CAPÍTULO VII
PROCESAMIENTO DE LOS DATOS

Este capítulo es para los coordinadores de encuesta y recursos humanos técnicos. Le proporciona orientación para ayudarle a:

- ✔ Realizar la entrada de datos.
- ✔ Verificar y editar los datos y crear un archivo "limpio" de los datos para análisis.
- ✔ Evaluar la calidad de los datos de la encuesta.
- ✔ Producir estimaciones para los indicadores.

INTRODUCCIÓN

El procesamiento de los datos debe tener lugar tan pronto como sea posible. Los primeros resultados de la encuesta, el informe preliminar, deben aparecer rápidamente, a las pocas semanas después de que termine el trabajo de campo. Un sesión entera de discusión con el equipo de la encuesta para la revisión de los resultados iniciales contribuye a un fuen informe final. El entusiasmo e interés generado por el trabajo de campo se evaporará rápidamente cuante éste termine, así que asegúrese de mantener la ventaja que le pueden dar los resultados de la encuesta al reportarlos rápidamente.

Como parte de los procedimientos de campo para reducir los errores, las entrevistadoras deben verificar los cuestionarios cuando completan la entrevista. Los errores se pueden corregir a menudo en el terreno, sin la necesidad de una reentrevista. Las supervisoras también deben revisar los cuestionarios inmediatamente después que los reciben de las entrevistadoras. ¿Están completos? ¿Se siguieron las instrucciones de "pase" correctamente? El administrador de la encuesta también debe revisar todos los cuestionarios en las primeras etapas del trabajo de campo y continuar revisando muestras de los cuestionarios a lo largo de la encuesta.

Lo mejor es empezar la entrada de datos mientras las entrevistadoras están todavía en el terreno. Esto permitirá detectar y corregir errores que ciertas entrevistadoras o equipos pueden estar cometiendo. Problemas serios que pueden escaparsele a la supervisora pueden ser reconocidos rápidamente, a tiempo para reentrenar el personal de campo y corregir errores graves. Las supervisoras de campo son responsables por revisar que los cuestionarios estén completos y que sean consistentes, y por clasificar las respuestas que las entrevistadoras no pudieron codificar. Excepto por la revisión del número de cuestionarios que llegan del terreno y por completar los números de identificación en los formularios, no se requiere que el personal de oficina critique o codifique los cuestionarios que llegan a la oficina.

El sistema de entrada de datos debe ser cuidadosamente diseñado y el personal de digitación debe estar bien entrenado y motivado. Los programas para entrada de datos, verificación y análisis deben ser instalados y verificados utilizando datos del estudio piloto antes

de que empiece la etapa principal del trabajo de campo. Cualquier problema debe ser limado y los programas modificados si fuese necesario, antes de que empiece a llegar información del terreno.

Cuatro son las principales etapas para el procesamiento computarizado de los cuestionarios:

1. entrada de datos
2. revisión y crítica de los datos
3. distribuciones de frecuencia
4. contrucción de variables y tabulaciones

Varios paquetes de computador se encuentran disponibles para realizar estas funciones. Un programa completo para entrada de datos, verificación y análisis para calcular todos los indicadores de las siete Metas de Mitad de Década (es decir, aquéllas que pueden medirse con una encuesta de indicadores múltiples) acompaña este manual para ser utilizado con el paquete EPI INFO (Versión 6). Los programas también se listan en el Apéndice 3.

Si se han hecho modificaciones al cuestionario, los programas de entrada y análisis de datos tendrán que ser modificados por un programador con excelente conocimiento del programa EPI INFO

EPI INFO fue desarrollado por los Centros para el Control y Prevención de Enfermedades en colaboración con la OMS (disponible de USD, Inc., 2075 A West Park Place, Stone Mountain, GA 30087, USA) y puede realizar todas las cuatro fases del procesamiento de datos. Una copia de EPI INFO (Versión 6) se incluye con este manual.

