INEI
INSTITUTO NACIONAL DE ESTADISTICA E INFORMATICA

GUÍA DE PROCEDIMIENTOS PARA EVALUAR BASE DE DATOS DE REGISTROS ADMINISTRATIVOS

Dirección Técnica de Demografía e Indicadores Sociales

Doc. SIRA.02
Lima, mayo 2018

PRESENTACIÓN

La producción estadística es un insumo necesario para el diseño de políticas públicas, utilizando metodologías sólidas y estandarizadas. En este marco, el Instituto Nacional de Estadística e Informática, ha elaborado el documento **"GUÍA DE PROCEDIMIENTOS PARA EVALUAR BASES DE DATOS PROVENIENTES DE REGISTROS ADMINISTRATIVOS"**, teniendo en consideración la importancia de fortalecer los registros administrativos y transformarlos en estadísticas.

El presente documento, constituye una guía general con orientaciones técnicas y metodológicas; emite lineamientos, estándares, buenas prácticas estadísticas y directrices, garantizando de esta manera mejorar la calidad estadística de la información sobre registros administrativos.

Es importante destacar la necesidad de analizar la cobertura y calidad de la información estadística, de este modo, se conoce la completitud de los datos, el nivel de desagregación y la calidad de los datos, que redundará en mejorar la disponibilidad, calidad, coherencia y comparabilidad.

El contenido de este documento comprende tres capítulos, el primero, describe aspectos generales como finalidad, objetivos, principios de la calidad estadística, entre otros; el segundo, refiere los procedimientos para la revisión, identificación de variables, cobertura, diccionario de variables, tratamiento de las unidades de registro y tratamiento de las variables. El último capítulo incorpora definición básica y finalmente anexo sobre diccionario de datos.

.

ÍNDICE

PRESENTACIÓN **3**

I. ASPECTOS GENERALES **9**

1.1. Finalidad 9

1.2. Objetivos 9

1.3. Alcance 9

1.4. Base Legal 9

1.5. Criterios de calidad estadística 9

1.6. Elementos clave a considerar en un sistema de registros administrativos 11

1.6.1. Base conceptual y metodológica 11

1.6.2. Clasificaciones y variables investigadas 11

1.6.3. Cobertura 11

1.6.4. Procesos de recolección de datos y frecuencia de disponibilidad 12

1.6.5. Proceso de diagnóstico 12

II. PROCEDIMIENTOS PARA REVISAR BASES DE DATOS **17**

2.1. Determinación de las unidades de registro 17

2.2. Cobertura geográfica 18

2.3. Diccionario de datos 18

2.3.1. Pasos para elaborar diccionario de datos o variables 21

2.3.2. Diseño del diccionario de datos o variables 32

2.4. Base de datos 38

2.5. Tratamiento de las unidades de registro 39

2.5.1. Revisión de duplicados 39

2.5.2. Revisión de omisiones 40

2.6. Tratamiento de las variables 47

III. DEFINICIONES BÁSICAS **69**

ANEXOS **79**

REFERENCIAS BIBLIOGRÁFICAS **87**

I. ASPECTOS GENERALES

I. ASPECTOS GENERALES

1.1. Finalidad

Brindar lineamientos técnicos para revisar y evaluar bases de datos provenientes de diversas fuentes de información como encuestas y registros administrativos, y de esta manera contribuir en el mejoramiento de la calidad de información estadística.

1.2. Objetivos

- Fortalecer la producción estadística a fin de contribuir al mejoramiento de los registros administrativos, mediante inclusión de lineamientos técnicos, y tener un mejor aprovechamiento de la información estadística.

- Contar con lineamientos estandarizados, buenas prácticas y directrices para analizar bases de datos provenientes de registros administrativos.

- Apoyar a las entidades públicas y privadas con un documento guía, con criterios técnicos que sirva de consulta para revisar y evaluar bases de datos.

1.3. Alcance

- Funcionarios de las instituciones públicas y privadas, encargados del análisis y revisión de las bases de datos.

1.4. Base Legal

- Ley Nº 30693, Ley de Presupuesto del Sector Público para el Año Fiscal 2018. Artículo 18.
- Ley N° 28411, Ley General del Sistema Nacional de Presupuesto. Artículos 83 y 84.
- Decreto Supremo Nº072-2012-PCM que aprueba el "Código de Buenas Prácticas Estadísticas del Perú".
- Decreto Legislativo Nº 604, Ley de Organización y Funciones del INEI.

1.5. Criterios de calidad estadística

La calidad estadística se define como el conjunto de propiedades que debe tener tanto el proceso como el producto estadístico para satisfacer las necesidades de información de los usuarios y las partes interesadas[1]. La evaluación de la calidad estadística se realiza a partir de los siguientes atributos:

[1] DANE, Evaluación y certificación de la calidad estadística, pag.11, Bogotá, 2012.

Credibilidad, se define como la confianza de la opinión pública en la información estadística producida por las entidades.

Oportunidad/puntualidad, es el tiempo transcurrido desde que se tiene disponible la información del registro administrativo y la presentación de resultados. La Puntualidad se relaciona con el cumplimiento de las fechas para la publicación de la información, previamente establecidas en un calendario.

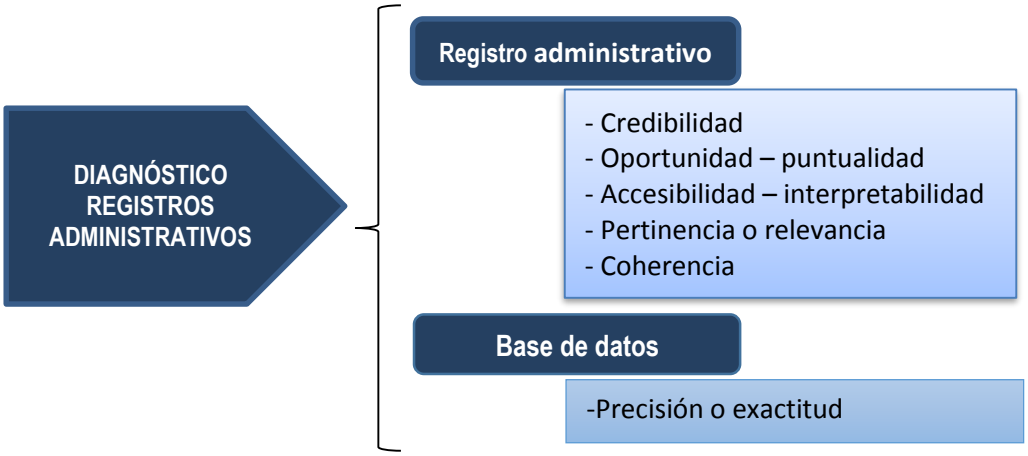
Accesibilidad, es la idoneidad en que están disponibles los datos, metadatos, los medios de divulgación, las metodologías y servicios de apoyo al usuario. El acceso a la información se refiere al conjunto de técnicas para buscar, categorizar, modificar y acceder a la información que se encuentra en un sistema, bases de datos, bibliotecas, archivos, Internet[2].

Pertinencia o relevancia, es una medida cualitativa del valor aportado por la información, es decir, el grado de utilidad para satisfacer el propósito por el cual la información fue buscada o solicitada. La medición de la relevancia de un producto estadístico requiere de la identificación de su grupo de usuarios y sus necesidades.

Coherencia, es el proceso de mantener información uniforme, los datos de los registros administrativos estén conectados y consistentes. Las estadísticas son coherentes cuando todos los elementos de un sistema son uniformes en un momento específico en el tiempo[3], y dentro del conjunto de datos al que pertenecen, y con otros conjuntos de datos.

Precisión/exactitud, es el grado de coincidencia entre los resultados, es decir, cuando los datos entregados a los usuarios describen correctamente las características que deben medir.
La exactitud se refiere a qué tan cerca están del valor real las mediciones de un sistema de medición. La precisión se refiere a qué tan cerca están las mediciones entre ellas.

Gráfico N°1
Criterios de análisis para los registros administrativos y las bases de datos



[2] https://es.wikipedia.org/wiki/Acceso_a_la_información.
[3] http://www.prucommercialre.com/que-es-la-coherencia-de-datos/

1.6. Elementos clave a considerar en un sistema de registros administrativos

Un buen sistema de registros administrativos, como fuente de información estadística, depende de los elementos siguientes:

1.6.1. Base conceptual y metodológica

En primer lugar, es necesario definir los objetivos y priorizar los temas a incluir, esta priorización determinará la demanda de información en cada etapa del proceso estadístico. La demanda permitirá buscar las fuentes de información, es decir, los registros administrativos necesarios para satisfacer los requerimientos. La evaluación de cada una de las etapas del proceso de registros administrativos, está enmarcada en normas, directrices y buenas prácticas aceptadas internacionalmente.

También se revisarán los conceptos y definiciones que caracterizan el hecho registrado y que influyen en la temática para su uso con fines estadísticos.

Por otro lado, es importante que exista documentación de los procesos realizados en cada una de las etapas de la producción estadística, es decir, debe existir una metodología que integre los procesos del registro administrativo, como recolección, procesamiento y difusión de la información, asimismo, el manejo de la base de datos, procesos de compilación, validación y metadato.

En el marco de la metodología, es pertinente considerar el origen del registro administrativo: solicitud, objetivos, conceptos, leyes, políticas y normas que fundamentan el registro y las necesidades de adecuación para su proceso estadístico.

1.6.2. Clasificaciones y variables investigadas

Clasificación, es la acción de organizar o situar algo según una determinada directiva. Comprende la agrupación de la información de manera ordenada, sistemática para fines de comparabilidad. Los clasificadores especifican las clasificaciones que se utilizan y a que conceptos o variables clasifican.

Las variables investigadas se diseñan de acuerdo al objeto de estudio. Las clasificaciones y las variables deben estar incluidas en el marco conceptual.

1.6.3. Cobertura

La cobertura temática enumera los temas que se cubren, así como los conceptos y variables considerados. La cobertura permite determinar el alcance temático para su uso con fines estadísticos.

La cobertura geográfica indica si la información es representativa a nivel nacional, departamental, provincial y/o distrital; hace referencia a la delimitación de la población objetivo del registro administrativo y de su base de datos en términos geográficos. Este alcance nacional, regional, local permite efectuar la la desagregación geográfica de resultados estadísticos.

Es posible también, revisar y evaluar el ámbito de acción en el que el registro administrativo se desarrolla, así como su participación en los sectores económico, social, medio ambiente y recursos naturales, entre otros.

1.6.4. Procesos de recolección de datos y frecuencia de disponibilidad

Los procesos de recolección y procesamiento de los datos deben responder a ciertas características técnicas y procesos aplicados que permitan obtener los resultados deseados. Durante la etapa de diagnóstico es fundamental identificar procesos así como bases de datos que cuenten con procedimientos documentados, controles, archivos de datos o sistemas de información y que generan reportes estadísticos para fines de control.

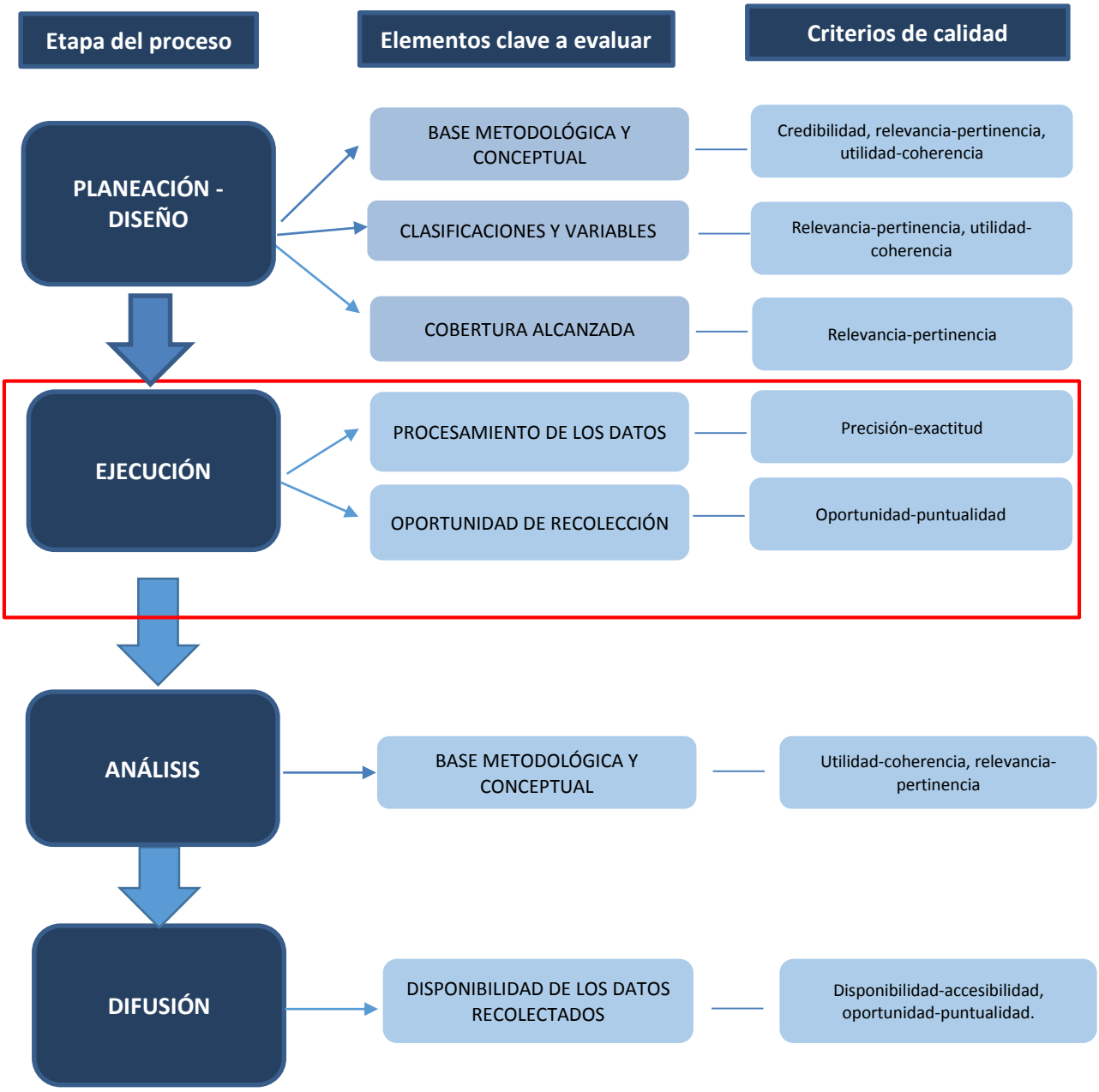
Los datos y los resultados deben estar disponibles a tiempo y se pueda acceder fácilmente a ellos, que sean de fácil interpretación y útiles para la toma de decisiones.

1.6.5. Proceso de diagnóstico

Consiste en evaluar los elementos clave del registro administrativo, determinar el diagnóstico o estado actual de los registros administrativos.

Aplicar los criterios de calidad estadística y relacionar con las etapas del proceso estadístico, como planeación, diseño, producción, análisis y difusión.

Gráfico N°2
Etapas del proceso del registro administrativo, elementos clave a evaluar y criterios de calidad estadística



II. PROCEDIMIENTOS PARA REVISAR BASES DE DATOS

II. PROCEDIMIENTOS PARA REVISAR BASES DE DATOS

2.1. Determinación de las unidades de registro

La unidad de registro constituye el elemento básico de un sistema de registro, sobre dichos elementos se registran las características, teniendo en cuenta el propósito del registro administrativo. Es importante determinar las unidades de registro, que serán de utilidad en el análisis estadístico ya que permitirán definir las unidades de análisis; en consecuencia, se analizan las características registradas.

