-2, t-3) + \beta_2 \text{Vhe} (t-11, t-12, t-13) + \beta_3 \text{Vha} (t-1, t-2, t-3) + \beta_4 \text{Vha} (t-11, t-12, t-13)$$

Donde,

Variación_t = variación a estimar

Vhe = variación histórica promedio por establecimiento industrial de los períodos $t-1, t-2$ y $t-3$ o la variación estacional anual en los períodos $t-11, t-12$ y $t-13$

Vha = variación histórica promedio por clase industrial de los últimos k períodos o la variación estacional anual en los períodos $t-11, t-12$ y $t-13$

β_i para $i=1, 2, 3$ o 4 , son coeficientes de ponderación

El modelo describe la imputación de la variación del

dato faltante como un promedio ponderado de las variaciones de los datos en el establecimiento y en la clase industrial, donde los β_i son los coeficientes de ponderación de las variaciones.

Como se expone en Lozano, "el propósito es estimar los parámetros desconocidos β_i , utilizando un método iterativo con el modelo de mínimos cuadrados y restringiéndolos a que la suma sea igual a 1, para que haya convergencia en la imputación".

Utilización del modelo

Se tendrán en cuenta los supuestos planteados en la metodología de imputación "Estimación de novedades en estado de deuda," de Andrés Lozano (2000), para el buen desempeño del modelo, los cuales son el de homogeneidad de los datos dentro de la clase industrial y la autocorrelación temporal entre las variaciones de los datos de la serie histórica dentro de cada establecimiento y dentro de cada clase industrial, supuestos que se utilizan en la estructura del modelo.

III- Ejemplo

Para explicar la metodología, se procederá a presentar un ejemplo de imputación de la variación y del total de las variables empleados permanentes y producción, en mayo de 2001, para el establecimiento con número de orden 509.

Utilizando la base de datos en la cual se encuentran los datos históricos por establecimiento, se calcularon las variaciones que han tenido las variables por establecimiento y por cada clase industrial en los últimos 13 períodos; luego se calculó el promedio de las variaciones de marzo y abril de 2001 para conocer la evolución de la variable en los dos meses anteriores, dentro del establecimiento y dentro de la clase industrial; también se calculó el promedio de las variaciones de abril y mayo de 2000 para determinar la estacionalidad de las variaciones dentro del establecimiento y dentro de la clase industrial.

Empleando diferentes combinaciones de parámetros, con la restricción exigida que la suma sea igual a 1, se procedió a imputar las variaciones y el total de las variables mencionadas anteriormente.

En el cuadro 2 se relacionan los datos reales del total y de las variaciones de las variables dentro del establecimiento y de la clase industrial, datos que para observar la bondad del ajuste del modelo, se quieren imputar.

Cuadro 2
Datos de los totales y de las variaciones reales para las variables total de empleados permanentes y producción

Total empleados permanentes	Variación del total de empleados permanentes	Total de empleados permanentes por clase industrial	Variación del total de empleados permanentes por clase industrial	Producción (miles de pesos)	Variación de la producción	Producción por clase industrial (miles de pesos)	Variación de la producción por clase industrial
112	1,00	406	1,01	5.601.469	1,04	25.033.372	1,17

Las imputaciones obtenidas para la variación y para el total de la variable total de empleados permanentes, se presentan en los cuadros 3 y 4 respectivamente, y se muestra en la gráfica 1, el dato real y la imputación más

grande y la más pequeña del total de empleados permanentes. En este caso sólo se observa una imputación, debido a que por redondeo todas las imputaciones fueron iguales.

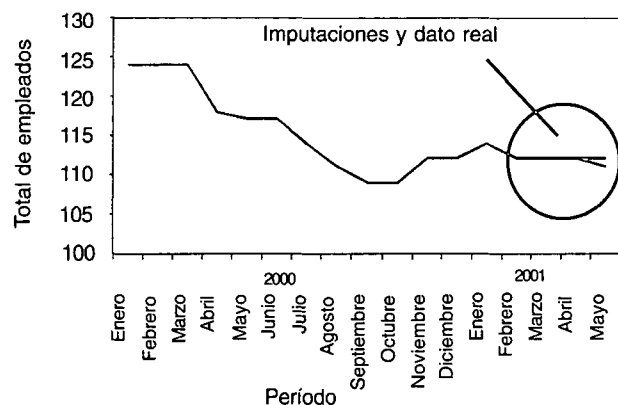
Cuadro 3
Imputaciones de la variación del total de empleados permanentes para distintas combinaciones de parámetros

Variaciones imputadas	Combinaciones de parámetros								
	$\beta_1 = \beta_3 = 0.05$ $\beta_2 = \beta_4 = 0.45$	$\beta_1 = \beta_3 = 0.1$ $\beta_2 = \beta_4 = 0.4$	$\beta_1 = \beta_3 = 0.15$ $\beta_2 = \beta_4 = 0.35$	$\beta_1 = \beta_3 = 0.2$ $\beta_2 = \beta_4 = 0.3$	$\beta_1 = \beta_3 = 0.25$ $\beta_2 = \beta_4 = 0.25$	$\beta_1 = \beta_3 = 0.35$ $\beta_2 = \beta_4 = 0.15$	$\beta_1 = \beta_3 = 0.35$ $\beta_2 = \beta_4 = 0.15$	$\beta_1 = \beta_3 = 0.4$ $\beta_2 = \beta_4 = 0.1$	$\beta_1 = \beta_3 = 0.45$ $\beta_2 = \beta_4 = 0.05$
	0,9933	0,9929	0,9926	0,9923	0,9920	0,9917	0,9914	0,9911	0,9908

