PROCEDIMIENTOS BASICOS EN EL REGISTRO Y PROCESO DE DATOS

Miguel A. Guzmán[1], *Ricardo Sibrián*[2] *y Rafael Flores*[2]

Instituto de Nutrición de Centro América y Panamá (INCAP), Guatemala, Guatemala, C.A.

RESUMEN

Se describen los principios generales que deben aplicarse como marco de referencia para identificar procedimientos apropiados en el manejo de datos recolectados en estudios de gran escala.

La descripción se hace en función de dos aspectos fundamentales: 1) el registro de datos y 2) el proceso de datos. En la primera parte (registro de datos), se definen las diferentes etapas a desarrollar en la preparación de formularios y/o cuestionarios, y en la recolección de datos. En la segunda parte

---

Manuscrito modificado recibido: 29–3–83.

1 En la época en que se llevó a cabo este trabajo, el Dr. Guzmán ocupaba el cargo de Jefe de la División de Estadística del Instituto de Nutrición de Centro América y Panamá (INCAP), Guatemala, C.A. Su título actual es: Professor, Department of Pathology and Department of Biometry, Louisiana State University Medical Center, Perdido St., New Orleans, La., 70112, U.S.A.

2 Estadísticos, División de Estadística del INCAP, Apartado Postal 1188, Guatemala, Guatemala, C.A.

Publicación INCAP E-1109.

(proceso de datos) se delinean las etapas a ser implementadas para lograr el control de calidad, realizar la codificación, y efectuar el ingreso de datos para su análisis. Con relación a este último aspecto, se propone un esquema de trabajo continuo por etapas, y se identifican algunos de los métodos estadísticos de uso común al someter a prueba diversas hipótesis.

El proceso se ilustra usando diagramas de flujo para algunas fases del proceso, en los que se señala en particular que el proceso y análisis de datos deben constituir una acción continua de principio a fin. Este debe definirse cuidadosamente en un protocolo de operaciones que norme todas las acciones del proceso.

INTRODUCCION

Los datos constituyen el nexo entre el diseño experimental y los resultados de una investigación. Por lo tanto, los aspectos relacionados con el registro y proceso de datos ameritan cuidadosa atención.

Los conocimientos requeridos para el manejo y análisis de datos sugieren que especialistas en estos campos sean incluidos en el equipo de la investigación, participando desde un principio en la planificación, ejecución, y evaluación de resultados.

En este trabajo se describen conceptos básicos para el registro y proceso de datos, los cuales pueden usarse como marco de referencia para el establecimiento de procedimientos adecuados en investigaciones específicas. La Figura 1 constituye un resumen de las etapas que se discuten seguidamente.

REGISTRO DE DATOS

1. Objetivo

Su objetivo es almacenar temporal o permanentemente la información recolectada. Las hipótesis y objetivos de la investigación definen la naturaleza de los datos a recolectar, pero en todo caso, hay necesidad de elaborar formularios y/o cuestionarios para anotar esa información. Es preferible el uso de formularios precodificados ya que permite el registro directo de información con mayor rapidez y minimiza posibles fuentes de error.