Ejemplo 1: Registro administrativo: Padrón nominal

El propósito del padrón nominal es contar con un registro de niños/as menores de 6 años de todo el país, consolidado en una base de datos autenticada, de naturaleza dinámica y se actualiza permanentemente; proporciona información del número de niños y niñas por departamento, provincia, distrito y centro poblado, con lo cual se promueve el acceso del niño/a los diferentes servicios que brinda el estado, así mismo, permite analizar la brecha no cubierta por estos, contribuyendo de esta manera al ejercicio de derechos fundamentales y reduciendo las desigualdades.

Unidad de registro: El niño o niña de 6 años a menos de edad.

Ejemplo 2: Registro administrativo: Registro de Atenciones del Sistema Integral de Salud -SIS

El propósito del SIS, comprende:

- Evaluar el nivel de calidad de la oferta de los prestadores de servicios y la satisfacción en la atención de salud de los beneficiarios.
- Facilitar servicios de calidad a los beneficiarios del Seguro Integral de Salud.
- Dirigir los procesos de afiliación y operación del Seguro Integral de Salud en todos los niveles.

Unidad de registro: Es la atención de salud del beneficiario SIS.
Subunidad de registro: El beneficiario del SIS.

Ejemplo 3: Registro administrativo: Registro de Facturación de la SUNASS

El propósito del registro de facturación, es evidenciar una gestión transparente y el mejoramiento de la calidad y acceso de los servicios de agua potable y alcantarillado en todo el territorio nacional, a través de la optimización y vigencia del sistema de regulación tarifaria como también de la función supervisora y fiscalizadora de la prestación de los servicios de agua potable y alcantarillado, así como el fortalecimiento de capacidades de los actores involucrados en la prestación de los servicios de saneamiento.

Unidad de registro: Unidad de uso (Medidor).
Subunidad de registro: El usuario de la unidad de uso.

2.2. Cobertura geográfica

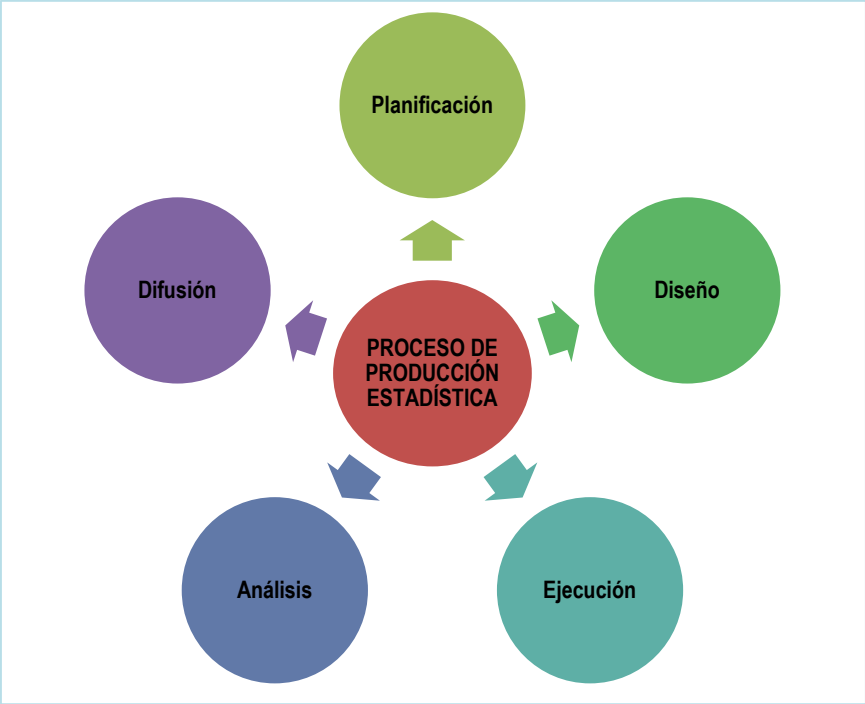
Es el alcance de registros relacionado con el territorio cubierto, es decir, podría ser nacional, departamental, provincial, distrital, centro poblado etc. La determinación de la cobertura geográfica es importante para el análisis que se realicen en las estimaciones.

Ejemplos:

- Padrón Nominal, su alcance es nacional, provincial y distrital.
- Registro de Catastro Distrital: su alcance seria solo distrital, pues su cobertura solo llega a este ámbito.

2.3. Diccionario de datos

El proceso de producción estadística requiere establecer y proveer información estandarizada que permita una buena comunicación entre las fases del proceso y mejorar la eficiencia de la institución.

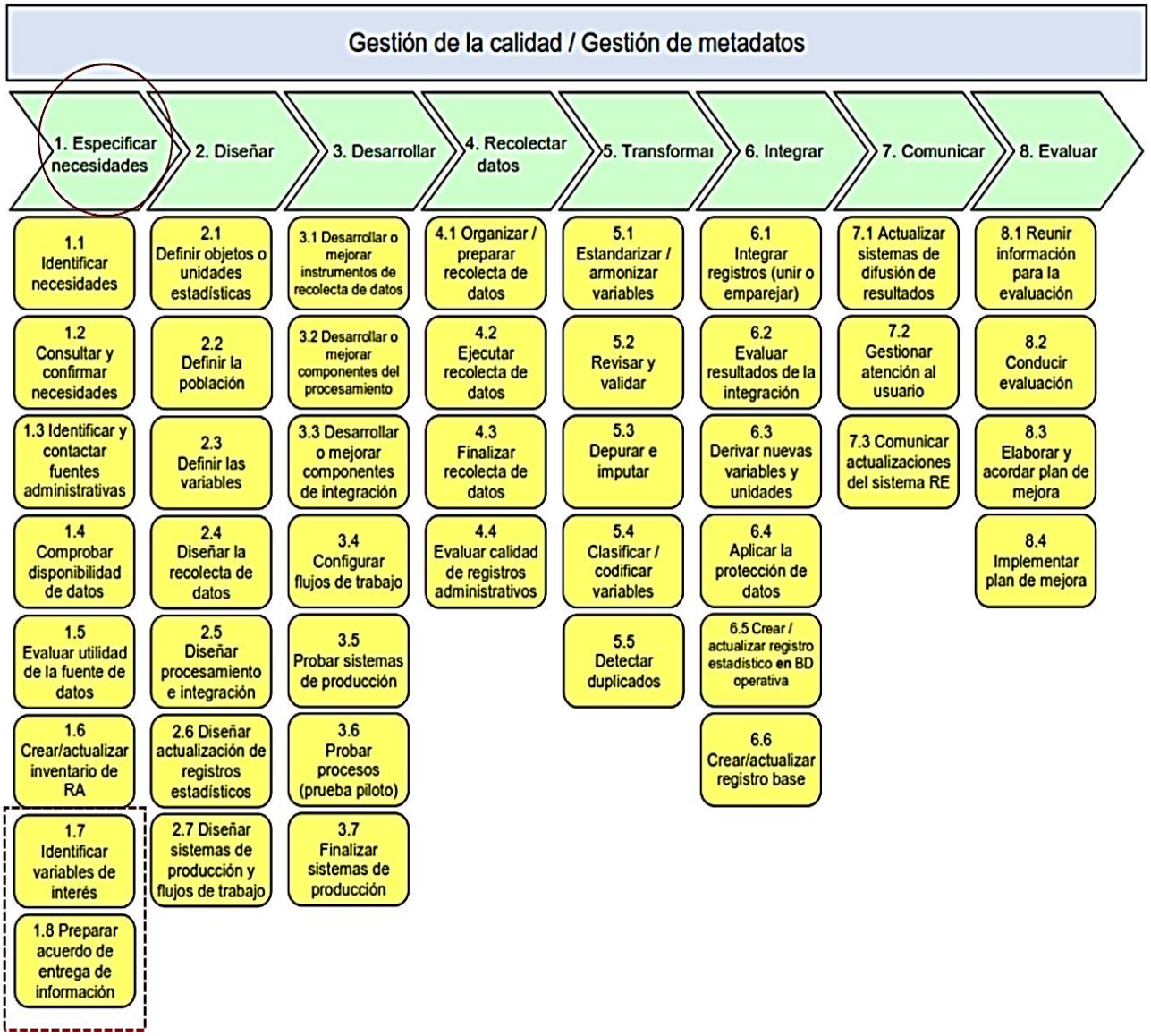


En este sentido, la elaboración del diccionario de datos[4] forma parte del proceso de producción estadístico, es decir, de la fase "Especificar necesidades" donde se identifica una necesidad de nuevas estadísticas o cuando la retroalimentación sobre las estadísticas producidas actualmente inicia una revisión[5].

[4] El diccionario de datos contiene información de la base de datos como el nombre de la tabla o archivo de donde viene la variable, el nombre del campo, la descripción del campo, el tipo y medida del dato y la longitud del campo.
[5] Federico Segui Stagno "Marco Conceptual y Metodológico que Sustenta el Diseño, Desarrollo e Implementación de un Sistema Integrado de Registros Estadísticos de Población e Inmuebles", 2016, pág. 31.

El objetivo del diccionario de datos o de variables es definir con precisión las variables que se manejan en una base de datos a fin de evitar errores en la interpretación.

Conocer el grado de detalle de las tareas que se desarrollan en cada proceso, significa visibilizar la importancia de la construcción del diccionario en la producción estadística y calidad de la información. Para ello, se utilizará el Modelo Genérico de Procesos de Producción de Registros Estadísticos (GSRBPM).



Fuente: Elaboración del consultor Federico Segui.

El diccionario de datos corresponde al subproceso ***"Identificar variables de interés y preparar acuerdo de entrega de información***", en el cual se registran las variables que son consideradas como característica de una unidad observada. Los elementos que componen su estructura son los siguientes:

a) Nombre de la variable, se puede dar un nombre a la variable que tenga que ver con su contenido, o simplemente un nombre o alias como var1, o v1.

De una forma u otra, el nombre debe contener como máximo 8 caracteres que pueden ser alfabéticos, numéricos, o el símbolo de subrayado (_). No puede contener espacios en blanco ni caracteres especiales como!, ?, etc. El primer carácter debe ser alfabético, $, ó #. No pueden existir dos variables con el mismo nombre, y debe tener en cuenta el hecho de que el sistema no distingue entre mayúsculas y minúsculas (siendo la misma variable TIEMPO que tiempo).

b) **Tamaño o anchura,** comprende el número de caracteres que contiene la variable.

c) **Tipo o formato,** los más utilizados son numérico y cadena. Generalmente las variables a escala se definen como numéricas, mientras que las categóricas como numéricas o cadena. En el primer caso, los valores de la variable categórica son las modalidades, y en el segundo, se consideran unos valores numéricos arbitrarios (variable nominal) o que indican un orden (variable ordinal), posteriormente se asocia a cada uno de ellos una etiqueta de valor.

d) **Etiqueta de variable,** permiten describir las variables hasta 256 caracteres de longitud, sobre todo cuando el nombre no es identificativo. El nombre de las etiquetas puede llevar cualquier tipo de signo o símbolo, acentos,(?), (!), etc., pero no puede exceder de 120 caracteres.

e) **Código o valor de variable,** es el número que le corresponde a cada categoría.

f) **Nombre del código o etiqueta de valor de variable, comprende las categorías de las variables.**

ESTRUCTURA DEL DICCIONARIO DE VARIABLES

A. Nombre	D. Etiqueta	E. Código	F. Nombre del código

ESTRUCTURA DEL DICCIONARIO DE DATOS

A. Nombre	B. Tamaño o anchura	C. Tipo o formato	D. Etiqueta E. Código F. Nombre del código

2.3.1. Pasos para elaborar diccionario de datos o variables

Luego de haber aplicado un instrumento de recolección de información (censos, encuestas, registros administrativos) se definen los pasos siguientes:

Paso 1: Listar todas las variables que debe contener el diccionario tanto para uso interno o para compartir o difundir, a fin de visibilizar los conceptos o elementos del sistema.

Ejemplo 1:

PADRON DE BENEFICIARIOS DEL PROGRAMA VASO DE LECHE
I SEMESTRE 2017

UBICACIÓN GEOGRAFICA	
DEPARTAMENTO	LORETO
PROVINCIA	MAYNAS
DISTRITO	INDIANA

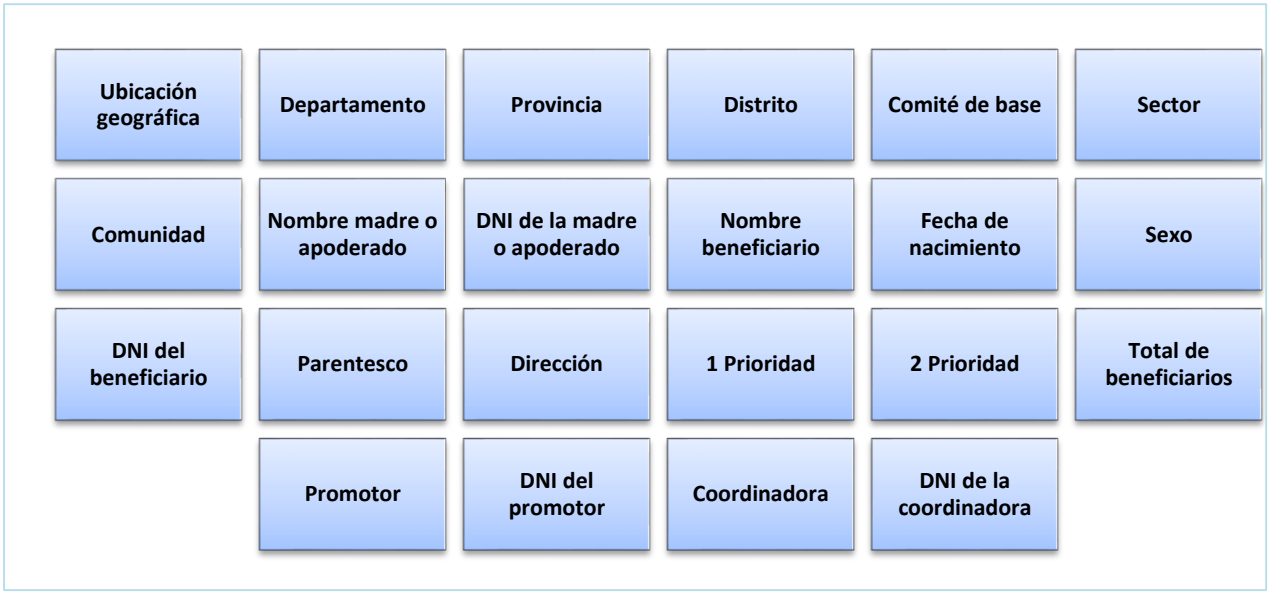
NOMBRE DEL COMITÉ DE BASE O COMITÉ SECTORIAL	
COMITÉ DE BASE	
SECTOR	
COMUNIDAD	

N°	NOMBRE MADRE O APODERADO	DNI	NOMBRE BENEFICIARIOS	FECH NAC.	SEXO	DNI	parentesco	DIRECCION	BENEFICIARIOS												
									1 PRIORIDAD									2 PRIORIDAD			
									1	2	3	4	5	6	GES	LAC	DIS				
TOTAL BENEFICIARIOS:																					