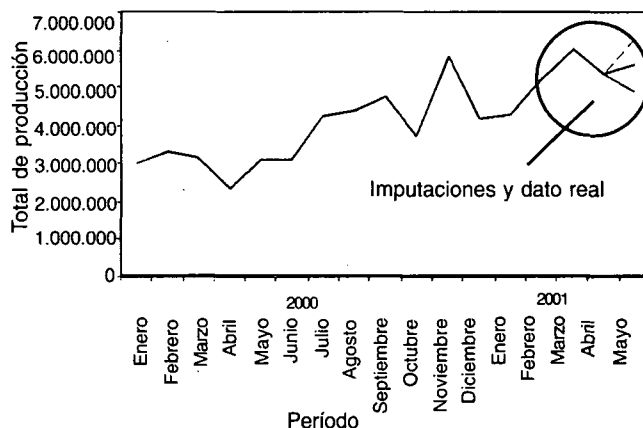
Cuadro 4
Imputaciones del total de empleados permanentes para distintas combinaciones de parámetros

Variaciones imputadas	Combinaciones de parámetros								
	$\beta_1 = \beta_3 = 0.05$ $\beta_2 = \beta_4 = 0.45$	$\beta_1 = \beta_3 = 0.1$ $\beta_2 = \beta_4 = 0.4$	$\beta_1 = \beta_3 = 0.15$ $\beta_2 = \beta_4 = 0.35$	$\beta_1 = \beta_3 = 0.2$ $\beta_2 = \beta_4 = 0.3$	$\beta_1 = \beta_3 = 0.25$ $\beta_2 = \beta_4 = 0.25$	$\beta_1 = \beta_3 = 0.35$ $\beta_2 = \beta_4 = 0.15$	$\beta_1 = \beta_3 = 0.35$ $\beta_2 = \beta_4 = 0.15$	$\beta_1 = \beta_3 = 0.4$ $\beta_2 = \beta_4 = 0.1$	$\beta_1 = \beta_3 = 0.45$ $\beta_2 = \beta_4 = 0.05$
	111	111	111	111	111	111	111	111	111

Gráfica 1
Imputaciones y dato real del total de empleados permanentes
2000 - 2001 (mayo)



Gráfica 2
Imputaciones mínima, máxima y dato real del total de producción
2000 - 2001 (mayo)



Los cuadros 5 y 6 presentan respectivamente las imputaciones de la variación y del total para la variable producción; la imputación más grande y más pequeña

y el dato real del total de producción se muestran en la gráfica 2.

Cuadro 5
Imputaciones de las variaciones del total de producción
para distintas combinaciones de parámetros

Combinaciones de parámetros									
$\beta_1=\beta_3=0.05$ $\beta_2=\beta_4=0.45$	$\beta_1=\beta_3=0.1$ $\beta_2=\beta_4=0.4$	$\beta_2=\beta_4=0.15$ $\beta_2=\beta_4=0.35$	$\beta_1=\beta_3=0.2$ $\beta_2=\beta_4=0.3$	$\beta_1=\beta_3=0.25$ $\beta_2=\beta_4=0.25$	$\beta_1=\beta_3=0.35$ $\beta_2=\beta_4=0.15$	$\beta_1=\beta_3=0.35$ $\beta_2=\beta_4=0.15$	$\beta_1=\beta_3=0.4$ $\beta_2=\beta_4=0.1$	$\beta_1=\beta_3=0.45$ $\beta_2=\beta_4=0.05$	
Variaciones imputadas	0,97	0,96	0,96	0,95	0,94	0,93	0,93	0,92	0,91

Cuadro 6
Imputaciones del total de producción
para distintas combinaciones de parámetros

Combinaciones de parámetros									
$\beta_1=\beta_3=0.05$ $\beta_2=\beta_4=0.45$	$\beta_1=\beta_3=0.1$ $\beta_2=\beta_4=0.4$	$\beta_2=\beta_4=0.15$ $\beta_2=\beta_4=0.35$	$\beta_1=\beta_3=0.2$ $\beta_2=\beta_4=0.3$	$\beta_1=\beta_3=0.25$ $\beta_2=\beta_4=0.25$	$\beta_1=\beta_3=0.35$ $\beta_2=\beta_4=0.15$	$\beta_1=\beta_3=0.35$ $\beta_2=\beta_4=0.15$	$\beta_1=\beta_3=0.4$ $\beta_2=\beta_4=0.1$	$\beta_1=\beta_3=0.45$ $\beta_2=\beta_4=0.05$	
Totales imputados (miles de pesos)	5.231.285	5.189.393	5.147.501	5.105.609	5.063.717	5.021.825	4.979.933	4.938.041	4.896.149

El cálculo de los coeficientes de variación para las dos variables dio como resultado: para la variable total de empleados permanentes, un coeficiente de variación igual a 0,18 y para la variable producción, un coeficiente de variación igual a 0,21, indicando con esto que los

datos no son muy homogéneos pero aun sin el cumplimiento de este supuesto las imputaciones de los totales que se presentan bajo las diferentes combinaciones de parámetros, no difieren demasiado de los datos reales.

BIBLIOGRAFÍA

Roderick, J.A. Models for Nonresponse in Sample Surveys. Journal of the American Statistical Association. 1982.

Martin, D. et. al. Alternative Methods for CPS Income Imputation. Journal of the American Statistical Association. 1986.

Guerrero, V. Análisis estadístico de series de tiempo económicas. México. 1991.

Lozano, A. Estimación de novedades en estado de deuda. DANE. 2000.

METODOLOGÍA DE RANGOS DE VARIACIÓN PARA EL CONTROL DE CALIDAD DE INFORMACIÓN DE LAS VARIABLES QUE CONFORMAN LA MUESTRA DE COMERCIO AL POR MENOR

I- Marco conceptual

José Elías Mendivelso (Estadístico)
Andrés René Lozano (Estadístico)

El éxito de las estimaciones en la inferencia estadística aplicada a las encuestas económicas depende de la calidad de los datos que reportan los establecimientos fuentes, al DANE. Por esto se pretende controlar los datos de las encuestas a través del análisis de las variaciones que tienen los mismos, a través del tiempo. La manera más común de llevar a cabo un análisis de calidad es construyendo límites de control que permitan detectar cuándo una variación está por fuera del comportamiento corriente de la variable.