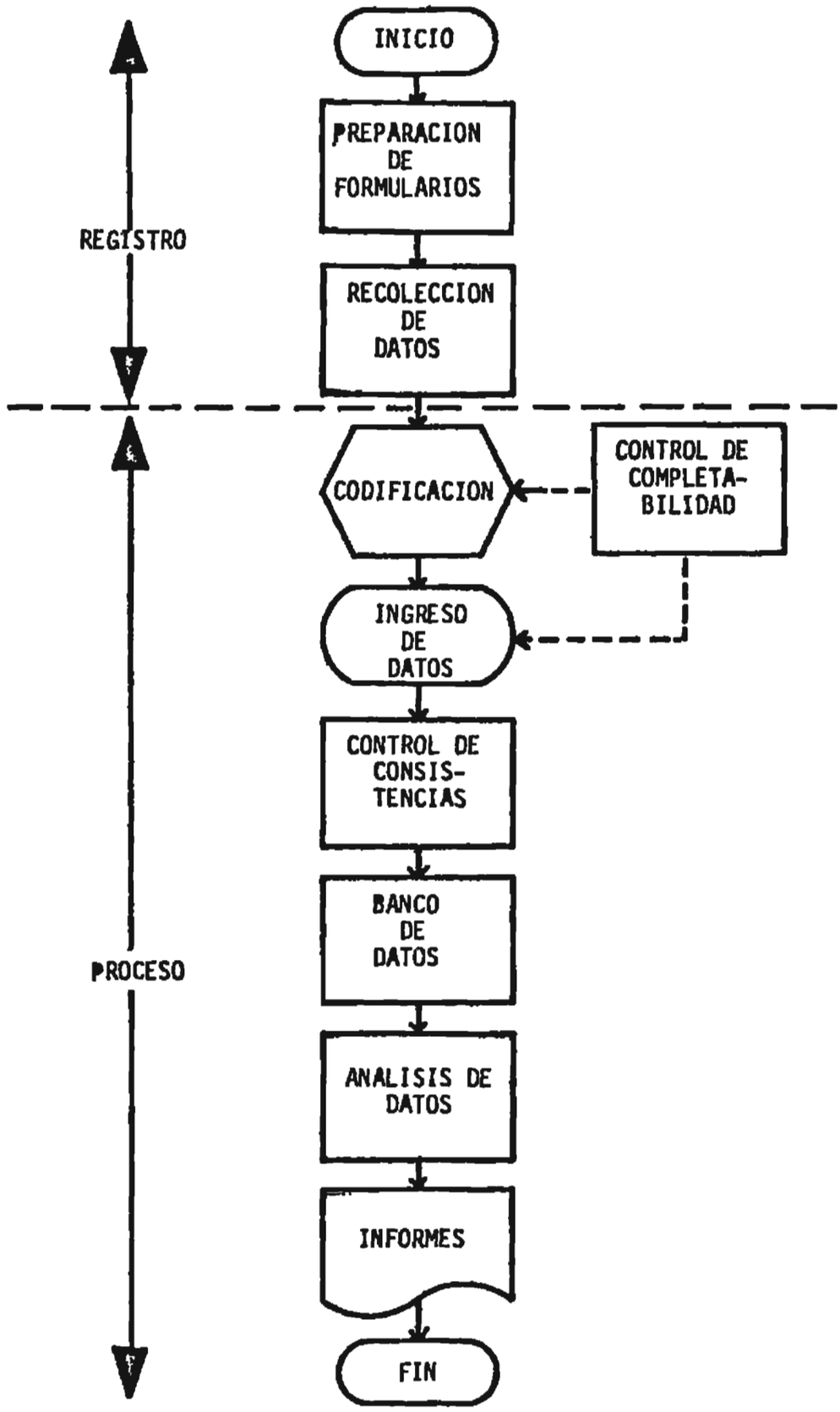

FIGURA 1

Etapas del registro y proceso de datos

2. *La Preparación de Formularios*

Esta consiste en producir en su versión final, todos los formularios y/o cuestionarios que servirán para recolectar los datos y que, en general, constan de encabezado y cuerpo. El encabezado contiene información útil para el personal encargado del proceso de datos, y el cuerpo, espacio para el registro de datos pertinentes a cada una de las diferentes áreas de la investigación. Las fases a cubrir en esta etapa generalmente se refieren al diseño de formularios, preparación de instructivos y manuales de codificación, pre-pruebas y revisiones. En este proceso existe opción para el entrenamiento y estandarización del personal que se encargará de recolectar los datos. Para determinar la confiabilidad de los formularios, así como para la estandarización del personal, se requiere el uso de técnicas estadísticas apropiadas (1-4). Una muestra de 20 a 30 formularios suele ser generalmente suficiente para este propósito.

3. *La Recolección de Datos*

Esta etapa puede iniciarse cuando los investigadores estén satisfechos con el nivel de entrenamiento y estandarización del personal encargado de la recolección de datos y con la versión final de los formularios y/o cuestionarios. Debe integrarse en un manual de operaciones de la investigación (MOI) (2) una descripción de cómo anotar la información, y de las técnicas y procedimientos a seguir para la recolección de datos. Varios son los errores que pueden surgir en esta etapa de recolección, los cuales es factible que alteren la interpretación de los resultados. Es usual que tales errores estén asociados con omisiones al llenar los formularios cuando se efectúan las entrevistas, ausencia de datos, mala interpretación por parte del entrevistado, desconocimiento de la respuesta por parte del mismo, mala interpretación intencionada por parte del entrevistado y errores del propio entrevistador (5). Con miras a controlar la calidad de los datos a este nivel, se requiere de una buena supervisión en el lugar de la recolección (6).