El supervisor de procesamiento de datos debe instalar el sistema de entrada de datos por adelantado. En el Apéndice 3 se encuentran consejos de como hacerlo.

ENTRADA DE LOS DATOS EN COMPUTADOR Y PRODUCCIÓN DE ARCHIVOS DE DATOS "LIMPIOS"

Entrada de Datos

La entrada de datos empieza tan pronto como llegan a la sede los datos de un conglomerado. La entrada de datos debe ejecutarse en lotes pequeños (e.g., conglomerado por conglomerado).

Los datos del cuestionario para cada hogar deben digitarse en el orden en que se recolectan: los datos del hogar de la página de listado, los datos para los niños del listado de niños, los módulos de agua y saneamiento y de yodización de la sal, datos de educación para cada niño en edad escolar, el módulo del toxoide tetánico para cada madre, el módulo de atención de enfermedades respiratorias agudas (AERA) para cada madre, y por último, los módulos de salud del niño para cada niño. Una vez se completa la entrada de datos para un hogar, el digitador pasa a entrar los datos para el próximo hogar.

Note que la información faltante durante el trabajo de campo se registra como nueves—por ejemplo 9, 99, 999, y así sucesivamente. Los campos que estén en blanco cuando se hace la entrada de datos y respuestas no aplicables (e.g., para respuestas a preguntas opcionales o para preguntas que deben ser saltadas) se almacenan como campos en blanco en la base de datos. A medida que se entran los datos, el programa realiza la *verificación de rangos*. Esto asegura que para un rubro no se ingresa un valor que está por fuera del rango dado para las respuestas a ese rubro.

✎ **EJEMPLO:**
La fecha de la entrevista debe registrarse durante el período de trabajo de campo. Una fecha registrada como 01 11 93 está por fuera del rango de fechas permitidas para esta encuesta. Si el programa de entrada de datos ha establecido el rango para esta variable, el computador reconocerá que esta entrada de fecha es inválida.

Hay dos enfoques al ingreso de datos. Los datos se pueden entrar dos veces al computador, por dos digitadores diferentes. Los dos archivos se comparan entonces para inconsistencias y también se realizan otros chequeos. Otro enfoque consiste en digitar los datos sólo una vez pero realizando un determinado número de chequeos de inconsistencias para prevenir errores durante la digitación. La verificación interactiva de datos hace más lento el proceso de entrada de datos, pero puede evitar muchos errores comunes que cometen los digitadores. Luego se corren los chequeos de inconsistencias y se corrige el archivo. Para estas encuestas de indicadores múltiples, recomendamos el último enfoque, el cual se utiliza en los programas que acompañan el manual.

Recuadro 7.1
VENTAJAS Y DESVENTAJAS DE LA DOBLE ENTRADA

Ventajas

Con doble entrada, los errores de entrada de datos prácticamente se eliminan si se hace correctamente. Esta opción, precedida por chequeo de errorres por lotes, puede resultar más barata y rápida si se utiliza personal profesional para la entrada de datos.

Desventajas

La doble entrada duplica el costo de la entrada de datos y complica el flujo de cuestionarios en la oficina, lo cual también puede llevar a errores.

Un posible término medio consiste en entrar algunos cuestionarios dos veces como una forma de verificar ahí mismo la exactitud del personal de entrada de datos y de identificar los errores sistemáticos que se están cometiendo.

Verificación de los Datos

La experiencia ha demostrado que los errores de rango son casi siempre errores de entrada de datos y pueden ser chequeados y corregidos inmediatamente durante la digitación. Los errores de consistencia, por otro lado, deben resolverse regresando a los hogares o mediante un examen cuidadoso de los cuestionarios. Es mejor entonces realizar la verificación de inconsistencias como una etapa separada, listando los errores en una copia dura para marcar las correcciones en los cuestionarios.

El programa EPI INFO de entrada de datos que viene con este manual, aplica automaticamente un gran número de chequeos a los datos, bien al alertar al digitador cada vez que se entra una respuesta por fuera de rango, o al informar las inconsistencias en el pase final por el programa de chequeo de datos.