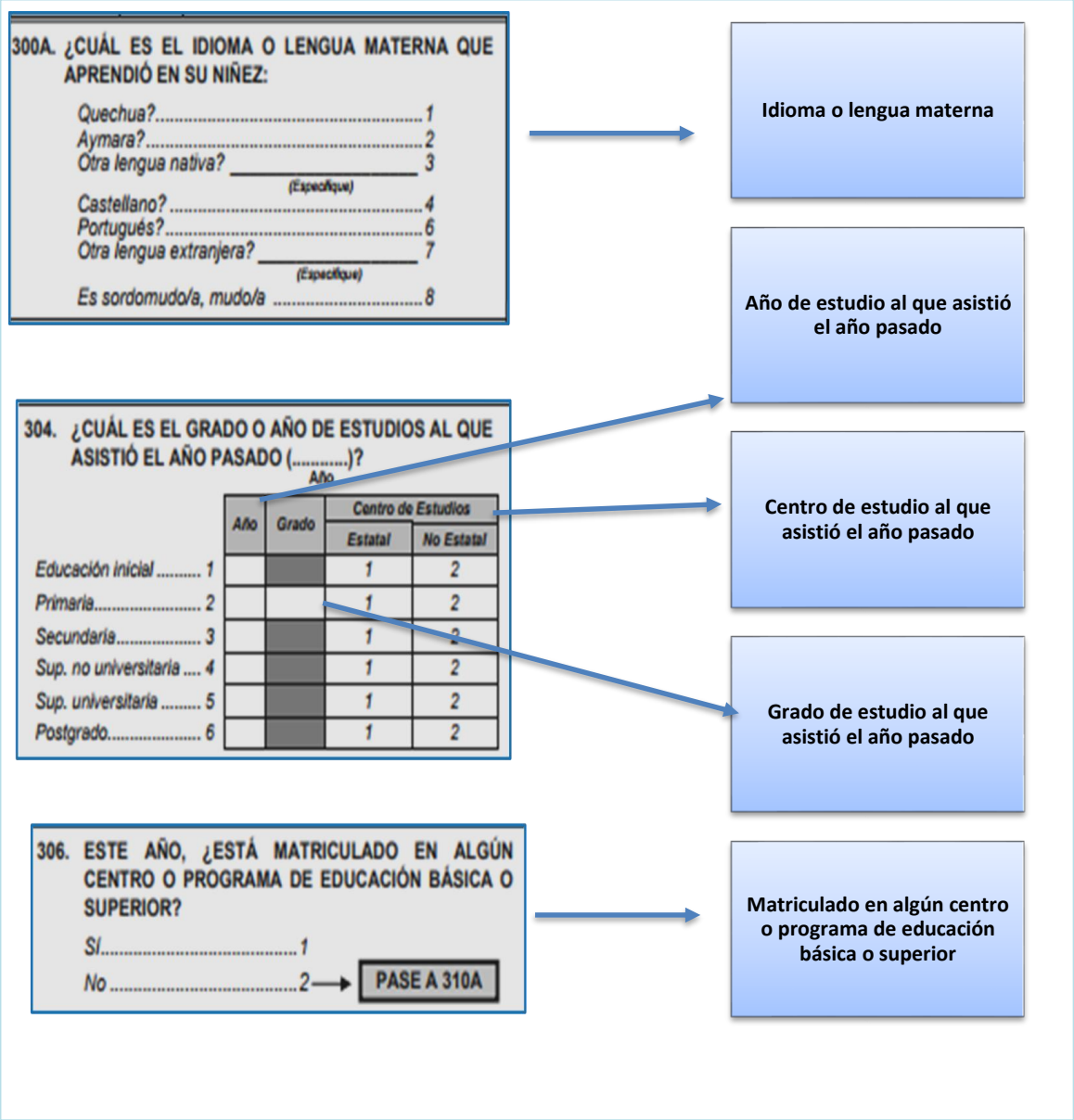
PROMOTOR	
DNI	
FIRMA	

COORDINADORA	
DNI	
FIRMA	

Listado de variables:



Ejemplo 2: **Listado de variables:**



Recomendación para listar las variables:

- ✓ Identificar las variables que pertenezcan a una sola característica.
- ✓ Cualquier persona debe identificar la variable.
- ✓ Si se usan siglas estas deben ser definidas.

Paso 2: Definir nombre de variable (usar un sinónimo u otra denominación), esta nueva denominación tiene que reflejar a la variable de origen.

Ejemplo 1:

PADRON DE BENEFICIARIOS DEL PROGRAMA VASO DE LECHE
I SEMESTRE 2017

UBICACIÓN GEOGRAFICA	
DEPARTAMENTO	LORETO
PROVINCIA	MAYNAS
DISTRITO	INDIANA

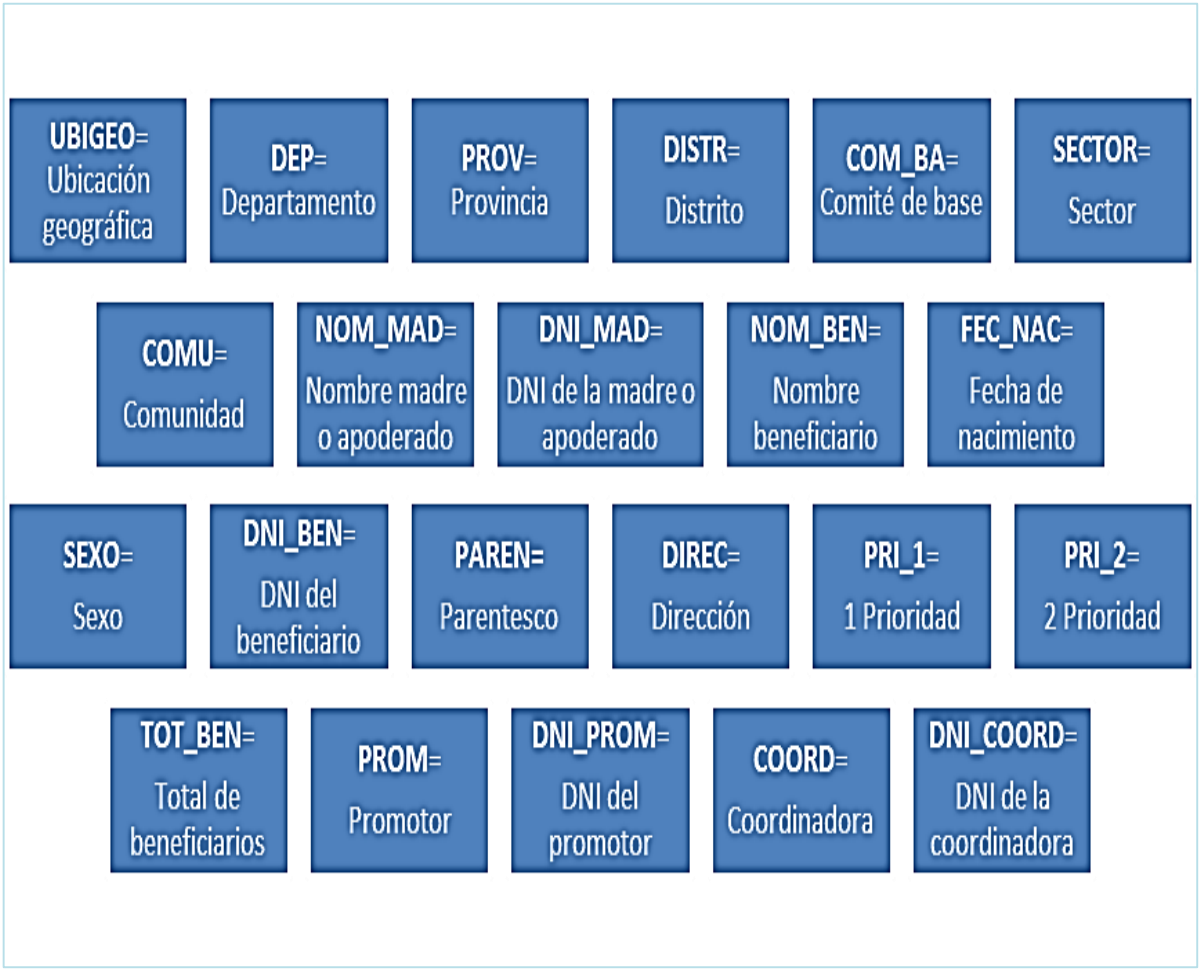
NOMBRE DEL COMITÉ DE BASE O COMITÉ SECTORIAL	
COMITÉ DE BASE	
SECTOR	
COMUNIDAD	

N°	NOMBRE MADRE O APODERADO	DNI	NOMBRE BENEFICIARIOS	FECH NAC.	SEXO	DNI	parentesco	DIRECCION	BENEFICIARIOS												
									1 PRIORIDAD									2 PRIORIDAD			
									1	2	3	4	5	6	GES	LAC	DIS				
TOTAL BENEFICIARIOS:																					

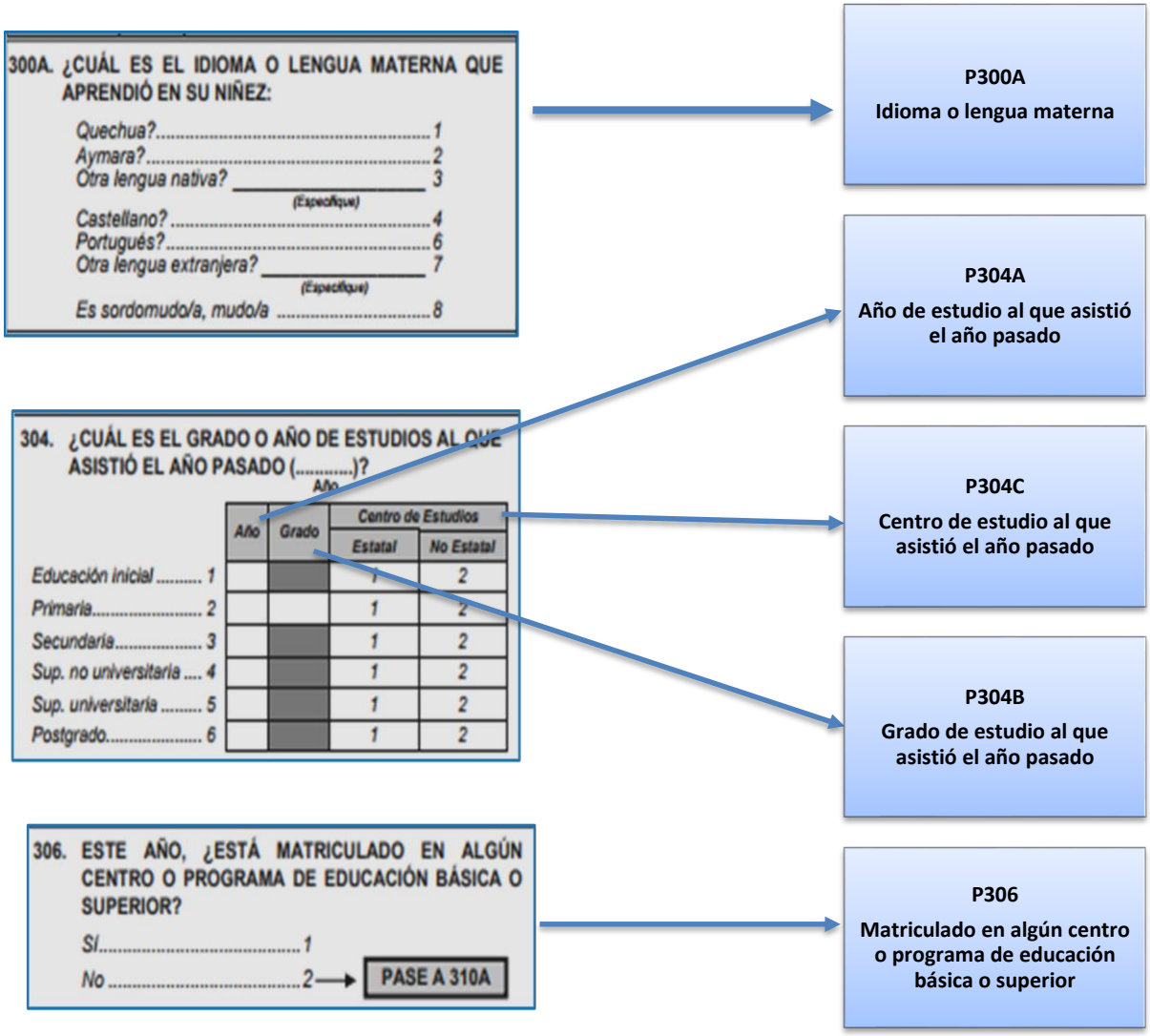
PROMOTOR	
DNI	
FIRMA	

COORDINADORA	
DNI	
FIRMA	

Definición del nombre de la variable:



Ejemplo 2: **Definición del nombre de la variable:**



Recomendación para describir el nombre de la variable:

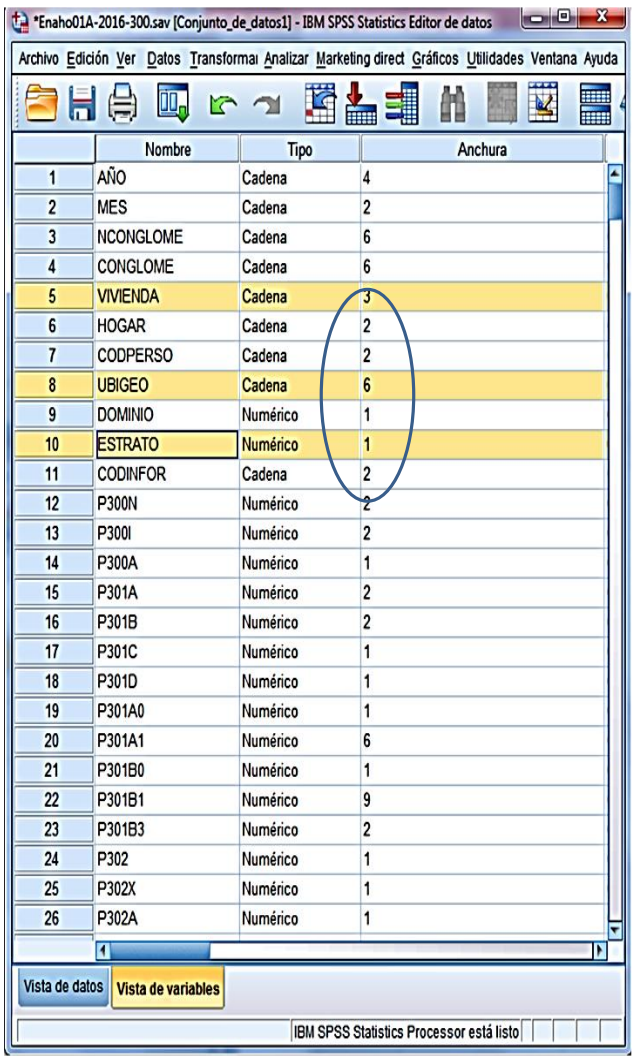
- ✓ El nombre de la variable puede tener máximo ocho caracteres.
- ✓ Son caracteres válidos: todas las letras, todos los números, el punto y los caracteres @, #, $, _
- ✓ No pueden contener espacios en blanco ni caracteres especiales como signos de admiración o interrogación (!, ?), el apostrofe (´) y el asterisco (*).
- ✓ El nombre con caracteres compuestos por dos grupos diferentes deben estar unidos por un guión (DNI_PROM).
- ✓ Definir si los nombres estarán en mayúsculas o minúsculas.
- ✓ El nombre de las variables nuevas deben comenzar siempre con una letra, pero pueden terminar con cualquier carácter válido exceptuando el punto.
- ✓ Dos variables no pueden tener el mismo nombre.