PROCESAMIENTO DE DATOS

El procesamiento de datos es el tratamiento de la información después de recolectada. En estudios pequeños este tratamiento es básicamente manual, mientras que en aquéllos de mayor

escala se requiere del uso de medios electrónicos. Esta etapa generalmente exige atención en los siguientes rubros.

1. *Codificación*

La codificación se puede iniciar aun antes de empezar la recolección de datos. Por ejemplo, es factible pre-codificar algunos rubros en el encabezado de los formularios, usando para ello las facilidades de computación. Sin embargo, y en general, los formularios y/o cuestionarios son codificados después de recolectar los datos. En este caso es aconsejable que la codificación se realice tan pronto como sea posible, de preferencia el mismo día de la recolección.

2. *Entrada de Datos*

Esta operación puede variar desde la tradicional tarjeta perforada de 80 columnas hasta la entrada directa desde el aparato de medición. Cuando la recolección y el proceso de datos se lleva a cabo en el mismo lugar, es factible establecer sistemas automáticos para la entrada de datos directamente en línea. Cuando la investigación requiere encuestas de gran envergadura y el origen de los datos es remoto al procesamiento, un sistema de perforación-cinta, sería mucho más apropiado que un sistema interactivo.

3. *Control de Calidad de los Datos*

Este rubro comprende dos aspectos: la completabilidad y la consistencia. La completabilidad se refiere al hecho de verificar que cada pregunta haya sido respondida y que los elementos de identificación del formulario (encabezado) estén completos. En general, aunque no necesariamente, la completabilidad se ejecuta después de la codificación. Esta revisión puede incluir procedimientos simples tales como el control de mínimos y máximos para cada elemento o procedimientos complejos como la revisión contra un inventario detallado de combinaciones correctas de códigos para unidades primarias, secundarias y terciarias de acuerdo a su composición. La revisión de completabilidad para el cuerpo del formulario se refiere tanto a la ausencia como a la presencia de respuestas cuando estas situaciones proceden. Especial atención debe darse a situaciones en las que existen conjuntos de respuestas relacionadas entre sí. Cuando la revisión de completabilidad se desarrolla inmediatamente después de la recolección de datos, es posible

recuperar bastante información; en tal sentido esta alternativa deberá especificarse como una acción rutinaria dentro del MOI.

La consistencia concierne a la detección de contradicciones originadas en los datos. Los procedimientos para aplicar este control usualmente se relacionan con rangos admisibles, relaciones aritméticas, lógicas y especiales entre dos o más respuestas. Los controles de rangos admisibles se aplican con mayor frecuencia a datos básicos que a datos derivados. En general, los controles de relaciones aritméticas se aplican a respuestas de naturaleza continua. Estas relaciones comprenden: a) valores de respuestas mayores en una pregunta con respecto a otra ($A > B$); b) valores de respuestas mayores o iguales de una pregunta con respecto a otra ($A \geq B$); y c), valores de respuestas iguales ($A = B$). En la Tabla 1 se proporciona un resumen de algunas relaciones aritméticas. Los

TABLA 1

ESQUEMA Y EJEMPLOS DE DIFERENTES CONTROLES ARITMETICOS Y LOGICOS DE DATOS

ARITMETICOS.

1) $A > B$ 2) $A \geq B$ 3) $A = B$

LOGICOS Entre dos respuestas y entre más de dos respuestas

Prototipo de una vía:	Prototipo de doble vía:
Si A = x entonces B = y	A = x si y sólo si B = y
a. Si a = 1 entonces B = 1, 3, 5	e. A = 1 si y sólo si B = 2, 5, 7
b. Si A = 2, 3 entonces B = 4, 6	f. A = 1, 3, 6 si y sólo si B = 3-18
c. Si A = 1-10 entonces B = 3-18	g. A = 1-10 si y sólo si B = 5
d. Si A = 1-14, 8-12 entonces B = 2-12	h. A = 1-4 si y sólo si B = 5-7, 15-30
i. Si A = 1 y B = 2 entonces C = 5, 10, 11	l. A = 1 y B = 2 si y sólo si C = 5, 10, 11
j. Si A = 1-5 y B = 10-20 entonces C = 2-8	m. A = 1-5 y B = 10-20 si y sólo si C = 2-8
k. Si A = 1-5 y B = 1-5 entonces D = 5-10	