☞ Una vez que los datos están en la sede, no es usualmente posible regresar a los hogares para corregir las respuestas inconsistentes. El coordinador del estudio o el principal supervisor tendrán que hacer los "mejores supuestos" a partir de la información existente para corregir el cuestionario. Si eso no es posible, la variable tendrá que dejarse como "faltante."

Una vez que se han ingresado los datos para todo un conglomerado, deben revisarse y editarse las inconsistencias del lote. El programa chequea para asegurar que todas las variables apropiadas para el caso están en el archivo (es decir, que la estructura del archivo es correcta) y que están dentro de los rangos correctos. El programa realiza chequeos internos de inconsistencia para asegurar que los datos han sido registrados y entrados correctamente. Las instrucciones de entrada de datos al final del Apéndice 3 discute los procedimientos a seguir cuando se presentan errores. La consulta del cuestionario original en este punto puede ayudar a resolver los problemas en los datos.

✎ **EJEMPLO:**

Los chequeos de consistencias verifican que se llenan los siguientes requisitos:

— Que la fecha de nacimiento de cada niño precede la fecha de la entrevista.

— Que las fechas de nacimientos de todos los niños deben estar dentro de los últimos 15 años.

— Que si la fuente de agua es "vivienda con tubería," entonces la distancia debe ser "en la vivienda."

— Que si un niño recibió "SÓLO pecho," entonces no deben aparecer respuestas afirmativas para otros rubros.

— Que la fecha de la primera DPT debe preceder la fecha de la segunda DPT y ésta debe preceder la tercera DPT.

— Que los índices antropométricos no son alertados como inválidos debido a errores de entrada de la edad o el peso.

Puede ser útil mirar el número de errores por equipo o entrevistadora, de tal manera que se pueda realizar más entrenamiento en las primeras fases del trabajo de campo.

Una vez que se han definido las correcciones para las inconsistencias reportadas, se actualizan los registros correspondientes en el computador. El programa de chequeo debe correrse de nuevo y se repite el proceso hasta que se han rectificado todos los errores que pueden corregirse en la oficina. Una vez que se completan estos chequeos, los datos quedan "limpios" y los archivos listos para ser analizados.

EVALUACIÓN DE LA CALIDAD DE LOS DATOS DE LA ENCUESTA

El próximo paso es preparar *la distribución de frecuencias* de todas las variables en el archivo de datos y los posibles valores que pueden tomar (véase el Cuadro 7.1). Esto le permite chequear por códigos inapropiados en los datos y notar el número de valores faltantes en los datos:

Cuadro 7.1 Ejemplo de una distribución de frecuencias

Rubro número	Fuente de suministro de agua para el hogar	Distribución
1	Agua de tubería	146
2	Pila pública	440
3	Grifo o pozo	102
4	Aljibe con protección	95
5	Aljibe sin protección	660
6	Estanque, rio, arroyo	35
7	Carro tanque, vendedor (aguatero)	55
9	Otra fuente	15
.	Sin información	6
Total		1554

Es posible que sea necesario que usted mire los cuestionarios originales para chequear los valores en los datos. Es posible que sea necesario corregir el archivo crudo de datos y correr de nuevo las frecuencias.

☞ Con relación al ejemplo en el Cuadro 7.1: Supongamos que usted desea examinar los 15 cuestionarios con "otra fuente" de agua. Usted necesitará programar el computador para seleccionar los registros que tienen el valor "9" (otra fuente) para la variable "fuente de agua." Usted tiene que pedirle al computador que liste los números de identificación de los hogares con esos valores. Usted puede usar los números de identificación para

examinar esos cuestionarios.

Usted debe llevar un registro de cuantos casos en el archivo estaban incompletos, cuantos hogares fueron listados pero no contactados y que errores se notaron en los datos. El porcentaje de hogares elegibles que no fueron contactados debe ser examinado e informado. Una tasa de respuesta por debajo de 90-95 por ciento significa que la encuesta no necesariamente representa la población total.