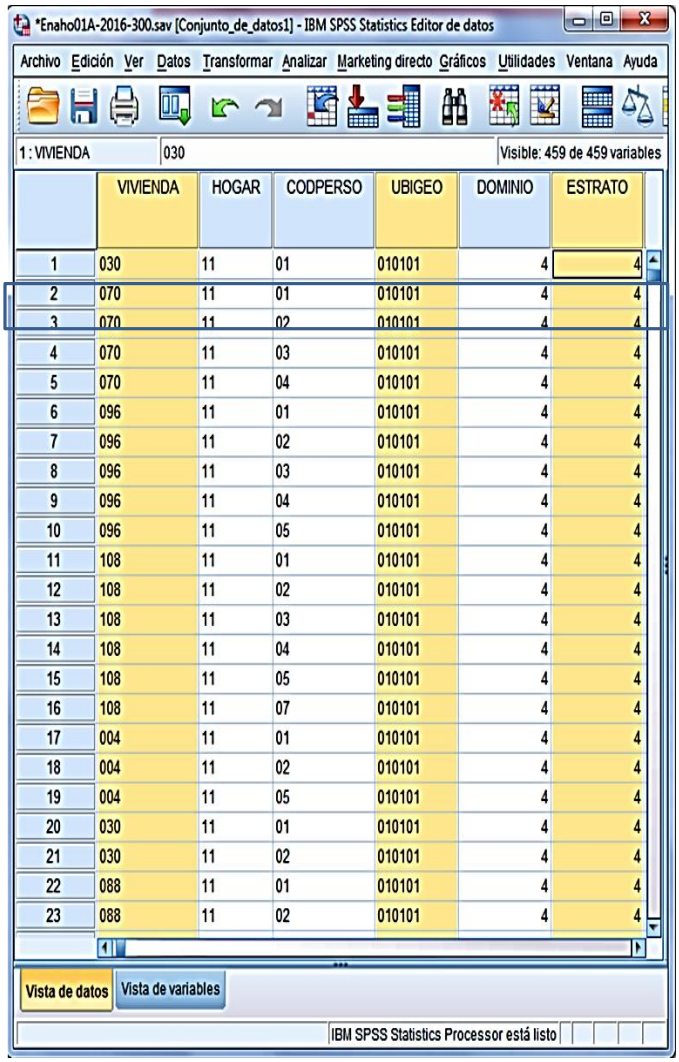
Paso 3: Definir tamaño o anchura del dato de la variable (si se trabaja con algún software estadístico). El tamaño comprende en establecer el número total de dígitos que se desea o que se requiera para la variable, incluyendo una posición para el separador decimal. Por otro lado, la anchura máxima permitida para las variables numéricas es 40; el número máximo de decimales es 16.

Cuando se trabaja con archivos diferentes a software estadísticos no se requiere definir este campo; sin embargo, se recomienda exportar a algún tipo de programa.

Ejemplo:

VIVIENDA= 3 Equivale a: **003**	**UBIGEO**= 6 Equivale a: **010101**	**ESTRATO**= 1 Equivale a: **4**

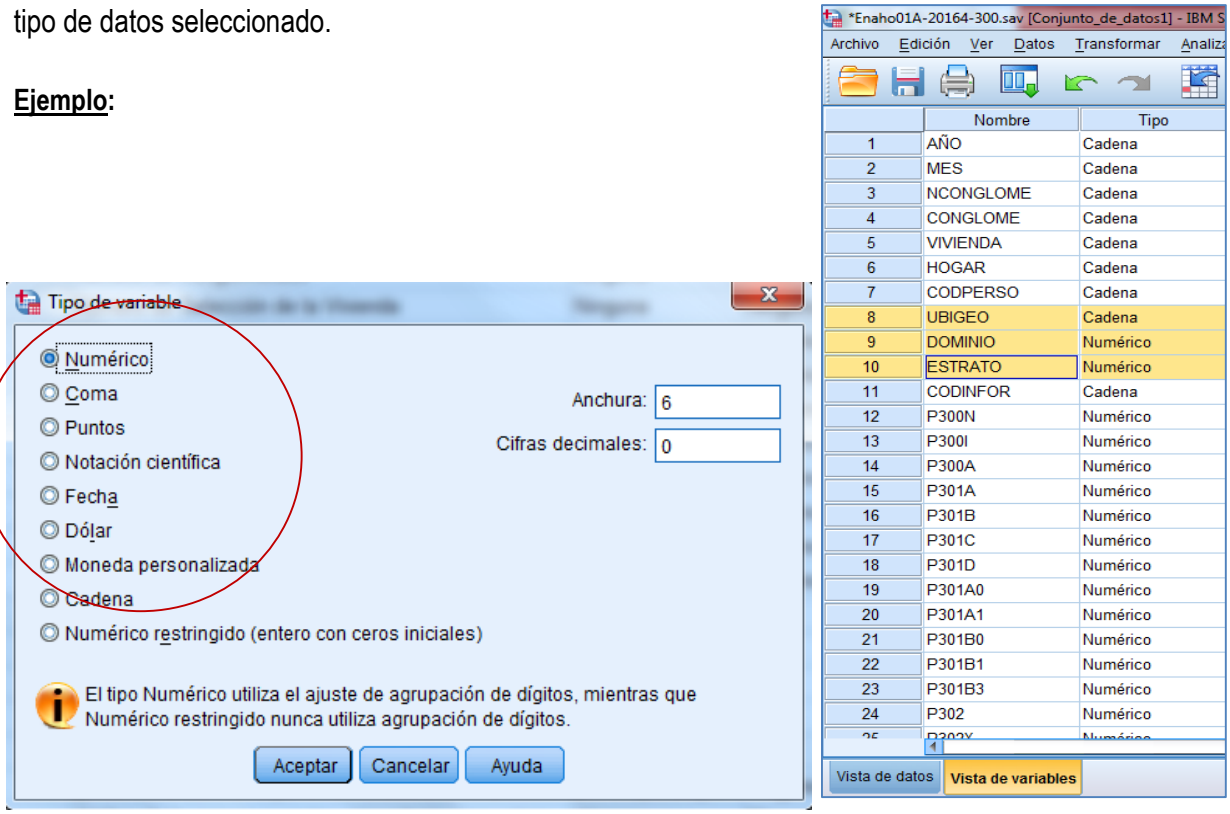




Paso 4: El tipo de variable se encuentra disponible en un software estadístico, donde el sistema le asigna el nombre (numérico o cadena entre otras). De forma predeterminada, se asume que todas las variables nuevas son numéricas. Se puede utilizar tipo de variable para cambiar el tipo de datos.

El contenido del cuadro de diálogo tipo de variable depende del tipo de datos seleccionado.

Ejemplo:



Numérico, es una variable cuyos valores son números. Los valores se presentan en formato numérico estándar.

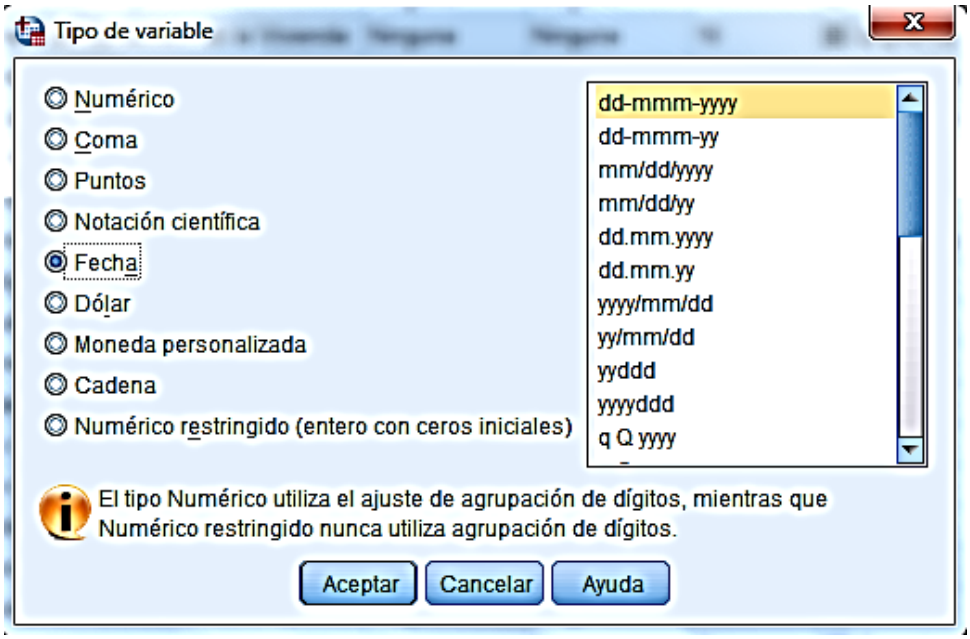
Coma, una variable numérica cuyos valores se muestran con “comas” que delimitan cada tres posiciones y con el punto como delimitador decimal. Los valores no pueden contener comas a la derecha del indicador decimal.

Punto, una variable numérica cuyos valores se muestran con puntos que delimitan cada tres posiciones y con la coma como delimitador decimal. El Editor de datos acepta valores numéricos para este tipo de variables con o sin puntos, o bien en notación científica. Los puntos separadores de los millares son automáticamente insertados.

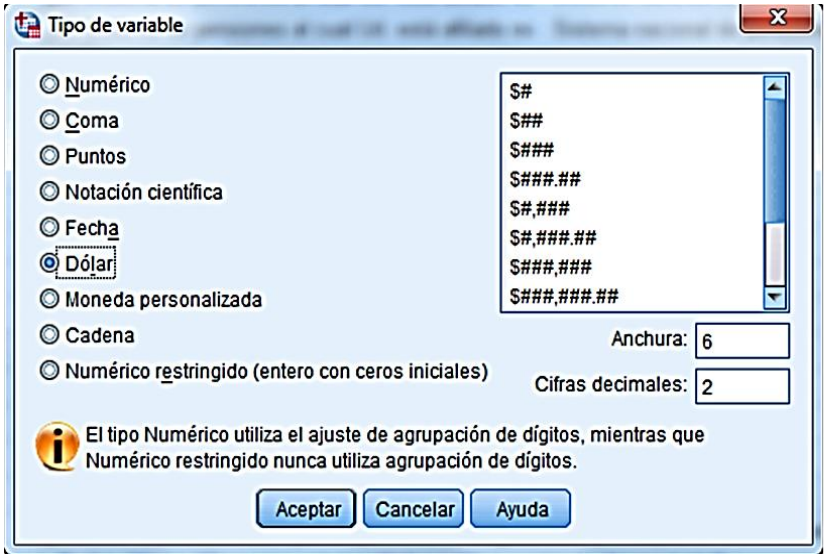
Notación científica, una variable numérica cuyos valores se muestran con una E intercalada y un exponente con signo que representa una potencia de base diez. El Editor de datos acepta para estas variables valores numéricos con o sin el exponente. El exponente puede aparecer precedido por una E o una D con un signo opcional, o bien sólo por el signo (por ejemplo, 123, 1,23E2, 1,23D2, 1,23E+2 y 1,23+2).

Ejemplo: Anchura 8 y decimales 2
1’000.000,00 = 1.0E+6

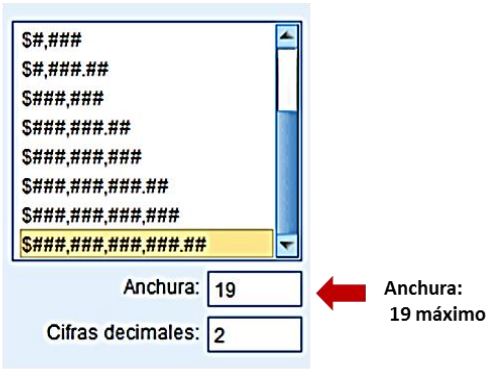
Fecha, una variable numérica cuyos valores se muestran en uno de los diferentes formatos de fecha-calendario u hora-reloj. Es necesario seleccionar un formato de la lista. Para introducir la fecha se pueden utilizar barras inclinadas, guiones, puntos, comas o espacios.



Dólar, una variable numérica que se muestra con un signo dólar inicial ($), comas que delimitan cada tres posiciones y un punto como delimitador decimal. Se pueden introducir valores de datos con o sin el signo dólar inicial, y se emplea en variables numéricas cuyos valores representan sumas de dinero en dólares.



Nombre	Tipo
INGRESO_ENERO	Dólar
INGRESO_FEBRERO	Dólar
INGRESO_MARZO	Dólar



Moneda personalizada, esta variable se emplea cuando los valores representan sumas de dinero diferentes al dólar; seleccionar uno de los formatos existentes de moneda personalizados. La diferencia con el tipo dólar es que permite trabajar con cinco diferentes tipos de moneda.

Cadena, una variable cuyos valores no son numéricos y, por lo tanto, no se utilizan en los cálculos de los estadísticos. Los valores pueden contener cualquier carácter siempre que no se exceda la longitud definida. Las mayúsculas y las minúsculas se consideran diferentes. Este tipo también se conoce como variable alfanumérica.

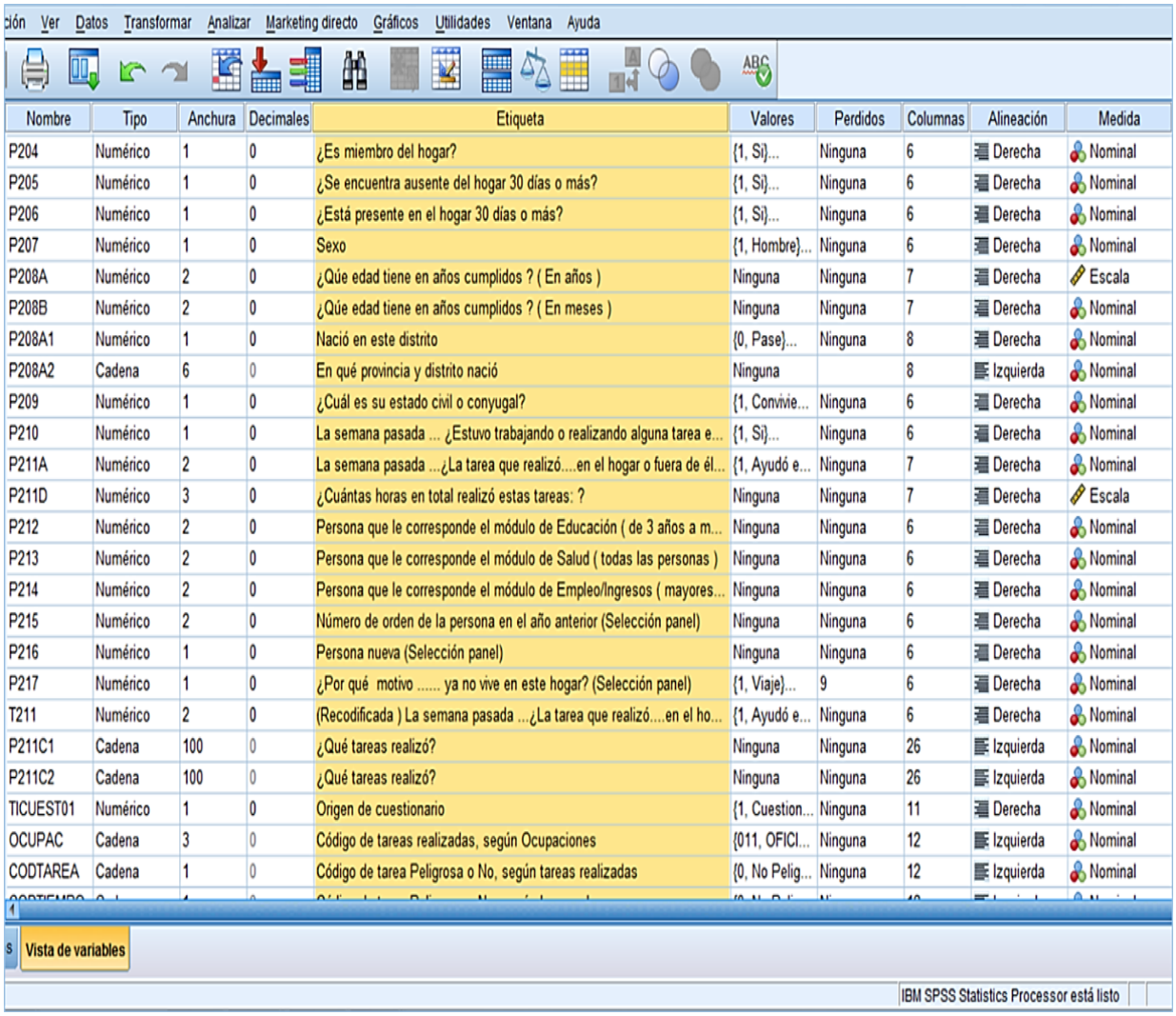
SPSS Statistics Tipo de variable	Formato de datos de Excel
Numérico	0.00; #,##0.00; ...
Coma	0.00; #,##0.00; ...
Dólar	$#,##0_); ...
Fecha	d-mmm-aaaa
Hora	hh:mm:ss
Cadena	General

Recomendación para los tipos de variable:

- ✓ En el tipo de variable COMA, los valores no pueden contener comas a la derecha del indicador decimal.
- ✓ En el tipo de variable PUNTO, los valores no pueden contener puntos a la derecha del indicador
- ✓ Cuando se trabaja con archivos Windows dependerá del objetivo de estudio para usar estos tipos de variables.