Nota: A, B, C y D son respuestas que están relacionadas.

controles de relaciones lógicas se fundamentan en la dependencia de una respuesta con respecto a otra u otras; por ejemplo, el peso corporal depende del sexo y la edad, por lo que un sujeto de sexo masculino de 30 meses de edad no podrá tener un peso fuera de determinados valores. Se pueden implementar también otros controles más sofisticados, tales como el control de rutas de respuestas, basado en la construcción de "árboles" que describen "rutas" de respuestas consideradas como aceptables. En la misma Tabla se incluyen algunos ejemplos de relaciones lógicas.

Al encontrar inconsistencias internas, surgen las siguientes alternativas: a) revisión de datos originales con el fin de decidir su recuperación y aceptación o rechazo, b) eliminación automática del dato en cuestión, c) eliminación automática del dato en cuestión y revisión con respecto a la decisión de otros datos relacionados, y d) eliminación completa del formulario. Es necesario elaborar un catálogo de errores a fin de facilitar la toma de decisión con respecto a cada error identificado.

4. *Banco de Datos*

El banco de datos está formado por el conjunto de archivos limpios, adecuadamente documentados y generados en la etapa anterior. A partir del banco de datos se puede crear un archivo maestro uniendo los diferentes archivos en base a sus diferentes claves de identificación (encabezado). Una vez formado el archivo maestro, debidamente documentado, se puede dar inicio a la etapa de análisis. En sistemas de computación con capacidad para el manejo de bases de datos, el banco de datos es la fuente original para crear la estructura de la base (7). La etapa de banco de datos no es estática; por el contrario, es muy dinámica y requiere acción continua mientras no se termine el análisis y la interpretación de los datos.

5. *Análisis de Datos*

Relaciona la información y las hipótesis de la investigación. Una descripción básica de cada una de las variables continuas en términos de promedios, desviación estándar, mínimos y máximos y número de casos, resulta muy informativa, mientras que para variables discretas, son más útiles las tablas de frecuencia para cada categoría. Con esta descripción básica se corroboran los procedimientos de limpieza de la etapa de control de calidad, y al mismo tiempo, se obtiene un patrón preliminar del comportamiento de

las variables. Una vez asegurada la limpieza, es factible iniciar el análisis de datos con las pruebas de las hipótesis, utilizando las pruebas de "t" de Student y el análisis de varianza (prueba de F) para comparaciones entre dos y más clases, respectivamente (8-10), asumiendo que los supuestos requeridos para estas técnicas han sido satisfechos. Las tendencias y asociaciones entre dos y más variables, se pueden examinar a través del análisis de regresión y de correlación simple y múltiple (11-13). La clasificación e identificación de grupos de observaciones pueden obtenerse a través de técnicas de análisis discriminatorio y de agrupación (14-18), mientras que la identificación de interrelaciones entre variables de diferente naturaleza puede examinarse por medio del análisis de factores (19). Finalmente, es factible realizar el estudio de interrelaciones entre varias variables dependientes e independientes valiéndose de técnicas de análisis canónico (17). En los casos en que no sea posible satisfacer los supuestos para las anteriores pruebas, se utilizarán técnicas no paramétricas (20-22).

Es preciso mencionar aquí que para el análisis de datos, si ello es factible, es muy conveniente utilizar paquetes estadísticos tales como: SAS, BMDP y SPSS, ya que los mismos aumentan la productividad dada la gran cantidad de técnicas disponibles para el análisis y la velocidad de ejecución. Si estos paquetes están disponibles, los recursos de programación son mínimos y el manejo de la información para pruebas de hipótesis es un proceso relativamente sencillo. En la Figura 2 se ilustra gráficamente un resumen de las acciones por lo general requeridas para el análisis de datos (23).

6. *Elaboración de Informes*

Los informes se elaboran en términos generales para toda la investigación, y en forma separada para cada una de las áreas sometidas a estudio.