A partir de este objetivo se desarrolla una metodología utilizando la información existente para la muestra mensual de comercio al por menor, creando sus límites de control. La forma más eficiente de establecer los límites, es por medio de intervalos de confianza, asumiendo que las variaciones presentan una distribución normal.

La muestra mensual de comercio al por menor cuenta con un número suficiente de datos, que hace posible la verificación del supuesto de normalidad que deben cumplir las variaciones para la construcción de los intervalos de confianza; como las variaciones se distribuyen normales, se construyen intervalos de confianza simétricos alrededor de la media.

II- Metodología general de los rangos de variación

En el caso de que algunas variaciones no cumplan con este supuesto, se establecerán intervalos asimétricos, de acuerdo al coeficiente de asimetría o sesgo.

Para medir el comportamiento de la serie de datos en la muestra mensual de comercio al por menor, se utiliza la variación o razón de crecimiento de los datos, definida como:

$$\text{Variación} = \frac{X_t}{X_{t-1}} \quad (1)$$

Donde,

X_t = dato en el período t

X_{t-1} = dato en el período anterior $t-1$

La variación está dada en términos del cociente entre la observación presentada en el período t y la observación del período anterior ($t-1$).

El cuadro 1 presenta un ejemplo, realizado para la empresa con número de orden 20198 y que pertenece a la actividad con código 5231, de las variaciones calculadas para las variables ventas de: farmacéuticos, artículos de aseo personal, y del total de ventas.

Bajo estos términos se construyen límites de control de calidad para la variación de los datos. Para establecer los límites de variación se tiene en cuenta el comportamiento histórico de la serie de variaciones en cada empresa.

La metodología para construir los límites de control comienza con una revisión de la tendencia o estacionariedad que tiene la serie de variaciones y con una transformación de las variaciones para homogeneizar la varianza; lo que se pretende con esta transformación es que se pueda hacer inferencia con los límites hallados, de tal forma que estas inferencias sean confiables.

1- Estacionariedad

Una serie de tiempo es estacionaria, en primer orden, si su media no se ve afectada en el tiempo, es decir, que aun cuando durante un cierto período de tiempo el proceso se aleje de la media, éste siempre regresa a ubicarse alrededor de ella.

Por consiguiente, el concepto de estacionariedad es de suma importancia cuando se analizan series temporales. Debido a que la mayoría de las series temporales que se trabajan en la práctica no son estacionarias, se utilizan técnicas estadísticas para transformarlas en series estacionarias.

En muchas situaciones la no estacionariedad se presenta en el nivel, es decir, en la media, situación que es posible afrontar haciendo una diferenciación a la serie, lo cual significa restarle al dato del período t el dato del período $t-1$. Otra opción es la variación de los datos en el tiempo; ésta se utiliza en el presente trabajo.

Cuadro 1
Variaciones de las variables ventas de: farmacéuticos, artículos de aseo personal,
y del total de ventas
1999 - 2001 (agosto)

Período		Variación de ventas de farmacéuticos	Variación de ventas de artículos aseo personal	Variación del total ventas
1999	Febrero	0,77	0,77	0,77
	Marzo	1,19	1,19	1,19
	Abril	0,86	0,86	0,86
	Mayo	1,33	1,33	1,33
	Junio	0,95	0,95	0,95
	Julio	1,02	1,02	1,02
	Agosto	1,03	1,03	1,03
	Septiembre	0,96	0,96	0,96
	Octubre	1,04	1,04	1,04
	Noviembre	0,96	0,96	0,96
	Diciembre	1,14	1,14	1,14
2000	Enero	0,97	1,00	0,98
	Febrero	0,99	0,97	0,99
	Marzo	1,00	1,08	1,02
	Abril	0,95	0,93	0,94
	Mayo	1,19	1,05	1,15
	Junio	0,99	1,00	0,99
	Julio	1,03	1,02	1,03
	Agosto	1,04	1,01	1,03
	Septiembre	0,91	0,95	0,92
	Octubre	1,10	1,09	1,10
	Noviembre	0,98	0,90	0,96
	Diciembre	1,06	1,27	1,11
2001	Enero	1,11	0,98	1,07
	Febrero	0,81	0,82	0,81
	Marzo	1,09	1,13	1,10
	Abril	0,96	0,96	0,96
	Mayo	1,10	1,08	1,09
	Junio	1,02	1,01	1,02
	Julio	1,03	1,06	1,04
	Agosto	1,13	0,97	1,09

Fuente: DANE

2- Estabilización de la varianza

En muchas situaciones las series presentan grandes variaciones en los datos, es decir, la varianza no permanece constante, por lo que es necesario estabilizarla. Para ello se utiliza una transformación; el método propone elegir una potencia λ para que la relación

$$\frac{\sigma_t}{\mu_t^{1-\lambda}} \text{ sea constante para } t = 1, 2, \dots, N \quad (2)$$

Donde σ_t es la desviación estándar, μ_t es la media y N es el número de observaciones que tiene la serie. El método para estabilizar la varianza de la serie, es el siguiente:

se dividen las N observaciones de la serie en H grupos, cada uno de los cuales contiene $R=(N-n)/H$ observaciones próximas, dejando por fuera de los cálculos un total de n observaciones ($0 < n < R$) ya sean del principio o del final de la serie. En cada grupo se calculan la media y la desviación estándar y de esta manera se tienen H parejas de valores

$\{S_h, \bar{Z}_h\}$, para $h = 1, \dots, H$ donde $\{S_h, \bar{Z}_h\}$ son las estimaciones de σ_t y μ_t respectivamente, con los cuales se construye un coeficiente de variación para cada potencia λ , como se observa en el cuadro 2.