Una forma de chequear la calidad de los datos es examinar la consistencia interna de los datos recolectados. Los chequeos de inconsistencia corridos durante la entrada de datos pueden correrse de nuevo para toda la base de datos para determinar el número de registros que todavía tienen inconsistencias. (Véase el ejemplo de arriba sobre los requisitos de consistencia.)

Un número alto de inconsistencias sin resolver es un indicador de recolección floja de datos (y a veces falsificación de datos) y puede comprometer la validez de los resultados.

Un segundo indicador de la calidad de los datos es el porcentaje de respuestas con "no sabe" u "otro." Estos porcentajes se pueden obtener examinando la distribución de frecuencias para los rubros en la base datos. Un alto porcentaje de "no sabe" indica que las informantes no entendieron las preguntas o que la información solicitada es muy difícil de reportar. Los resultados basados en tales variables son cuestionables. Un porcentaje alto de "otras" respuestas indica que las categorías en el cuestionario no capturaron las respuestas más comunes que fueron dadas.

Otro indicador de calidad de los datos es que tan bien comparan con otros datos existentes. Por ejemplo, la distribución de niños y madres por edad y la razón entre niños y niñas deben ser similares a las distribuciones por edad y las razones de masculinidad provenientes de censos recientes u otras encuestas en la misma población. Aún en ausencia de encuestas o censos, el número de niños y niñas enumerados debe ser gruesamente el mismo. El número de niños en cada grupo de años simples también debe ser similar, con una ligera disminución del número con la edad.

CÁLCULO DE LOS INDICADORES

Los cuadros básicos para calcular cada indicador se listan en el Apéndice 4 y los programas para producir las estimaciones se presentan en el Apéndice 3. Todas las estimaciones se pueden producir con los programas que se encuentran en el disco magnético que acompaña este manual. De lo contrario, se puede utilizar un paquete estándar como el SPSS PC.

Usted puede obtener algunos de los indicadores directamente de las respuestas a las variables en el cuestionario—por ejemplo, el porcentaje de hogares con sal yodada. Para la mayoría de los indicadores, sin embargo, usted tendrá que examinar las respuestas a un número de preguntas para determinar la información que requiere el indicador. Por ejemplo, el acceso a agua potable se basa tanto en la fuente como en la distancia a esa fuente. En estos casos, debe crearse una variable que indica si el hogar, la madre *o* el niño satisfacen la definición del indicador.

☞ El programa EPI INFO que acompaña este manual tendrá que ser adaptado para informar sobre indicadores que varían de país a país—por ejemplo, agua y saneamiento.

La mayoría de los paquetes de análisis, incluyendo SPSS y EPI INFO, producirán un cuadro de información sobre las variables creadas para los indicadores. Para informar sobre el indicador, el número apropiado deberá ser extraído del cuadro.

✎ **EJEMPLO:**
Un cuadro sobre bajo peso para la edad puede ser generado, como sigue:

Rótulo del valor	Valor	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
≥ -2 DE	0	1,183	77.9	79.0	79.0
< -2 DE	1	314	20.7	21.0	100.0
	.	21	1.4	faltante	
Total =		1,518	100.0	100.0	
Casos válidos: 1,497					*Casos faltantes: 21*

DE = desviación estándar

El número a informar en este caso para el Indicador 11.1 será 21.0 por ciento, el porcentaje de niños cuyo peso para la edad fue bajo, entre aquellos que tenían un peso para la edad válido en la base de datos. En el paquete EPI INFO incluido con este manual, los indicadores se imprimen directamente con la función REPORT, de tal manera que sólo se muestra la información necesaria.