SUMMARY

BASIC PROCEDURES IN THE DATA RECORDING AND PROCESSING

This paper describes the general principles that must be used as a frame of reference to identify the appropriate procedures for the management of

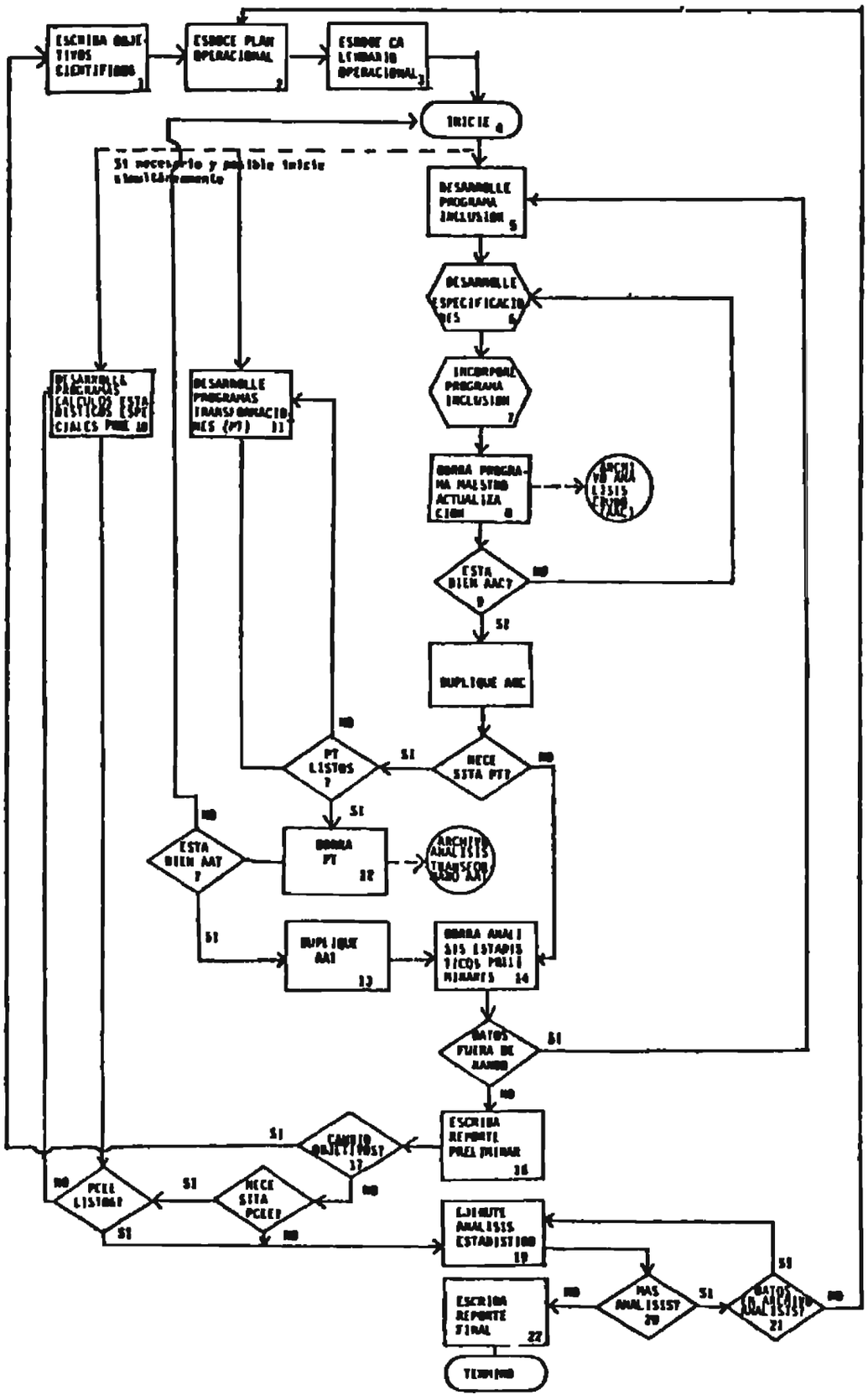

FIGURA 2

Diagrama de flujo para las acciones a ejecutar en el análisis de datos

data collected in large scale studies.