Paso 5: Definir las etiquetas que permiten describir las variables, sobre todo cuando el nombre no es identificativo. El nombre de las etiquetas puede contener cualquier tipo de signo o símbolo, acentos, espacios u otra forma, que no se admiten en los nombres de variable.

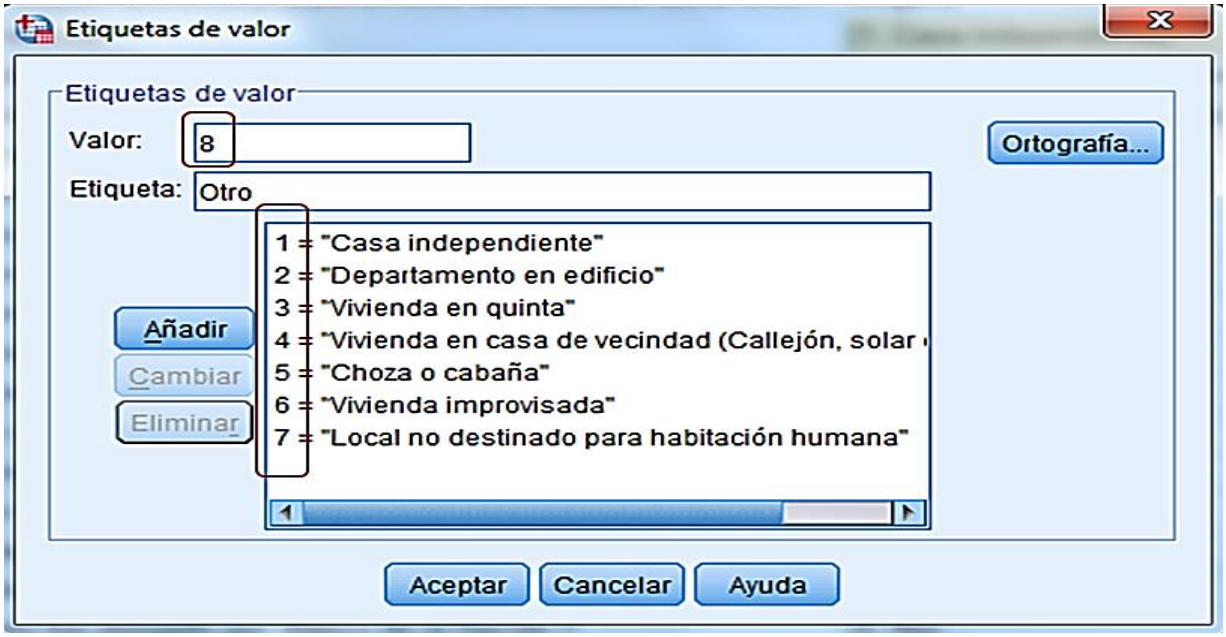
Ejemplo:



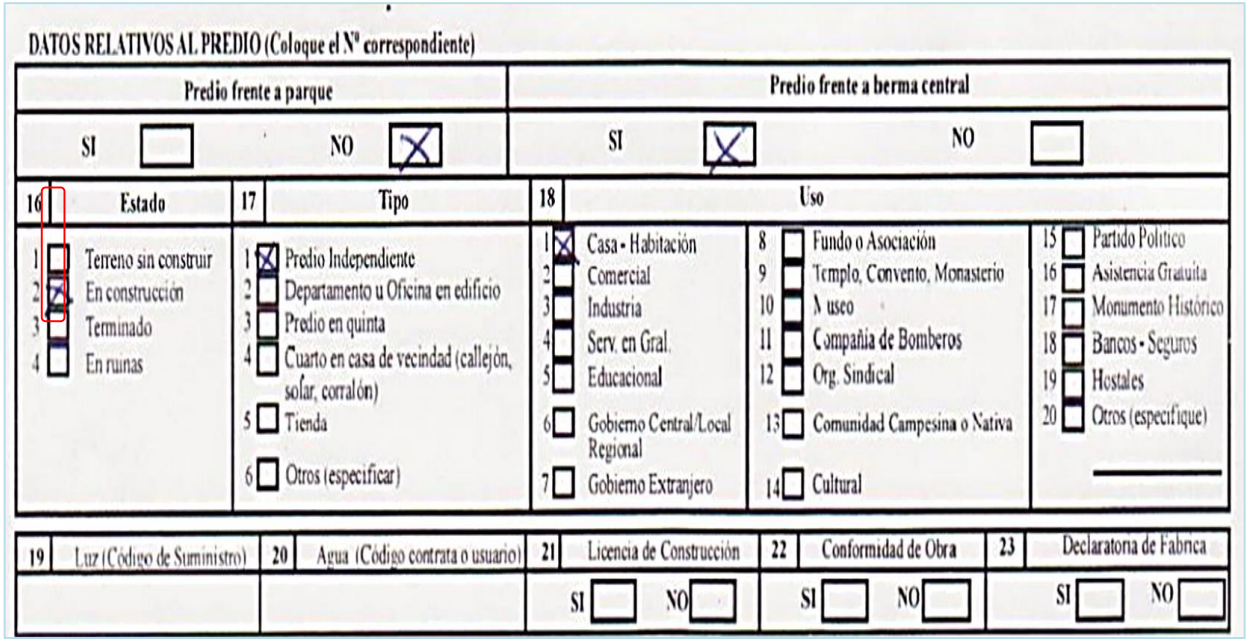
Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
P204	Numérico	1	0	¿Es miembro del hogar?	{1, Si}...	Ninguna	6	Derecha	Nominal
P205	Numérico	1	0	¿Se encuentra ausente del hogar 30 días o más?	{1, Si}...	Ninguna	6	Derecha	Nominal
P206	Numérico	1	0	¿Está presente en el hogar 30 días o más?	{1, Si}...	Ninguna	6	Derecha	Nominal
P207	Numérico	1	0	Sexo	{1, Hombre}...	Ninguna	6	Derecha	Nominal
P208A	Numérico	2	0	¿Qúe edad tiene en años cumplidos ? (En años)	Ninguna	Ninguna	7	Derecha	Escala
P208B	Numérico	2	0	¿Qúe edad tiene en años cumplidos ? (En meses)	Ninguna	Ninguna	7	Derecha	Nominal
P208A1	Numérico	1	0	Nació en este distrito	{0, Pase}...	Ninguna	8	Derecha	Nominal
P208A2	Cadena	6	0	En qué provincia y distrito nació	Ninguna		8	Izquierda	Nominal
P209	Numérico	1	0	¿Cuál es su estado civil o conyugal?	{1, Convivie...	Ninguna	6	Derecha	Nominal
P210	Numérico	1	0	La semana pasada ... ¿Estuvo trabajando o realizando alguna tarea e...	{1, Si}...	Ninguna	6	Derecha	Nominal
P211A	Numérico	2	0	La semana pasada ...¿La tarea que realizó....en el hogar o fuera de él...	{1, Ayudó e...	Ninguna	7	Derecha	Nominal
P211D	Numérico	3	0	¿Cuántas horas en total realizó estas tareas: ?	Ninguna	Ninguna	7	Derecha	Escala
P212	Numérico	2	0	Persona que le corresponde el módulo de Educación (de 3 años a m...	Ninguna	Ninguna	6	Derecha	Nominal
P213	Numérico	2	0	Persona que le corresponde el módulo de Salud (todas las personas)	Ninguna	Ninguna	6	Derecha	Nominal
P214	Numérico	2	0	Persona que le corresponde el módulo de Empleo/Ingresos (mayores...	Ninguna	Ninguna	6	Derecha	Nominal
P215	Numérico	2	0	Número de orden de la persona en el año anterior (Selección panel)	Ninguna	Ninguna	6	Derecha	Nominal
P216	Numérico	1	0	Persona nueva (Selección panel)	Ninguna	Ninguna	6	Derecha	Nominal
P217	Numérico	1	0	¿Por qué motivo ..... ya no vive en este hogar? (Selección panel)	{1, Viaje}...	9	6	Derecha	Nominal
T211	Numérico	2	0	(Recodificada) La semana pasada ...¿La tarea que realizó....en el ho...	{1, Ayudó e...	Ninguna	6	Derecha	Nominal
P211C1	Cadena	100	0	¿Qué tareas realizó?	Ninguna	Ninguna	26	Izquierda	Nominal
P211C2	Cadena	100	0	¿Qué tareas realizó?	Ninguna	Ninguna	26	Izquierda	Nominal
TICUEST01	Numérico	1	0	Origen de cuestionario	{1, Cuestion...	Ninguna	11	Derecha	Nominal
OCUPAC	Cadena	3	0	Código de tareas realizadas, según Ocupaciones	{011, OFICI...	Ninguna	12	Izquierda	Nominal
CODTAREA	Cadena	1	0	Código de tarea Peligrosa o No, según tareas realizadas	{0, No Pelig...	Ninguna	12	Izquierda	Nominal

Paso 6: El código o valor de variable se define de acuerdo con el número de las categorías que corresponda a una variable y que son ordenadas en base al diseño de las preguntas del cuestionario o ítem del formulario. Cuando son preguntas abiertas será con relación al objetivo del estudio.

Ejemplo: Tipo de vivienda tiene de 1 a 8 categorías.



Ejemplo: Estado del predio tiene de 1 a 4 categorías.



DATOS RELATIVOS AL PREDIO (Coloque el Nº correspondiente)

Predio frente a parque		Predio frente a berma central	
SI ☐	NO ☒	SI ☒	NO ☐

16 Estado
1 ☐ Terreno sin construir
2 ☒ En construcción
3 ☐ Terminado
4 ☐ En ruinas

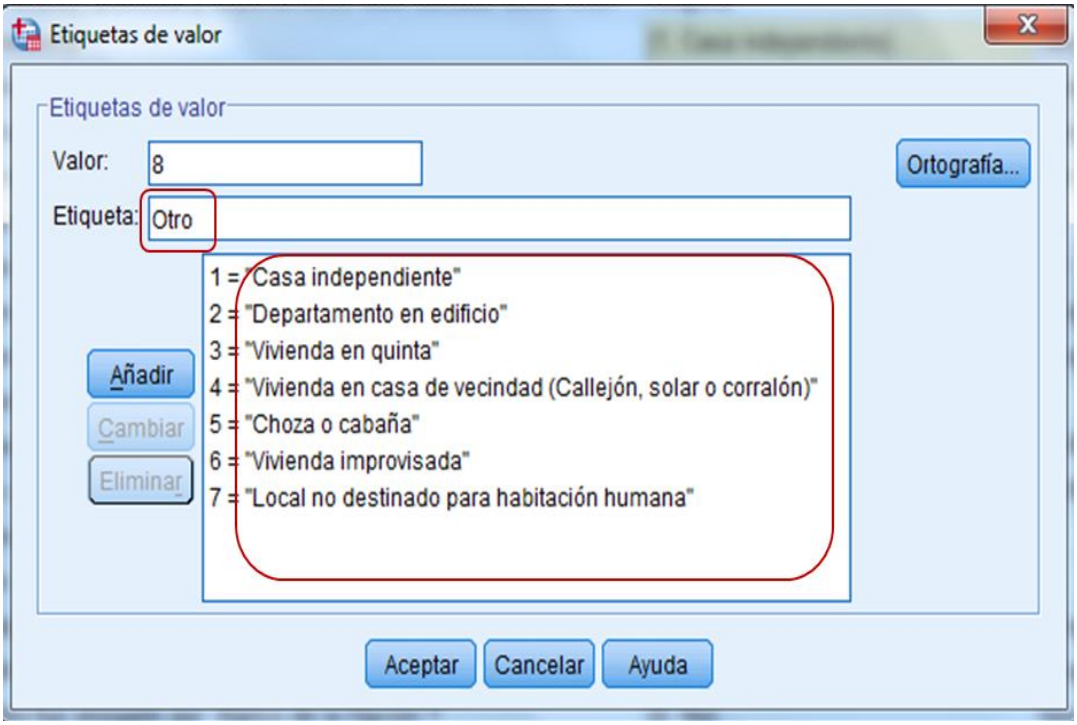
17 Tipo
1 ☒ Predio Independiente
2 ☐ Departamento u Oficina en edificio
3 ☐ Predio en quinta
4 ☐ Cuarto en casa de vecindad (callejón, solar, corralón)
5 ☐ Tienda
6 ☐ Otros (especificar)

18 Uso
1 ☒ Casa - Habitación
2 ☐ Comercial
3 ☐ Industria
4 ☐ Serv. en Gral.
5 ☐ Educacional
6 ☐ Gobierno Central/Local Regional
7 ☐ Gobierno Extranjero
8 ☐ Fundo o Asociación
9 ☐ Templo, Convento, Monasterio
10 ☐ Museo
11 ☐ Compañía de Bomberos
12 ☐ Org. Sindical
13 ☐ Comunidad Campesina o Nativa
14 ☐ Cultural
15 ☐ Partido Político
16 ☐ Asistencia Gratuita
17 ☐ Monumento Histórico
18 ☐ Bancos - Seguros
19 ☐ Hostales
20 ☐ Otros (especifique)

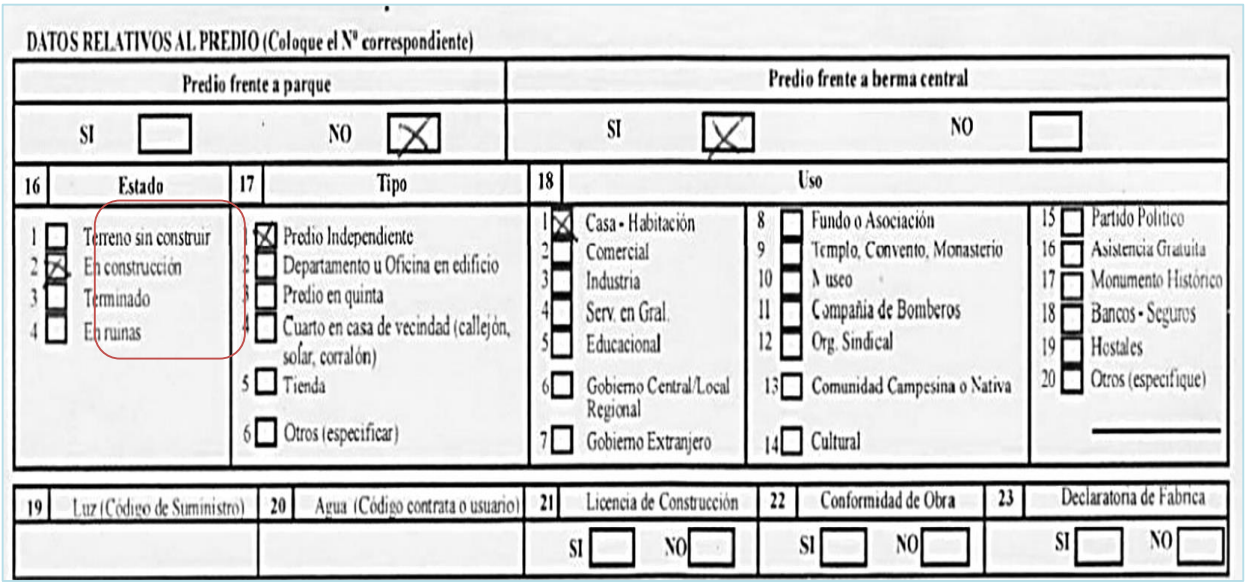
19 Luz (Código de Suministro)	20 Agua (Código contrata o usuario)	21 Licencia de Construcción	22 Conformidad de Obra	23 Declaratoria de Fábrica
		SI ☐ NO ☐	SI ☐ NO ☐	SI ☐ NO ☐

Paso 7: Nombre del código o etiqueta del valor: Comprende los nombres de las categorías de la variable.