Cuadro 2
Método de selección de la potencia para estabilizar la varianza con diferentes valores de λ por grupo

Grupo	λ				
	-1	-0.5	0	0.5	1
1	$S_1/\bar{Z}_1^2$	$S_1/\bar{Z}_1^{1.5}$	$S_1/\bar{Z}_1$	$S_1/\bar{Z}_1^{0.5}$	S_1
2	$S_2/\bar{Z}_2^2$	$S_2/\bar{Z}_2^{1.5}$	$S_2/\bar{Z}_2$	$S_2/\bar{Z}_2^{0.5}$	S_2
...	...	...	...	...	...
h	$S_h/\bar{Z}_h^2$	$S_h/\bar{Z}_h^{1.5}$	$S_h/\bar{Z}_h$	$S_h/\bar{Z}_h^{0.5}$	S_h
...	...	...	...	...	...
H	$S_H/\bar{Z}_H^2$	$S_H/\bar{Z}_H^{1.5}$	$S_H/\bar{Z}_H$	$S_H/\bar{Z}_H^{0.5}$	S_H
CV	$CV(-1)$	$CV(-0.5)$	$CV(0)$	$CV(0.5)$	$CV(1)$

Se pretende que la relación (2) sea constante; entonces con los valores de la serie se elige la potencia que proporcione el mínimo coeficiente de variación y con esta potencia se hace la transformación a la serie, siendo la transformación de la siguiente manera:

$$T(Z_t) = \begin{cases} Z_t^\lambda & \text{si } \lambda \neq 0 \\ \log(Z_t) & \text{si } \lambda = 0 \end{cases}$$

En la práctica la transformación más utilizada es la logarítmica.

Luego de estabilizar la varianza de la serie, se procede a realizar la prueba de normalidad para poder construir los límites de control.

3 - Prueba de normalidad

Para determinar si la serie se distribuye de una forma normal, se utiliza la prueba de Shapiro Wilk. Esta prueba se realiza para cada línea de producción de todas las empresas a las cuales se pretende establecer los límites de control. Con base en este resultado y utilizando la regla empírica de que si un conjunto de datos tiene una distribución aproximadamente normal, se pueden utilizar las siguientes reglas prácticas para describir el conjunto de datos:

1. Aproximadamente el 68% de las observaciones están a una desviación estándar de su media (es decir, dentro del intervalo $(\bar{y} - s/\sqrt{n}, \bar{y} + s/\sqrt{n})$ para muestras o $(\mu_t - \sigma_t, \mu_t + \sigma_t)$ para poblaciones).
2. Aproximadamente el 95% de las observaciones están a dos desviaciones estándar de su media (es decir, dentro del intervalo $(\bar{y} - 2s/\sqrt{n}, \bar{y} + 2s/\sqrt{n})$ para muestras o $(\mu_t - 2\sigma_t, \mu_t + 2\sigma_t)$ para poblaciones).

3. Casi todas las observaciones quedan a tres desviaciones estándar de su media (esto es, dentro del intervalo $(\bar{y} - 3s/\sqrt{n}, \bar{y} + 3s/\sqrt{n})$ para muestras o $(\mu_t - 3\sigma_t, \mu_t + 3\sigma_t)$ para poblaciones).

Cuando los datos resultan ser aproximadamente normales, se construye el intervalo de control como se establece en el numeral dos de la regla empírica.

Para el caso de no-normalidad, por ejemplo si se presenta asimetría, la construcción de los límites de control se debe basar en el coeficiente de asimetría, especificando el extremo en el cual se presenta dicha asimetría para poder hacer mayor el corte. Los intervalos en esta situación se establecen de la siguiente manera:

Si el coeficiente de asimetría es positivo, el intervalo es:

$$(\bar{y} - 1.5S/\sqrt{n}, \bar{y} + 2S/\sqrt{n})$$

Si el coeficiente de asimetría es negativo, el intervalo es:

$$(\bar{y} - 2S/\sqrt{n}, \bar{y} + 1.5S/\sqrt{n})$$

El proceso anterior se comporta bien si las series no son estacionales; en el caso de que la serie sea estacional, se hace un ajuste para controlar las variaciones que se originan a causa de las alteraciones estacionales. Como por ejemplo el aumento de las ventas en diciembre para algunas líneas de producción, o en enero en las ventas de artículos escolares, etc.. Este tipo de series se presenta con mucha frecuencia en comercio y si no se les realiza un ajuste, los intervalos serán inexactos en algunos períodos de tiempo.

4 - Ajuste estacional

Una serie estacional es aquella que muestra variaciones que se repiten en períodos iguales de tiempo.

Una característica de las series estacionales con observaciones mensuales, es la similitud de comportamiento entre las observaciones del mismo mes en años consecutivos. De igual modo, si la serie estacional es de observaciones trimestrales, existirán semejanzas entre los mismos trimestres para años consecutivos. Por lo anterior, conviene hablar genéricamente de un período estacional que comprenda E observaciones contiguas, donde E no es necesariamente igual a 12 para series mensuales o igual a 4 para series trimestrales.

En la muestra de comercio al por menor la información con la que se cuenta, aunque es muy poca, puede dar una idea acerca de la estacionalidad de las series. La idea es identificar los períodos en los cuales se

presentan variaciones que son más grandes o más pequeñas a las variaciones normales y con base en estos resultados, se ajustan los rangos de variación. Esto permite, aunque de forma intuitiva, ajustar los intervalos a la evolución periódica de las variaciones que presentan las líneas de producción.

Para las series estacionales la construcción de los intervalos depende del período que se está considerando, como por ejemplo, si se van a construir los rangos para el mes de diciembre y este mes es estacional para la serie, la desviación estándar y la media que se utilizan son las que corresponden a este período que se espera pertenezcan a una clasificación de variaciones atípicas. Esto genera intervalos acordes con el comportamiento de la variación.