✎ **EJEMPLO:**
La salida generada por el programa REPORT que viene con este manual luce como sigue:

Proporción de Niños con Bajos Valores Z de Peso para la Edad (ZPPE)
(Indicador 11.1,2)

Total de Niños	ZPPE < -2	ZPPE < -3
1,497	21%	4%

☞ En algunos casos, el paquete de análisis no genera el indicador propiamente dicho. Por ejemplo, una encuesta no proporciona los números absolutos de población. Para convertir el porcentaje de población con acceso conveniente a agua potable (el valor informado por el programa REPORT) en números absolutos (Indicador 13.1), es necesario multiplicar este porcentaje por el tamaño de la población total.

Los índices antropométricos deben ser calculados para cada niño, a partir de las medidas de peso y talla. EPI INFO utiliza una rutina separada para comparar la edad y peso (o talla) del niño con un estándar internacional y asigna entonces una puntuación para cada niño según la desviación, con relación al estándar, de la mediana del valor (valor z). Los cálculos pueden hacerse a la entrada de datos, como se hace con el paquete EPI INFO o posteriormente, utilizando un paquete antropométrico tal como ANTHRO (antes CASP) o EPINUT. La ventaja de EPI INFO es que los índices pueden verse en la pantalla tan pronto como se digita la información de peso y talla de tal manera que el personal de digitación puede detectar inmediatamente índices que son inverosímilmente bajos o altos (generalmente debido a errores de digitación cuando se entran los datos). Indices del valor z por debajo de -6.0 o por encima de 6.0 son generalmente inverosímiles.

Estimación del Margen de Error para Cada Indicador

La mayoría de los paquetes de análisis estiman los errores estándar de los porcentajes y promedios, asumiendo que la muestra fue seleccionada como una "muestra aleatoria simple."[1] Sin embargo, los errores estándar se subestiman generalmente porque no tienen en cuenta que los datos provienen de una encuesta de conglomerados. Se requieren paquetes de análisis diseñados especificamente para trabajar con encuestas de muestras por conglomerados y estimar correctamente los errores estándar. EPI INFO (Versión 6) contiene un módulo llamado CSAMPLE para calcular los errores para porcentajes y promedios, aunque este módulo no está integrado con el módulo ANÁLYSIS. El Apéndice 3 contiene las instrucciones para generar los errores estándar y los intervalos de confianza para cada indicador que utilice este módulo.

El programa CSAMPLE de EPI INFO produce automaticamente la estimación del indicador y su margen de error. Véase el Apéndice 3 para las instrucciones sobre como utilizar este programa.

✎ **EJEMPLO:**

Se muestra el margen de error para el indicador y su intervalo de confianza del 95 por ciento. Por ejemplo, al analizar peso para la edad, el resultado fue:

El porcentaje con bajo peso para la edad es 21.0% con una intervalo de confianza de (19.3, 22.7).

Cálculo de los Factores de Ponderación para Muestras No Autoponderadas

Si se utilizaran marcos muestrales separados para las diferentes regiones en la primera etapa de

[1] La fórmula para calcular el *error estándar* (e.s.) para una proporción, asumiendo muestreo aleatorio simple es:

$$e.s. = \sqrt{[p \times (1 - p) \div n]}$$

donde p es la proporción de interés y n es el tamaño de la muestra en el grupo. El *margen de error* para esta proporción es (2 × e.s.). El *intervalo de confianza* para la proporción se calcula como $p \pm$ el margen de error.

muestreo, la muestra nacional no quedaría seleccionada con probabilidad proporcional al tamaño (PPT). Esto ocurre también si usted estratifica de acuerdo con otro factor (e.g., urbano/rural o tugurio/no tugurio) y toma diferentes fracciones (proporciones) de muestreo en cada estrato. Estas muestras no son "auto ponderadas" y usted debe ponderar su muestra al reportar las estimaciones nacionales. Es decir, debe asegurarse que cada submuestra separada—por ejemplo, cada región—contribuye solamente con lo que contribuiría si la muestra de la encuesta a nivel nacional hubiese sido seleccionada con PPT. Este procedimiento para calcular las ponderaciones se ilustra en el Cuadro 7.2 con un ejemplo hipotético.