Ejemplo: Tipo de vivienda (casa independiente, departamento en edificio, vivienda en quinta, etc.)


Etiquetas de valor
Etiquetas de valor
Valor:
8
Etiqueta:
Otro
Ortografía...
1 = "Casa independiente"
2 = "Departamento en edificio"
3 = "Vivienda en quinta"
4 = "Vivienda en casa de vecindad (Callejón, solar o corralón)"
5 = "Choza o cabaña"
6 = "Vivienda improvisada"
7 = "Local no destinado para habitación humana"
Añadir
Cambiar
Eliminar
Aceptar
Cancelar
Ayuda

Ejemplo: Estado del Predio (terreno sin construir, en construcción, terminado y en ruinas).



DATOS RELATIVOS AL PREDIO (Coloque el Nº correspondiente)

Predio frente a parque		Predio frente a berma central	
SI ☐	NO ☒	SI ☒	NO ☐

16 Estado	17 Tipo	18 Uso		
1 ☐ Terreno sin construir 2 ☒ En construcción 3 ☐ Terminado 4 ☐ En ruinas	1 ☒ Predio Independiente 2 ☐ Departamento u Oficina en edificio 3 ☐ Predio en quinta 4 ☐ Cuarto en casa de vecindad (callejón, solar, corralón) 5 ☐ Tienda 6 ☐ Otros (especificar)	1 ☒ Casa - Habitación 2 ☐ Comercial 3 ☐ Industria 4 ☐ Serv. en Gral. 5 ☐ Educacional 6 ☐ Gobierno Central/Local Regional 7 ☐ Gobierno Extranjero	8 ☐ Fundo o Asociación 9 ☐ Templo, Convento, Monasterio 10 ☐ Museo 11 ☐ Compañía de Bomberos 12 ☐ Org. Sindical 13 ☐ Comunidad Campesina o Nativa 14 ☐ Cultural	15 ☐ Partido Político 16 ☐ Asistencia Gratuita 17 ☐ Monumento Histórico 18 ☐ Bancos - Seguros 19 ☐ Hostales 20 ☐ Otros (especifique)

19 Luz (Código de Suministro)	20 Agua (Código contrata o usuario)	21 Licencia de Construcción	22 Conformidad de Obra	23 Declaratoria de Fábrica
		SI ☐ NO ☐	SI ☐ NO ☐	SI ☐ NO ☐

Paso 8: Determinar si existen valores perdidos que definen los valores de los datos identificados como perdidos por el usuario. Los valores de datos que se especifican como perdidos por el usuario aparecen marcados para un tratamiento especial, es decir, usar un valor que represente los datos perdido; además, se excluyen de la mayoría de los cálculos.

2.3.2. Diseño del diccionario de datos o variables

Todo archivo de base de datos debe tener un nombre y descripción sobre el contenido de los datos.

Ejemplo:

Nombre de archivo SPSS	Descripción
ENAHO01-2016-100.SAV	Características de la vivienda y del hogar
ENAHO01-2016-200.SAV	Características de los miembros del hogar
ENAHO01-2016-601.SAV	Gastos del Hogar
ENAHO01-2016-602.SAV	Alimentos para consumir dentro del hogar
ENAHO01-2016-603.SAV	Mantenimiento de la vivienda

Nombre de archivo EXCEL	Descripción
FUGAS_2014_2016.XLSX	Fugas intramuro y extramuro del penal
INTERNOS FALLECIDOS_2014_2016.XLSX	Internos fallecidos por actos violentos
INTRAMUROS_FALLECIDOS_2014_2016. XLSX	Intramuros_Causa de fallecimiento
EXTRAMUROS_FALLECIDOS_2014_2016. XLSX	Extramuros_Causa de fallecimiento
ENAHO01-2016-603.SAV	Mantenimiento de la vivienda

LISTADO DEL DICCIONARIO DE DATOS

A. De Archivo SPSS

Para obtener el diccionario de datos de un archivo SPSS, será a través del menú: Archivo / Mostrar información del archivo de datos, eligiendo archivo de trabajo o usando el siguiente comando:

GET
FILE='C: \Archivos de programa\SPSS\Enaho01-100.sav'.
DISPLAY DICCTIONARY o DISPLAY SORTED DICCTIONARY.

Primer momento: Se ejecuta la sentencia y en el visor de resultado o output se muestra lo siguiente:

Información sobre las variables									
Variable	Ubica-ción	Etiqueta	Nivel de medida	Papel	Ancho de columna	Alineación	Formato de impresión	Formato de escritura	Valores perdidos
AÑO	1	Ámbito geográfico	Nominal	Entrada	19	Derecha	F4	F4	
MES	2	Mes de Ejecución de la Encuesta	Nominal	Entrada	2	Izquierda	A2	A2	
CONGLOME	3	Conglomerado	Nominal	Entrada	7	Izquierda	A4	A4	
VIVIENDA	4	Número de Selección de Vivienda	Nominal	Entrada	5	Izquierda	A3	A3	
HOGAR	5	Hogar	Nominal	Entrada	5	Izquierda	A2	A2	
CODPERSO	6	Código de persona	Nominal	Entrada	5	Izquierda	A2	A2	
UBIGEO	7	Código de Ubicación Geográfica	Nominal	Entrada	6	Izquierda	A6	A6	
DOMINIO	8	Dominio	Nominal	Entrada	8	Derecha	F1	F1	
ESTRATO	9	Estrato Geográfico	Nominal	Entrada	8	Derecha	F1	F1	
CODINFOR	10	Código del informante	Nominal	Entrada	2	Izquierda	A2	A2	
P500N	11	Código de Persona	Nominal	Entrada	8	Izquierda	A2	A2	
P500I	12	Código de la Persona Informante	Nominal	Entrada	8	Izquierda	A2	A2	
P500A	13	Período de referencia (día)	Nominal	Entrada	8	Izquierda	A2	A2	
P500B	14	Período de referencia (mes)	Nominal	Entrada	8	Izquierda	A2	A2	
P500B1	15	Período de referencia (año)	Nominal	Entrada	8	Izquierda	A4	A4	
P500C	16	Período de referencia (día)	Nominal	Entrada	8	Izquierda	A2	A2	
P500D	17	Período de referencia (mes)	Nominal	Entrada	8	Izquierda	A2	A2	
P500D1	18	Período de referencia (año)	Nominal	Entrada	8	Izquierda	A4	A4	
P501	19	La semana pasada tuvo Ud., algún trabajo?	Nominal	Entrada	8	Derecha	F1	F1	9

Segundo momento: Debajo del primer resultado, aparece los valores de las variables:

Valores de las variables		
Valor		Etiqueta
DOMINIO	1	Costa Norte
	2	Costa Centro
	3	Costa Sur
	4	Sierra Norte
	5	Sierra Centro
	6	Sierra Sur
	7	Selva
	8	Lima Metropolitana
ESTRATO	1	Mayor de 100,000 viviendas
	2	De 20,001 a 100,000 viviendas
	3	De 10,001 a 20,000 viviendas
	4	De 4,001 a 10,000 viviendas
	5	De 401 a 4,000 viviendas
	6	Menos de 401 viviendas
	7	Área de Empadronamiento Rural - AER Compuesto
	8	Área de Empadronamiento Rural - AER Simple
P501	1	Si
	2	No

B. Diccionario de datos editado

Luego de obtener en el visor los resultados completos se procede a editar las tablas, teniendo en cuenta:

- Poner título al diccionario
- Ordenar las variables con sus componentes

Título ENAHO01-2017-100.SAV : Características de la Vivienda y del Hogar (Módulo 100)
Archivo: ENAHO01-2017-100.SAV

DICCIONARIO DE DATOS				
NOMBRE	**TAMAÑO**	**DECIMALES**	**FORMATO**	**ETIQUETA**
AÑO	4	0	F	Ámbito geográfico
MES	2	0	A	Mes de Ejecución de la Encuesta
CONGLOME	4	0	A	Conglomerado
VIVIENDA	3	0	A	Número de Selección de Vivienda
HOGAR	2	0	A	Hogar
CODPERSO	2	0	A	Código de persona
UBIGEO	6	0	A	Código de Ubicación Geográfica
DOMINIO	1	0	F	**Dominio geográfico**
				1 Costa Norte
				2 Costa Centro
				3 Costa Sur
				4 Sierra Norte
				5 Sierra Centro
				6 Sierra Sur
				7 Selva
				8 Lima Metropolitana
ESTRATO	1	0	F	**Estrato Geográfico**
				1 Mayor de 100,000 viviendas
				2 De 20,001 a 100,000 viviendas
				3 De 10,001 a 20,000 viviendas
				4 De 4,001 a 10,000 viviendas
				5 De 401 a 4,000 viviendas
				6 Menos de 401 viviendas
				7 Área de Empadronamiento Rural - AER Compuesto
				8 Área de Empadronamiento Rural - AER Simple
CODINFOR	2	0	A	Código del informante
P500N	2	0	A	Código de Persona
P500I	2	0	A	Código de la Persona Informante
P500A	2	0	A	Período de referencia (día)
P500B	2	0	A	Período de referencia (mes)
P500B1	4	0	A	Período de referencia (año)
P500C	2	0	A	Período de referencia (día)
P500D	2	0	A	Período de referencia (mes)
P500D1	4	0	A	Período de referencia (año)
P501	1	0	F	**La semana pasada tuvo Ud., algún trabajo?**
				1 Si
				2 No

C. Archivo STATA

En el caso de un archivo STATA, el diccionario de datos será a través de la siguiente ruta:

CD="C: \Archivos de programa\Enaho"

Use enaho01-2017-200.dta, clear

DESCRIBE o **DES**

variable name	storage type	display format	value label	variable label
identh03	str11	%11s		
identi03	str13	%13s		
aÑo	str4	%4s		año
mes	str2	%2s		mes
conglome	str6	%6s		conglomerado
vivienda	str3	%3s		n° de selección de la vivienda
hogar	str2	%2s		hogar
ubigeo	str6	%6s		código de distrito (ubicación geográfica)
dominio	byte	%8.0g	dominio	dominio
estrato	byte	%8.0g	estrato	estrato
codperso	str2	%2s		n° de orden de la persona
p201p	str17	%17s		código panel de persona
p203	byte	%8.0g	p203	relación de parentesco con el jefe del hogar
p203a	byte	%8.0g		número del núcleo familiar
p203b	byte	%8.0g	p203b	parentesco con el jefe del núcleo familiar
p204	byte	%8.0g	p204	¿es miembro del hogar?
p205	byte	%8.0g	p205	¿se encuentra ausente del hogar 30 días o más?
p206	byte	%8.0g	p206	¿está presente en el hogar 30 días o más?
p207	byte	%8.0g	p207	sexo
p208a	byte	%8.0g		¿ qúe edad tiene en año cumplidos ? (en años)
p208b	byte	%8.0g		¿ qúe edad tiene en año cumplidos ? (en meses)
p208a1	byte	%8.0g	p208a1	nació en este distrito
p208a2	str6	%6s		en qué provincia y distrito nació
p209	byte	%8.0g	p209	¿cuál es su estado civil o conyugal?
p210	byte	%8.0g	p210	la semana pasada ... ¿estuvo trabajando o realizando alguna labor?
p211a	byte	%8.0g	p211a	actividad de la semana pasada
p211d	byte	%8.0g		¿ cuántas horas en total realizó estas tareas: ?
p212	byte	%8.0g		persona que le corresponde el módulo de educación (de 3 años a más)
p213	byte	%8.0g		persona que le corresponde el módulo de salud (todas las personas)
p214	byte	%8.0g		persona que le corresponde el módulo de empleo/ingresos (mayores de 14 años)
p215	byte	%8.0g		número de orden de la persona en el año anterior (selección panel)
p216	byte	%8.0g		persona nueva (selección panel)

D. <u>Otros archivos</u>

El diccionario de variables debe contener información sobre definiciones de variables de acuerdo a la base de datos:

- Nombres de variable / Sinónimo o alias
- Etiqueta o descripción de la variable
- Código
- Nombre del código

LISTADO DEL DICCIONARIO DE VARIABLES

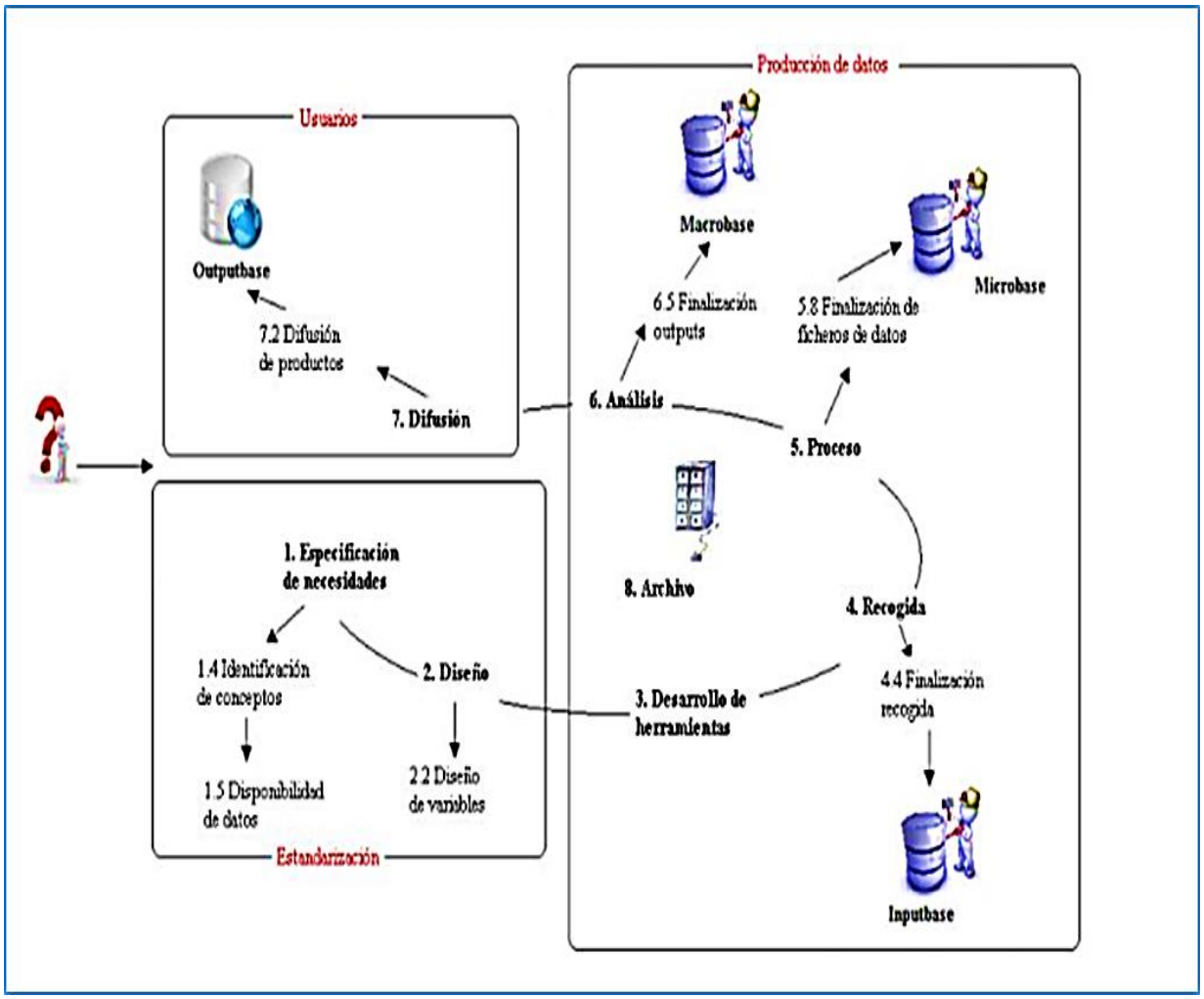
DICCIONARIO DE VARIABLES: REGISTRO NACIONAL DE DEFUNCIONES			
Variable	**Descripción de la variable**	**Cód.**	**Nombre del código**
ID_N	ID IDENTIFICATORIO		
UID_N	ID UNICO IDENTIFICATORIO		
AÑO_RENADE	AÑO RENADE		
CODREGISTR	CÓDIGO REGISTRO		
CODDISA	CÓDIGO DE DISA		
CORRED	CÓDIGO DE RED		
CODMICRED	CÓDIGO DE MICRORRED		
PTODIGI	PUNTO DE DIGITACIÓN		
NUMSERIE	NÚMERO DE SERIE DEL CERTIFICADO		
RG11DPTO	DEPARTAMENTO DONDE SE PROCESA EL INFORME ESTADÍSTICO		
RG12PROV	PROVINCIA DONDE SE PROCESA EL INFORME ESTADÍSTICO		
RG13DIST	DISTRITO DONDE SE PROCESA EL INFORME ESTADÍSTICO		
RG14CODLOC			
RG14DESLOC	LOCALIDAD DONDE SE PROCESA EL INFORME ESTADÍSTICO		
RG15LIBRO			
RG16ACTA			
RG17FECHA			
ID21NOMBRE	NOMBRES DEL FALLECIDO		
ID22APEPAT	APELLIDO PATERNO DEL FALLECIDO		
ID23APEMAT	APELLIDO MATERNO DEL FALLECIDO		
ID24APECAS	APELLIDO DE CASADA DE LA FALLECIDA		
ID25TIPDOC	TIPO DE DOCUMENTO DE IDENTIDAD	1	DNI
		2	LIBRETA MILITAR
		3	CARNÉ FF.AA/PNP
		4	PASAPORTE
		5	CARNÉ DE EXTRANJERÍA
		6	PARTIDA DE NACIMIENTO
		7	OTRO
		9	IGNORADO
ID25NUMDOC	NÚMERO DE DOCUMENTO DE IDENTIDAD		
DA31SEXO	SEXO DEL FALLECIDO	1	HOMBRE
		2	MUJER
		9	IGNORADO
DA32EDAD	EDAD DEL FALLECIDO		
DA32TIEM	TIPO EDAD	1	AÑOS
		2	MESES
		3	DÍAS
		4	HORAS
		9	IGNORADO
DA33CONYU	ESTADO CONYUGAL/MARITAL	1	CONVIVIENTE
		2	CASADO
		3	VIUDO
		4	DIVORCIADO
		5	SEPARADO
		6	SOLTERO
		9	IGNORADO