III- Consideraciones adicionales para los rangos de variación

Para el cálculo de los intervalos de variación, se deben tener en cuenta las siguientes restricciones sobre las variaciones que se van a controlar:

1 - Ceros en toda la serie

Cuando en la serie, a la cual se pretende establecer rangos de variación, todos los datos son cero, la variación en cada período se define como uno, en este caso no existe desviación estándar; por lo tanto no se le puede aplicar la metodología, así que se le fijan unos rangos de variación de uno, esto quiere decir que cualquier cambio, por pequeño que sea, se detecta.

2 - Datos constantes a través de toda la serie

Si los datos en la serie son constantes, su variación es igual a uno; en este caso se adopta la misma metodología anterior.

3 - Ceros en último trimestre de la serie

Si hay presencia de ceros en el último trimestre de la serie, se presentarán variaciones extremas, las cuales afectan los límites de control haciéndolos completamente erróneos. La solución cuando no se presenta información en el último trimestre, es considerar que sus rangos de variación son uno, tal como en el primer caso.

4 - Datos faltantes o ceros dentro de la serie

Cuando hay datos faltantes o datos ceros en diferentes períodos en la serie, la solución propuesta es imputarlos

con la mediana de las demás observaciones y que sean diferentes de cero; esto elimina el error que introducen los datos en cero sobre las variaciones.

IV - Ejemplo

Para explicar la metodología, se procede a presentar un ejemplo de construcción de rangos para la variación de las variables de ventas de farmacéuticos, artículos de aseo personal y del total de ventas en los meses de marzo de 2000 a agosto de 2001 para el establecimiento con número de orden 20198.

Utilizando la base de datos en la cual se encuentran los datos históricos por establecimiento, se calculan las variaciones que han tenido las variables por establecimiento en los últimos 13 períodos; luego se calculan el promedio y la varianza de las variaciones de los períodos $t-13$ hasta $t-2$. Por ejemplo si se quiere el rango para el mes de agosto de 2001, se calcula el promedio de los meses julio de 2000 a junio de 2001, dentro del mismo establecimiento; se toman estas variaciones para evitar datos inconsistentes del período anterior, además para tomar, en caso de estacionalidad, los datos correspondientes al período del año anterior.

En el cuadro 1, se relacionaron los datos y las variaciones de las variables ventas de: farmacéuticos, artículos de aseo personal, y del total de ventas, dentro del establecimiento, datos que se utilizarán para la construcción de los rangos.

Se procede a calcular el coeficiente de asimetría, para definir los límites de control.

- Si el coeficiente de asimetría es cero o cercano a cero, los límites de control, son:

(promedio - 2 * desviación , promedio + 2 * desviación)

- Si el coeficiente de asimetría es positivo los límites de control, son:

(promedio - 1,5 * desviación , promedio + 2 * desviación)

- Si el coeficiente de asimetría es negativo, los límites de control, son:

(promedio - 2 * desviación , promedio + 1,5* desviación)

donde desviación es la raíz cuadrada de la varianza.

1 - Rangos de variación para la variable ventas de farmacéuticos

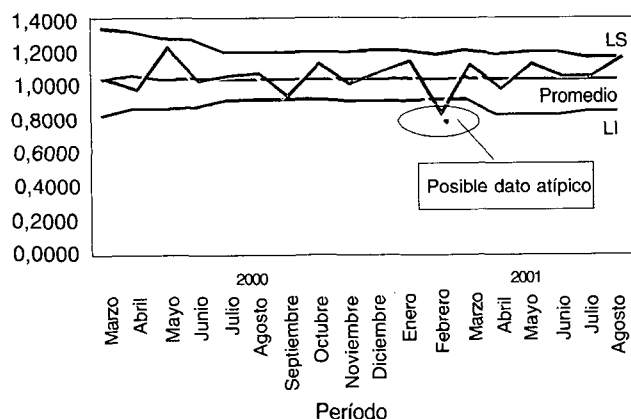
se muestra en la gráfica 1 la serie de las variaciones y los límites de control.

Los rangos obtenidos para la variación de la variable venta de farmacéuticos, se presentan en el cuadro 3, y

Cuadro 3
Rangos para la variación de la variable venta de farmacéuticos
2000 - 2001 (agosto)

Período	Coefficientes de asimetría	Promedio	Varianza	Límite inferior	Variación	Límite superior
2000	Enero				0,9697	
	Febrero				0,9925	
	Marzo	0,5763	1,0185	0,0219	0,7967	1,3143
	Abril	1,1609	1,0367	0,0162	0,8460	1,2909
	Mayo	1,7108	1,0206	0,0138	0,8445	1,2553
	Junio	2,218	1,0278	0,0119	0,8643	1,2457
	Julio	1,3843	1,0160	0,0058	0,9014	1,1688
	Agosto	1,4402	1,0194	0,0055	0,9084	1,1673
	Septiembre	1,3929	1,0203	0,0055	0,9092	1,1684
	Octubre	1,3482	1,0210	0,0055	0,9096	1,1695
	Noviembre	1,0156	1,0168	0,0063	0,8976	1,1758
	Diciembre	0,7865	1,0218	0,0069	0,8975	1,1876
2001	Enero	0,777	1,0235	0,0067	0,9009	1,1870
	Febrero	0,9642	1,0167	0,0055	0,9053	1,1653
	Marzo	0,5275	1,0283	0,0059	0,9126	1,1826
	Abril	-0,37	1,0129	0,0100	0,8127	1,1630
	Mayo	-0,564	1,0203	0,0105	0,8157	1,1738
	Junio	-0,601	1,0216	0,0103	0,8187	1,1737
	Julio	-1,106	1,0145	0,0084	0,8317	1,1517
	Agosto	-1,236	1,0176	0,0083	0,8354	1,1542

Gráfica 1
Rangos para la variación de la variable venta de farmacéuticos
2000 (marzo - diciembre) - 2001 (agosto)



Del cuadro 3 y de la gráfica 1, se puede establecer que las variaciones se encuentran entre un máximo de 1,1861 y un mínimo de 0,8072; también se aprecia que la media de las variaciones se encuentra alrededor de 1, y la varianza de las variaciones es constante a través del tiempo. La serie no presenta tendencia y además no presenta estacionalidad, por lo que se puede considerar la serie como estacionaria.