☞ Cree una nueva variable para representar el factor de ponderación. Este factor debe utilizarse para calcular las estimaciones *nacionales* a partir de las estimaciones subnacionales separadas. Cada caso debe ponderarse por esta variable.

Por ejemplo, suponga que el país tiene una población que comprende 1,333,415 hogares y que consiste en siete regiones, cada una de las cuales fue encuestada independientemente. La muestra total de hogares fue de 11,312.

Para calcular el factor aproximado de ponderación para cada muestra de la encuesta, produzca un cuadro como el Cuadro 7.2.

☞ Para calcular la cobertura actual para la muestra total, construya una varibale que representa el factor de ponderación:

IF (REGION = 1) THEN WTFACTOR = 2.73
IF (REGION = 2) THEN WTFACTOR = 0.48
IF (REGION = 3) THEN WTFACTOR = 0.99
IF (REGION = 4) THEN WTFACTOR = 0.26
IF (REGION = 5) THEN WTFACTOR = 1.77
IF (REGION = 6) THEN WTFACTOR = 0.41
IF (REGION = 7) THEN WTFACTOR = 0.23

El análisis debe ser entonces ponderado por la variable, WTFACTOR. El resultado producirá indicadores a nivel nacional adecuadamente ponderados para compensar por el diseño no autoponderado de la muestra.

☞ Este cálculo también se puede hacer con una calculadora manual. Multiplique cada estimación regional por su factor de ponderación, sume los resultados y divida por la suma de los factores de ponderación.

✎ **EJEMPLO:**

(1)	(2)	(3)	(4)
Región	**Estimación**	**Factor de ponderación**	**Estimación × WTFACTOR**
1	**0.85**	**1.7**	**1.445**
2	**0.75**	**0.2**	**0.15**
3	**0.70**	**1.0**	**0.70**
Total		**∑ = 2.9**	**2.295**

Estimación ponderada del promedio nacional: (suma de columnas 4 ÷ 2.9) = 0.79

Cuadro 7.2. Cuadro ilustrativo para el cálculo de los factores de ponderación

Estrato	(1) Tamaño de la población (hogares)	(2) Población del estrato como proporción de la población nacional	(3) Tamaño final de la muestra	(4) Muestra en el estrato como proporción de la muestra total	(5) Factor de ponderación
		(col. 1 ÷ 1,333,415)	*n (hogares)*	(col. 3 ÷ 11,312)	(col. 2 ÷ col. 4)
1	550,088	.412	1,707	.1509	2.730
2	93,088	.070	1,664	.1471	.476
3	192,580	.144	1,648	.1457	.988
4	47,873	.036	1,560	.1379	.261
5	330,312	.248	1,587	.1403	1.768
6	77,124	.058	1,592	.1407	.412
7	42,350	.032	1,544	.1374	.233
Total	1,333,415	1.000	11,312	1.000	

Estimaciones Específicas por Edad y Sexo

☞ Para calcular cada indicador por sexo del niño, usted debe seleccionar del archivo de datos solamente los niños del mismo sexo y hacer el análisis para este grupo por separado. Haga lo mismo para los niños del otro sexo.

Siga el mismo procedimiento si usted quiere presentar indicadores para grupos menos amplios de edad, o para subgrupos nacionales tales como grupos regionales, distritos o urbano/rurales. Seleccione cada grupo por separado y corra el programa de análisis para producir un informe para cada grupo. Repita el proceso para cada grupo para el cual usted desea calcular indicadores separados. Si usted no está usando el programa de análisis de EPI INFO, tal vez encuentre más facil hacer una tabulación cruzada por sexo del niño u otras variables similares que definan grupos subnacionales (por ejemplo, regional, urbano/rural).

☞ Asegúrese que usted rotula claramente los informes de computador (e.g., "sólo niños") para evitar confusiones más tarde.

Ahora usted está listo para escribir su informe preliminar y representar los indicadores gráficamente.