Cabe señalar que todo software estadístico tiene rutas para disponer de un diccionario de datos.

2.4. Base de datos

Las bases de datos se consideran una parte importante del proceso estadístico y dependerán del modelo que se adopte en la institución. Para elaborar se consideran diferentes elementos, como los datos, las metodologías usadas, las herramientas informáticas, etc., y para garantizar que esta información se produzca, elabore y difunda adecuadamente, es necesario disponer de bases de datos.

Las bases de datos en el proceso de producción estadística:



La estructura de una base de datos hace referencia a los tipos de datos, los vínculos o relaciones y las restricciones que deben cumplir estos datos; es diseñada empleando algún tipo de modelo de datos.

Cada fila de una tabla se llama “registro o casos”. Los registros incluyen datos sobre algo o alguien específico. En cambio, las columnas (conocidas como campos o atributos o variables) contienen un único tipo de información que aparece en cada registro, como nombre y apellidos, las direcciones, sexo, edad.

Con el fin de que los datos sean consistentes de un registro al siguiente, se asigna el tipo de datos apropiado a cada columna.

COLUMNAS

	Variable 1	Variable 2	Variable 3	Variable 4
Registro 1	1,1			
Registro 2	2,1	2,2	2,3	2,4
Registro 3	3,1			
Registro 4	4,1			
Registro 5	5,1			
Registro 6	6,1			

FILAS

2.5. Tratamiento de las unidades de registro

Para obtener una base de registros coherentes y consistentes, en primer lugar, se debe desarrollar el proceso de depuración y consistencia de variables, es decir, verificar la duplicidad de información, identificar omisiones y finalmente determinar la existencia de datos atípicos.

2.5.1. Revisión de duplicados

La existencia de unidades duplicadas en un registro administrativo o encuesta produce errores en los resultados; es por ello, la importancia de que estos casos sean detectados antes de realizar cualquier otro análisis, empezando con asegurar que las variables de identificación sean únicas, ya sea registros a nivel de vivienda, hogares o persona; obtenida una base de datos con registros únicos, se podrá identificar la duplicidad en el resto de variables, que servirán para los análisis posteriores.

Variables de identificación de registros únicos

Vivienda:

- Año
- Mes
- Conglomerado
- Vivienda

Hogar:

- Año
- Mes
- Conglomerado
- Vivienda
- Hogar

Persona:

Año
Mes
Conglomerado
Vivienda
Hogar
Persona

2.5.2. Revisión de omisiones

Para el caso de las omisiones, se debe verificar la existencia de la falta de información de una característica necesaria en la variable; el primer paso, es tener en cuenta los flujos (pases) y establecer un techo para cada variable, obtenido esto se puede generar los reportes y completar la información faltante en la medida de lo posible.

Caso práctico: Analizar base de datos con registros a nivel de vivienda usando el paquete estadístico SPSS.

Duplicados en las variables de identificación:

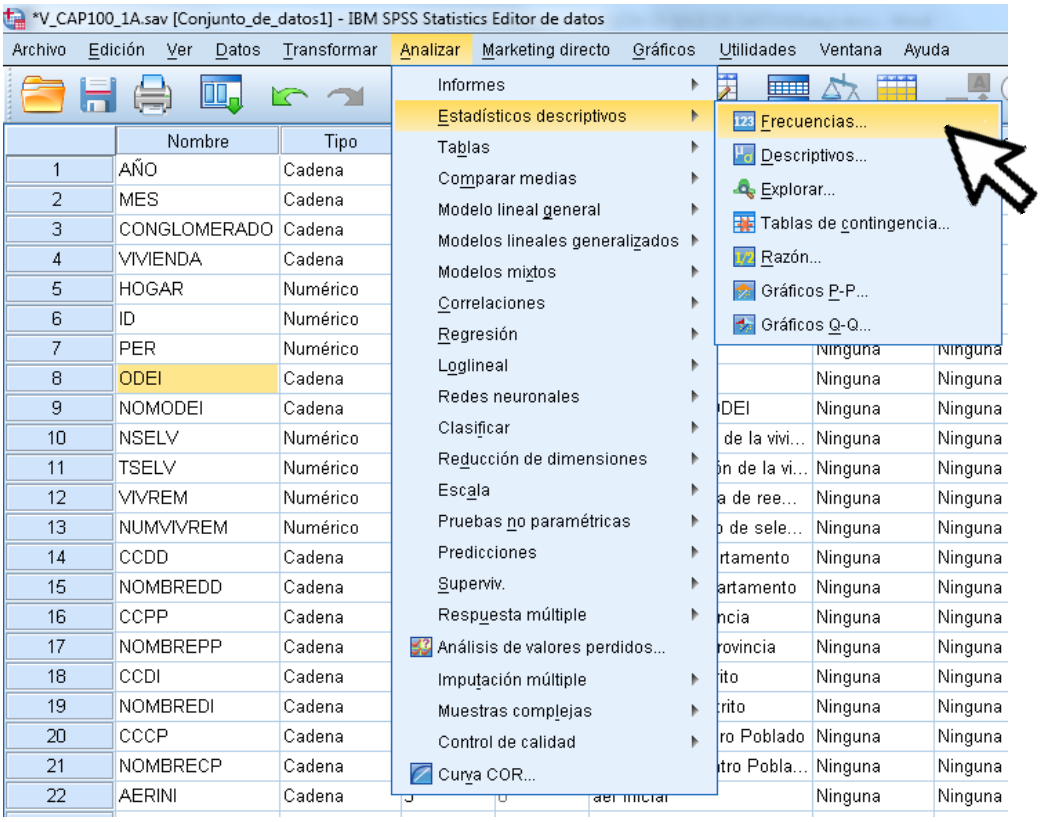
Paso 1: Listar las variables de identificación, para identificar los duplicados de las variables de identificación, primero es reconocer y listar las variables que forman parte de la identificación, por ejemplo:

Variables para la identificación de la vivienda:

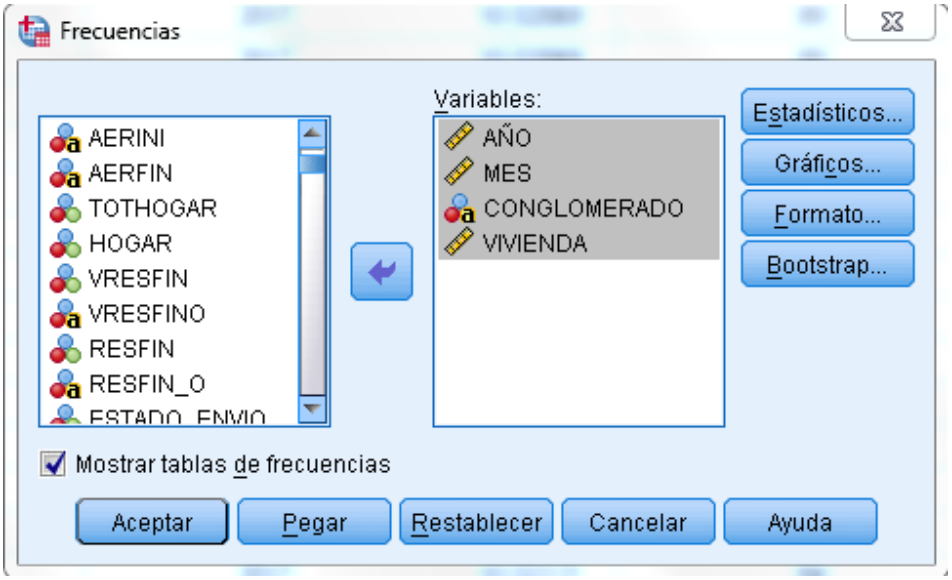
Año
Mes
Conglomerado
Vivienda

Paso 2: Frecuencias de cada variable, se debe realizar una frecuencia simple a cada variable para verificar que no existan omisiones, toda variable correspondiente a las variables de identificación deben ser completas.

Se analizará una base de datos con registros a nivel de vivienda, usando el paquete estadístico SPSS:
Ir al menú: Analizar / Estadísticos descriptivos / Frecuencias



Luego seleccionamos las variables de identificación: Año, mes, conglomerado y vivienda y pulsamos aceptar.



Paso 3: Análisis de los resultados, evaluar las variables con información faltante, en este caso de los resultados se observa que de las cuatro variables de identificación seleccionadas, dos de ellas presentan omisiones en el año y mes, con 15 registros omitidos para cada variable

Resultados:

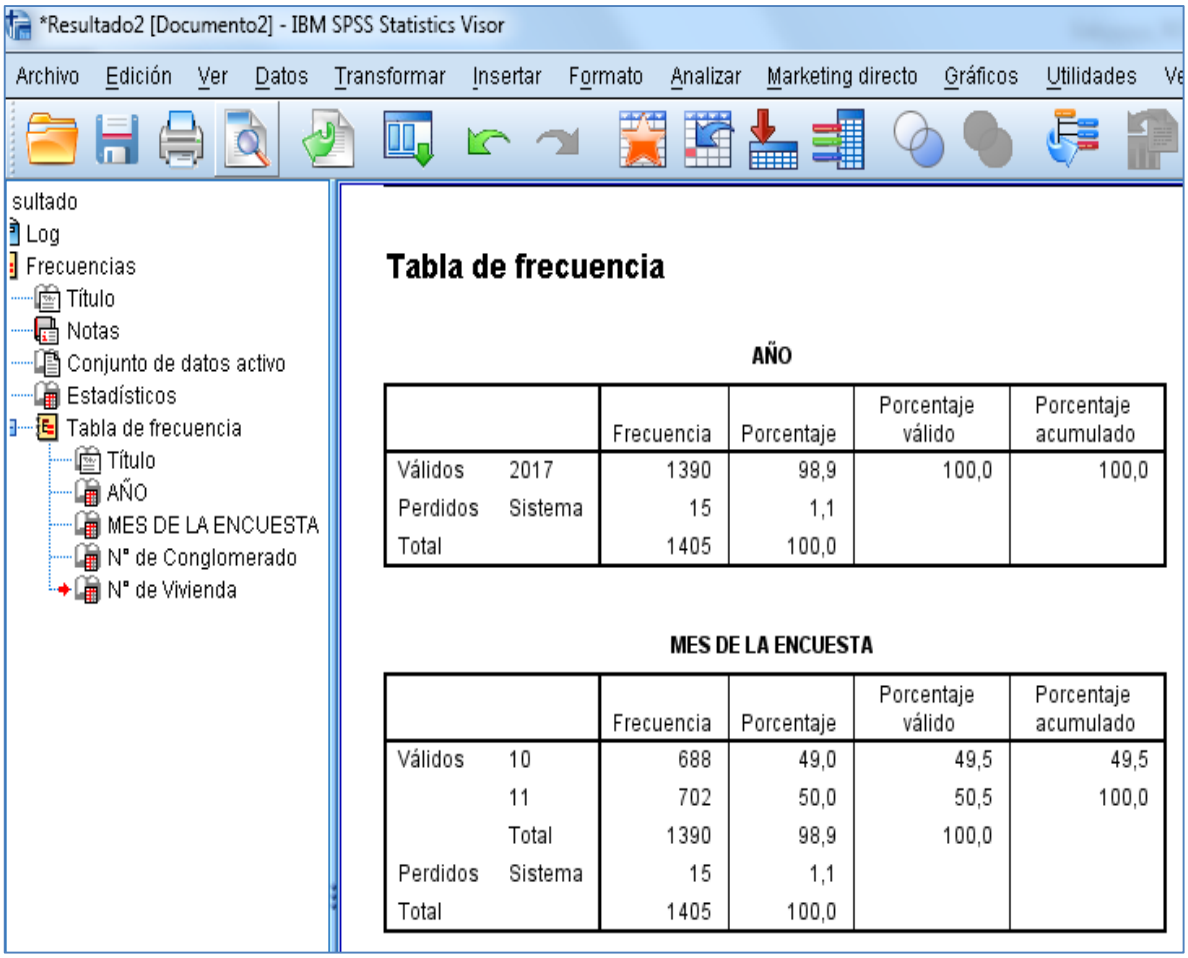


Tabla de frecuencia

AÑO

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	2017	1390	98,9	100,0	100,0
Perdidos	Sistema	15	1,1		
Total		1405	100,0		

MES DE LA ENCUESTA

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	10	688	49,0	49,5	49,5
	11	702	50,0	50,5	100,0
	Total	1390	98,9	100,0	
Perdidos	Sistema	15	1,1		
Total		1405	100,0		

Paso 4: Evaluación y corrección de los casos encontrados, visualizar los casos encontrados y verificar, si es posible corregirlas.

Para este caso, si es posible corregirlas, ya que tenemos un formato de programación de distribución de la carga de trabajo.