En la gráfica se observa que la variación correspondiente a febrero está fuera de los límites de control, por lo que se considera como un dato atípico y se deja a consideración del grupo temático evaluar si efectivamente lo es o no.

2 - Rangos de variación para la variable venta de artículos de aseo personal

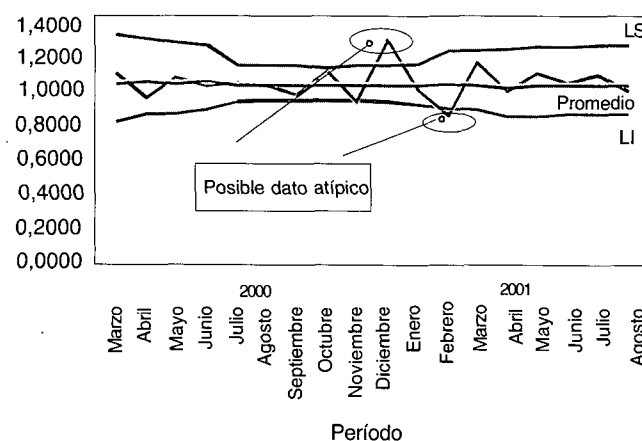
Los rangos obtenidos para la variación de la variable venta de artículos de aseo personal, se presentan en el cuadro 4 y se muestran en la gráfica 2, la serie de variaciones y los límites de control.

Para esta variable se puede establecer de la tabla 4 y de la gráfica 2 que las variaciones se encuentran entre 0,8156 y 1,2714. También se advierte que la media de las variaciones se encuentran alrededor de 1, la varianza de las variaciones es constante a través del tiempo, la serie no presenta tendencia ni estacionalidad, por lo que se puede considerar la serie como estacionaria.

Cuadro 4
Rangos para la variación de la variable venta de artículos de aseo personal
2000 - 2001 (agosto)

Período	Coefficientes de asimetría	Promedio	Varianza	Límite inferior	Variación	Límite superior
2000						
Enero					1,0027	
Febrero					0,9674	
Marzo	0,5209	1,0213	0,0217	0,8005	1,0766	1,3157
Abril	1,1477	1,0373	0,0161	0,8469	0,9331	1,2912
Mayo	1,476	1,0276	0,0139	0,8505	1,0482	1,2638
Junio	1,9332	1,0338	0,0121	0,8684	1,0013	1,2542
Julio	0,6903	1,0105	0,0037	0,9188	1,0167	1,1328
Agosto	0,6671	1,0150	0,0034	0,9279	1,0075	1,1311
Septiembre	0,6792	1,0148	0,0034	0,9277	0,9462	1,1308
Octubre	0,8261	1,0125	0,0033	0,9259	1,0852	1,1280
Noviembre	0,7774	1,0117	0,0034	0,9237	0,9003	1,1290
Diciembre	0,6121	1,0154	0,0038	0,9225	1,2714	1,1394
2001						
Enero	0,2551	1,0104	0,0048	0,9070	0,9822	1,1483
Febrero	1,5822	1,0214	0,0093	0,8767	0,8156	1,2143
Marzo	1,613	1,0197	0,0094	0,8742	1,1300	1,2137
Abril	0,7936	1,0070	0,0128	0,8375	0,9614	1,2330
Mayo	0,6822	1,0115	0,0137	0,8360	1,0755	1,2454
Junio	0,6599	1,0138	0,0133	0,8405	1,0135	1,2449
Julio	0,586	1,0161	0,0136	0,8413	1,0587	1,2492
Agosto	0,556	1,0171	0,0136	0,8424	0,9660	1,2500

Gráfica 2
Rangos para la variación de la variable venta de
artículos de aseo personal
2000 (marzo - diciembre) - 2001 (agosto)



En la gráfica 2, se observa que la variación correspondiente a diciembre se encuentra fuera de los límites de control; por tal razón se supone como dato atípico, pero considerando el mes en que se presenta esta variación, se deja al equipo temático determinar su calidad. La variación de febrero está fuera de los límites de control considerándola como un dato atípico e igualmente se deja a criterio del grupo temático tomarlo o no como un dato atípico.

3 - Rangos de variación para la variable total de ventas

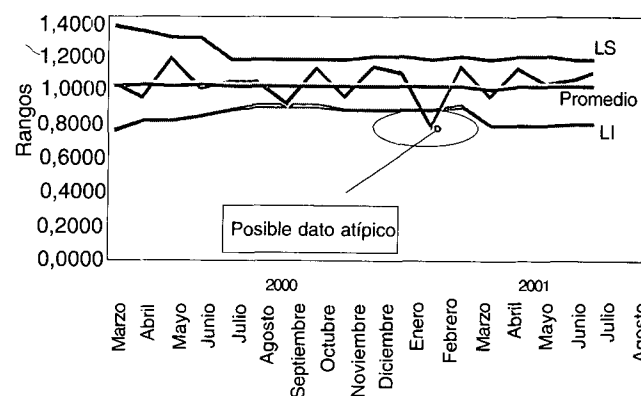
Los rangos obtenidos para la variación de la variable total de ventas se presentan en el cuadro 5 y se muestran en la gráfica 3 el dato real y los límites del rango.