It has two parts: 1) Data recording and, 2) data processing. In the first part (data recording), the various steps required for the preparation of forms and/or questionnaires and for the recollection of data are discussed. In the second part (data processing), the steps to achieve real quality control, proper coding and data entry for analysis are described. For the latter case, an outline of continuous general steps is proposed, and some statistical procedures commonly used for testing different hypotheses are identified.

The overall process is illustrated through flow charts, emphasizing the fact that the processing and analysis of data, must be a continuous action which requires careful definition in the standard protocol of operations.

BIBLIOGRAFIA

1 Selltiz, C., S. Wrightsman & S.W Cook **Research Methods in Social Relations.** 3rd ed. New York, N.Y., Holt, Rhinehart and Winston, 1976.

2. Guzmán, M A , C.A. McMaham, H C McGill, J.P. Strong, C. Tejada, C Restrepo, D.A Eggen, W.B Robertson & L A Solberg Selected methodologic aspects of the international atherosclerosis project. **Lab. Invest., 18**: 479-497, 1968.

3. Habitch, J-P. Estandarización de métodos epidemiológicos cuantitativos sobre el terreno **Bol. Of. San. Pan , 76.** 375-384, 1974.

4. Guzmán, M A. Some consideration in the design and execution of nutritional field studies. En **Amino Acid Fortification of Protein Foods.** N S. Scrimshaw and A.M. Altschul (Eds.). Cambridge, Mass., The MIT Press, 1971, p. 301-315

5 Feber, R , D.P. Sheatsley, A Turner & J Waksberg **What is a Survey?** Washington, D C , American Statistical Association, 1980.

6 Cosley, D J & D.A. Luny. **Data Collection in Developing Countries.** New York, N.Y., Oxford University Press, 1981.

7. Burch, J G & F R Strater **Information Systems Theory and Practice.** New York, N.Y., John Wiley & Sons Inc , 1974.

8. Scheffe, H. **The Analysis of Variance.** New York, N.Y., John Wiley & Sons Inc , 1969.

9. Snedecor, G.W & W G Cochran. **Statistical Methods.** 6th ed. Ames, Iowa, The Iowa State University Press, 1967.

10. Eisenhart, C. The assumption underlying the analysis of variance **Biometrics, 3.** 1-21, 1947.

11. Searle, S.R. **Linear Models.** New York, N.Y., John Wiley & Sons Inc., 1971.

12. Graybill, F. A **An Introduction to Linear Statistical Models.** Vol. 1. New York, N. Y., McGraw-Hill Book Company, 1961.
13. Neter, J. & W. Wasserman. **Applied Linear Statistical Models.** Homewood, Ill., Richard D. Irwin, Inc., 1974.
14. Hartigan, J. A. **Clustering Algorithms.** New York, N. Y., John Wiley & Sons Inc., 1975.
15. Tatsuoka, M. M. **Discriminant Analysis.** Champaign, Ill., Institute for Personality and Ability Testing, 1970.
16. Timm, N. M. **Multivariate Analysis with Application in Education and Psychology.** Belmont, Ca., Wadsworth, 1975.
17. Green, P. E. **Analyzing Multivariate Data.** Hinsdale, Ill., The Dryden Press, 1978.
18. Morrison, D. F. **Multivariate Statistical Methods.** 2nd ed. New York, N. Y., McGraw-Hill Book Company, 1976.
19. Lawley, D. N. & A. E. Maxwell. **Factor Analysis as a Statistical Method.** London, Butterworths, 1963.
20. Hollander, M. & D A. Wolfe. **Nonparametric Statistical Methods.** New York, N Y., John Wiley & Sons Inc., 1973.
21. Harshbarger, T. R. **Introductory Statistics.** 2nd ed New York, N. Y., MacMillan Publishing Co. Inc., 1977.
22. Marascuilo, L. A. & M. McSweeney. **Nonparametric and Distribution-free Methods for the Social Sciences.** Monterrey, CA., Brooks/Cole Publishing Co., 1977.
23. Helms, R. W. **Data Analysis Procedures for a Coordinating Center of a Large Collaborative Study.** Chapel Hill, N C., The Institute of Statistics, 1975 (Mimeo Series No. 1003).