La variable total de ventas resulta de la suma de las ventas de los productos que se comercian en cada establecimiento y la serie tiene el mismo comportamiento de las dos variables analizadas anteriormente, presentando las variaciones entre 0,8093 y 1,1470 con la media alrededor de uno, la varianza constante, sin presencia de tendencia y sin estacionalidad, por lo que se puede considerar la serie como estacionaria.

Las dos series analizadas anteriormente para las variables venta de farmacéuticos y venta de artículos de aseo personal, en febrero presentan variaciones fuera de los límites de control, y como la variable total de ventas es la suma de las ventas por establecimiento en cada período, esto puede estar causando que se presente la misma situación, por lo que se puede considerar esta variación como un dato atípico.

En conclusión, el grupo temático evaluará si los datos para el mes de febrero en el establecimiento, son datos atípicos.

Gráfica 3
Rangos para la variación de la variable venta de
artículos de aseo personal
2000 (marzo - diciembre) - 2001 (agosto)



Cuadro 5
Rangos para la variación de la variable total de ventas
2000 - 2001 (agosto)

Período		Coeficientes de asimetría	Promedio	Varianza	Límite inferior	Variación	Límite superior
2000	Enero					0,9786	
	Febrero					0,9855	
	Marzo	0,5628	1,0193	0,0218	0,7978	1,0206	1,3145
	Abril	1,1637	1,0368	0,0161	0,8464	0,9421	1,2907
	Mayo	1,6702	1,0225	0,0137	0,8469	1,1470	1,2565
	Junio	2,1789	1,0294	0,0118	0,8663	0,9917	1,2468
	Julio	1,0726	1,0143	0,0048	0,9109	1,0264	1,1523
	Agosto	1,1181	1,0180	0,0044	0,9186	1,0341	1,1504
	Septiembre	1,0825	1,0186	0,0044	0,9192	0,9159	1,1512
	Octubre	1,0846	1,0186	0,0044	0,9192	1,0963	1,1511
	Noviembre	0,8089	1,0152	0,0050	0,9094	0,9606	1,1564
	Diciembre	0,6089	1,0199	0,0055	0,9087	1,1093	1,1682
2001	Enero	0,6086	1,0198	0,0055	0,9086	1,0739	1,1682
	Febrero	0,5348	1,0173	0,0049	0,9120	0,8093	1,1577
	Marzo	0,189	1,0253	0,0050	0,9191	1,0997	1,1669
	Abril	-0,66	1,0106	0,0089	0,8222	0,9606	1,1519
	Mayo	-0,753	1,0172	0,0095	0,8219	1,0948	1,1637
	Junio	-0,808	1,0187	0,0093	0,8257	1,0219	1,1635
	Julio	-1,025	1,0144	0,0083	0,8319	1,0409	1,1513
	Agosto	-1,131	1,0169	0,0083	0,8349	1,0894	1,1534

BIBLIOGRAFÍA

GUERRERO, V. Análisis estadístico de series de tiempo económicas. México. 1991.

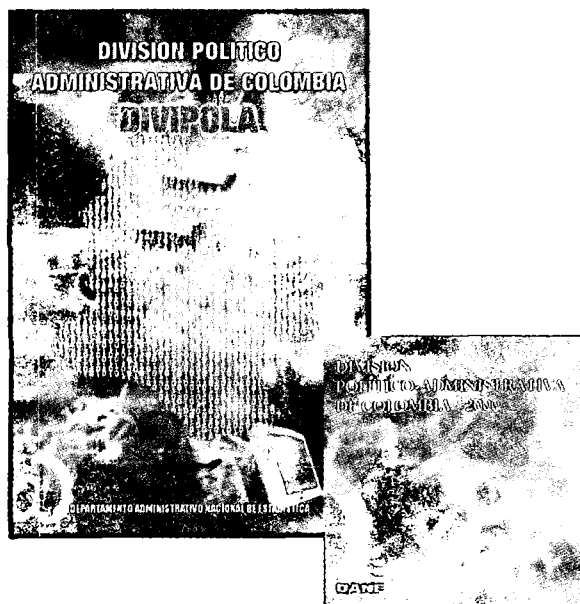
MENDENHALL, W. Probabilidad y Estadística para Ingeniería y Ciencias. USA. 1997

LOZANO, A. Estimación de novedades en estado de deuda. DANE. 2000.

PUBLICACIONES RECIENTES

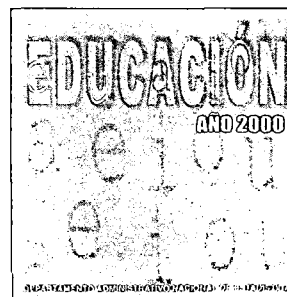
DIVIPOLA 2000

Disponible en publicación impresa
y CD ROM

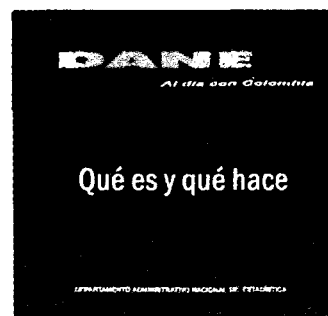


EDUCACIÓN 2000

Disponible en CD ROM

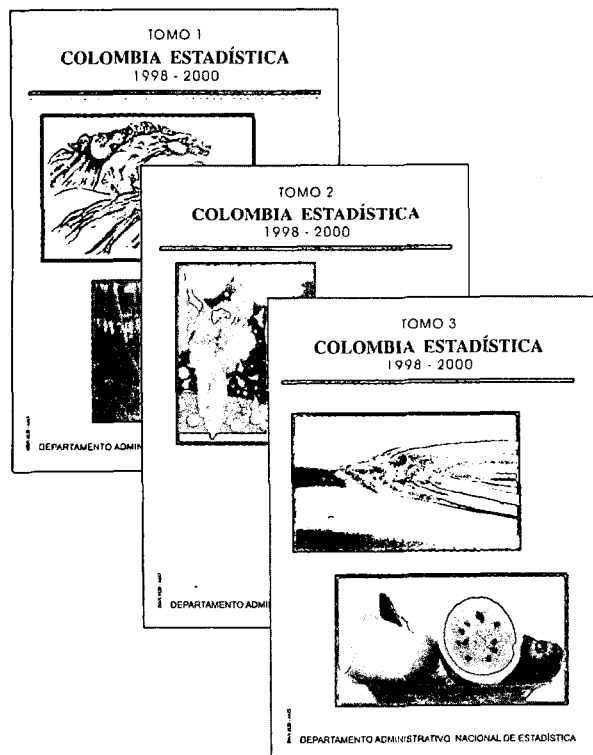


El DANE, qué es y qué hace
Disponible en publicación impresa
y CD ROM

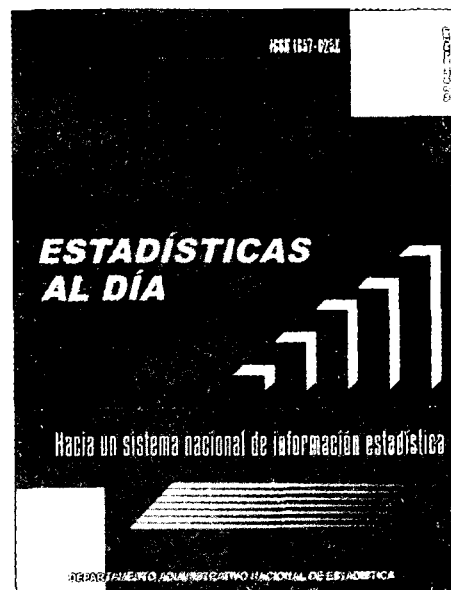


COLOMBIA ESTADÍSTICA 1998 - 2000

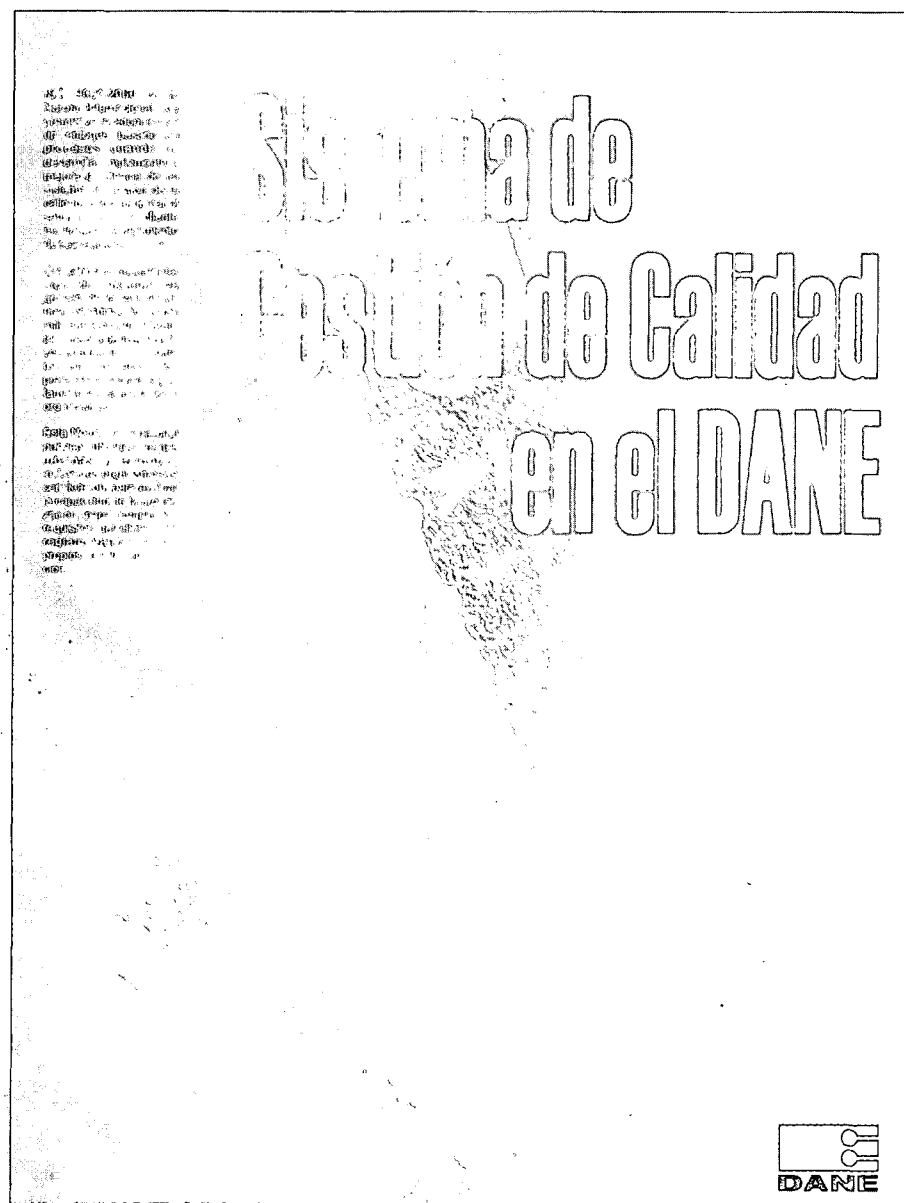
Disponible en publicación impresa
y CD ROM



ESTADÍSTICAS AL DÍA No. 2
Disponible en publicación impresa



PUBLICACIONES RECIENTES



Impreso en la Dirección de Mercadeo y Ediciones
Departamento Administrativo Nacional de Estadística - DANE
Bogotá, D.C., marzo de 2002