	AÑO	MES	CONGLOMERADO	VIVIENDA
1	.	.	08693	4
2	.	.	14714	11
3	.	.	16098	10
4	.	.	16098	11
5	.	.	16227	12
6	.	.	33144	12
7	.	.	33349	12
8	.	.	36351	9
9	.	.	37452	12
10	.	.	38439	7
11	.	.	45514	9
12	.	.	06755	12
13	.	.	08759	11
14	.	.	36558	12
15	.	.	36971	12
16	2017	10	02069	1
17	2017	10	02069	2
18	2017	10	02069	3
19	2017	10	02069	4
20	2017	10	02069	5

Programación de distribución de la carga de trabajo.

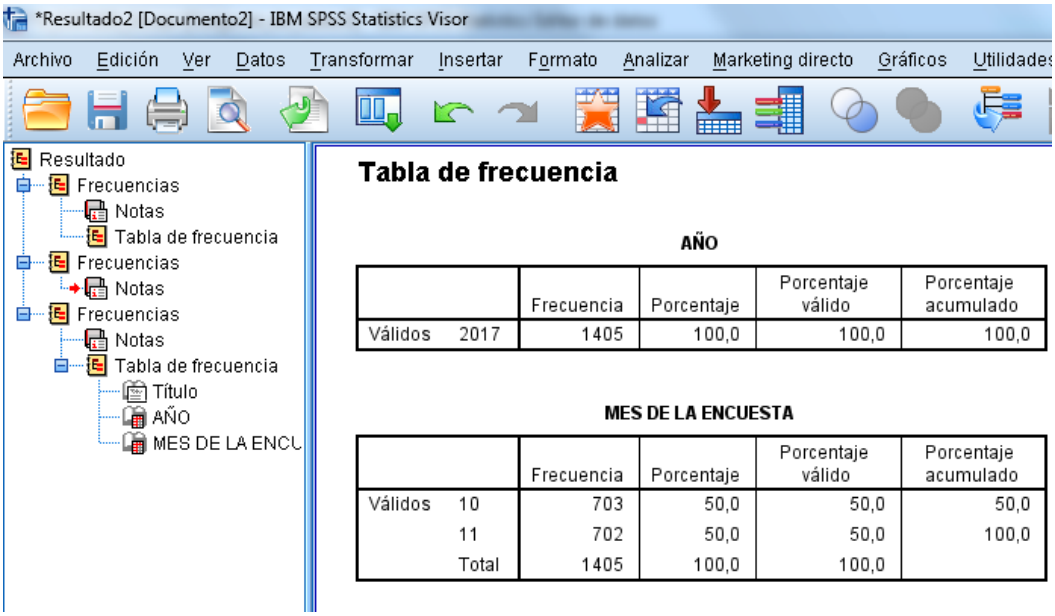
ODEI	AÑO	MES	PERIODO	CONGLOMERADO
AMAZONAS	2017	10	1	8693
AMAZONAS	2017	10	2	14714
AMAZONAS	2017	10	2	16098
AMAZONAS	2017	10	3	16098
AMAZONAS	2017	10	2	16227
AMAZONAS	2017	10	3	33144
AMAZONAS	2017	10	1	33349
AMAZONAS	2017	10	2	36351
AMAZONAS	2017	10	2	37452
AMAZONAS	2017	10	3	38439
AMAZONAS	2017	10	1	45514
AMAZONAS	2017	10	2	6755
AMAZONAS	2017	10	3	8759
AMAZONAS	2017	10	3	36558
LORETO	2017	10	1	36971

Contrastar con el documento y completar la información.

Luego la base de datos quedaría de la manera siguiente:

	AÑO	MES	CONGLOMERADO	VIVIENDA
1	2017	10	08693	4
2	2017	10	14714	11
3	2017	10	16098	10
4	2017	10	16098	11
5	2017	10	16227	12
6	2017	10	33144	12
7	2017	10	33349	12
8	2017	10	36351	9
9	2017	10	37452	12
10	2017	10	38439	7
11	2017	10	45514	9
12	2017	10	06755	12
13	2017	10	08759	11
14	2017	10	36558	12
15	2017	10	36971	12
16	2017	10	02069	1
17	2017	10	02069	2
18	2017	10	02069	3
19	2017	10	02069	4
20	2017	10	02069	5

Comprobar sacando frecuencias nuevamente a las variables identificadoras y observar que estas variables estén completas.



Paso 5: Verificar el tipo y ancho que debe tener la variable, usando el diccionario de variables, se debe contrastar la base de datos con el diccionario de variables, verificar el tipo y ancho de la variable, así como la etiqueta de la variable, véase el ejemplo:

Base de datos: Las variables se visualizan en la pestaña inferior "vista de variables"


Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades

	Nombre	Tipo	Anchura	Decimales	Etiqueta
1	AÑO	Numérico	11	0	AÑO
2	MES	Numérico	11	0	MES DE LA ENCUESTA
3	CONGLOMERADO	Cadena	5	0	N° de Conglomerado
4	VIVIENDA	Numérico	11	0	N° de Vivienda

Diccionario de variables:

DICCIONARIO DE VARIABLES

CAPÍTULO 100: CARACTERÍSTICAS DE LA VIVIENDA

Nro.	Nombre del Campo	Descripción	Valor	Tipo	Longitud
1	AÑO	AÑO		AN	4
2	MES	MES DE LA ENCUESTA		AN	2
3	CONGLOMERADO	N° de conglomerado		AN	5
4	VIVIENDA	N° de Vivienda		AN	6

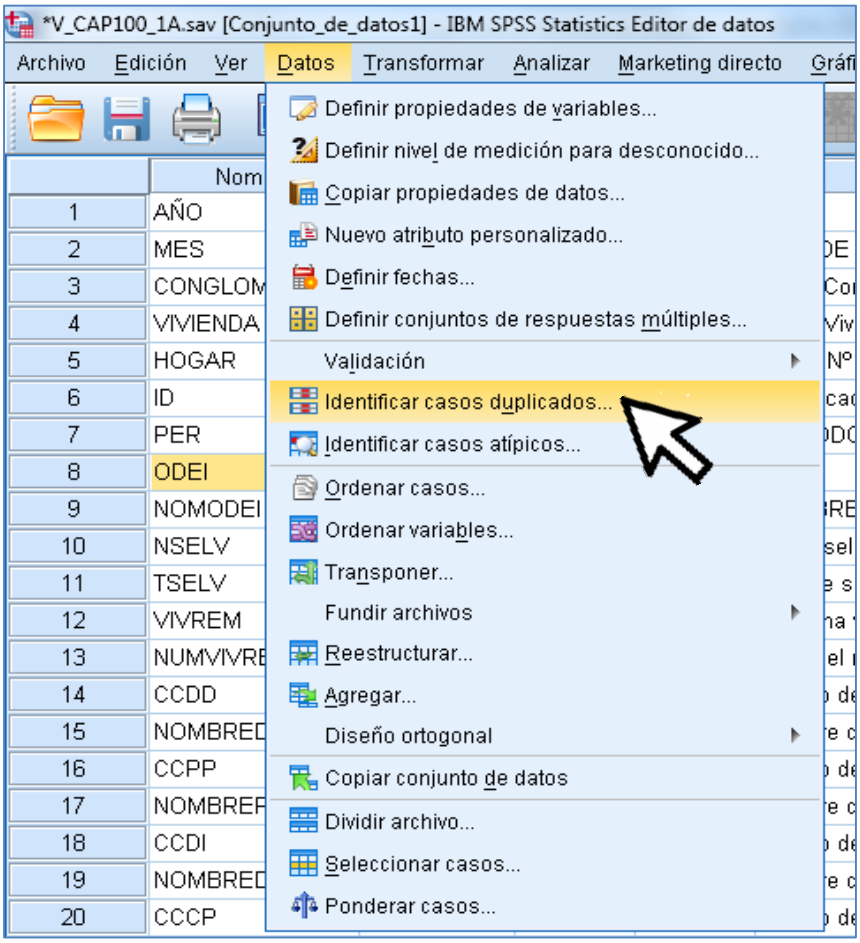
Corrigiendo el tipo y ancho de las variables se obtiene la base de datos con variables de identificación que cumpla con lo establecido.


Archivo Edición Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades

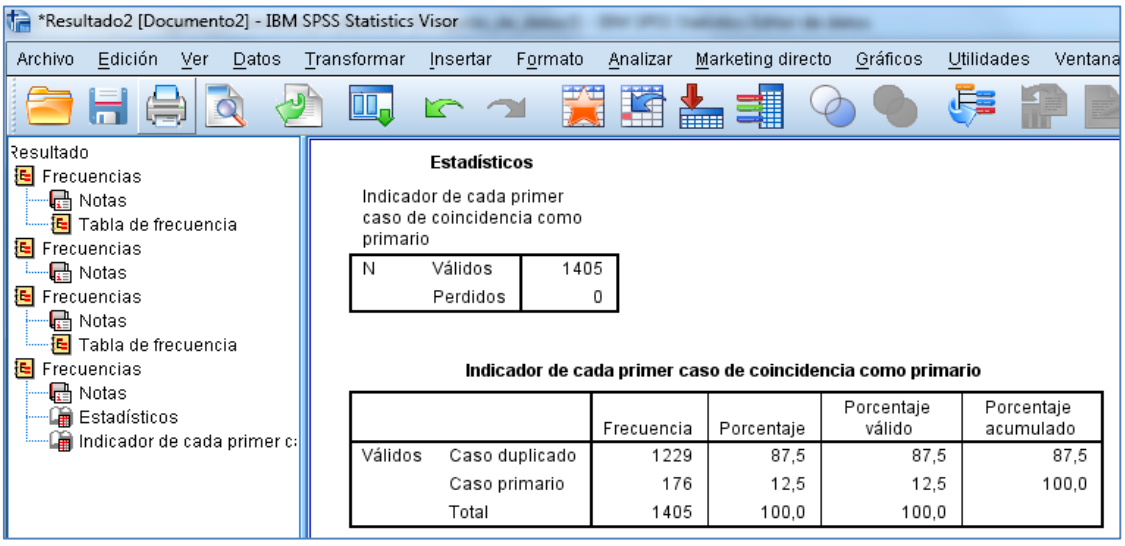
	Nombre	Tipo	Anchura	Decimales	Etiqueta
1	AÑO	Cadena	4	0	AÑO
2	MES	Cadena	2	0	MES DE LA ENCUESTA
3	CONGLOMERADO	Cadena	5	0	N° de Conglomerado
4	VIVIENDA	Cadena	6	0	N° de Vivienda

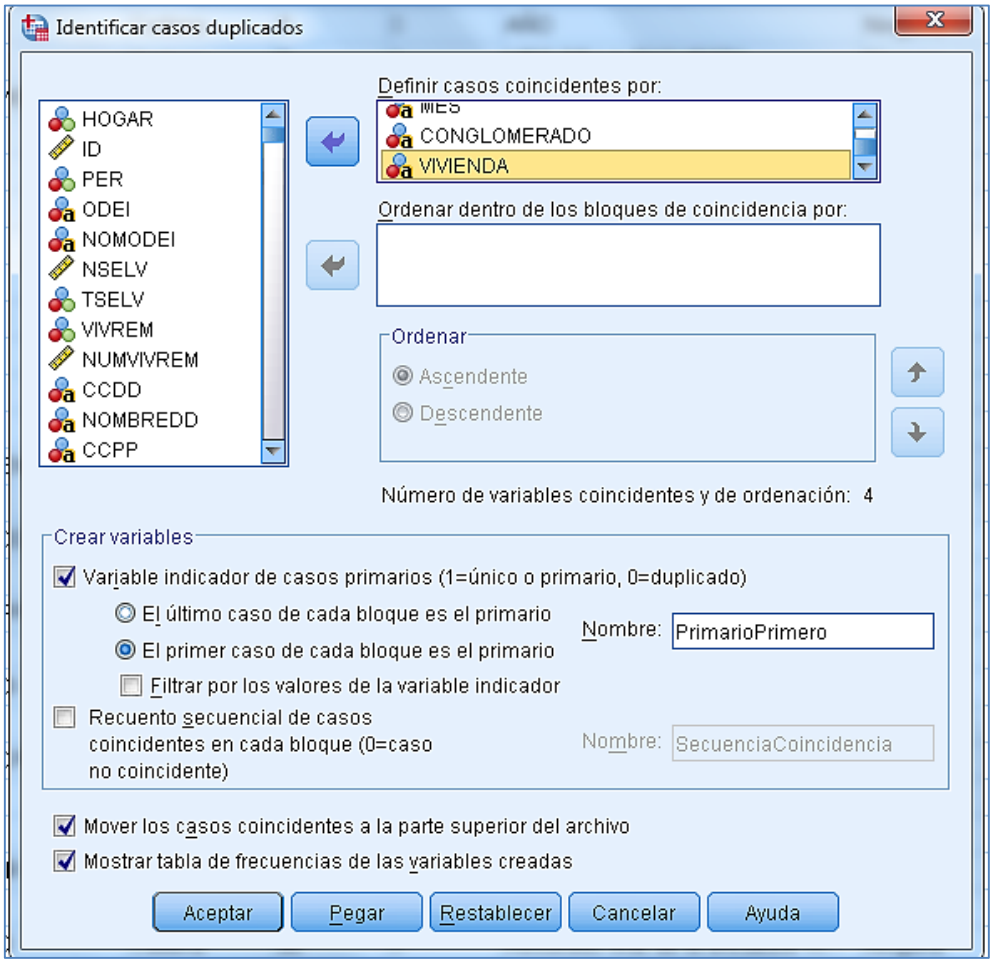
Vista de datos | Vista de variables

Paso 6: Verificar la existencia de duplicidad en la variable de identificación, siguiendo con el ejemplo usando el programa estadístico SPSS, se va al menú: Datos / identificar casos duplicados



Se introduce las variables de identificación





- Existen 176 casos duplicados.

Luego de ello, ver registros y analizar estos casos, una vez que se tiene la base de datos con variables de identificación únicas y sin omisiones, además de que esté debidamente etiquetada, tenga el tipo de variable y ancho establecido, recién ahí se podrá analizar todas las demás variables que servirán para los análisis respectivos.

2.6. Tratamiento de las variables

Una variable según Gonzáles (2004):
Es la característica de la muestra o población que se está estudiando. Los datos son el producto de su medición sobre los elementos o sujetos de estudio. (p.1).
Al conjunto de datos que puede tomar una variable se le denomina escala.

Ejemplos:

- ✓ Variable: SEXO
  Valores: Hombre y Mujer
- ✓ Variable: EDAD

Valores: 0, 1, 2, 3, ….20, 25

Se debe tener en cuenta que antes de realizar el tratamiento de las variables, deben estar bien definidas y medidas. Si se hace una definición incorrecta o una mala medición, la parte estadística elaborada hacia adelante estará incorrecta.

Tipo de variable

El tipo de variable es importante porque afecta al tipo de análisis. Asimismo, los métodos estadísticos que se utilicen dependen del tipo de variable.

Las variables pueden clasificarse por diferentes criterios, pero según su medición existen dos tipos de variables:

Variables cualitativas o categóricas

- Son aquellas variables que pueden tomar como valores cualidades o categorías.
- Las categorías son valores diferentes por una cualidad, no por una cantidad.
- Ningún valor es mayor o menor que otro.
- Se pueden dividir en dos tipos: nominales y ordinales.

Variables cualitativas nominales

Son aquellas que presentan modalidades no numéricas y que no admiten un criterio de orden.

✓ Variable: Sexo

Hombre
Mujer

✓ Variable: Estado Civil
Conviviente
Separado
Casado
Viudo
Divorciado
Soltero

✓ Variable: Color
Blanco
Rojo
Azul

Variables cualitativas ordinales

Son aquellas que se les clasifica dentro de un orden o jerarquía. También se les conoce como variables semi-cuantitativas.

Ejemplos:

- ✓ Variable: Salud
  Buena
  Regular
  Mala

- ✓ Variable: Clase social
  Alta
  Media
  Baja

- ✓ Variable: Medalla deportiva
  Oro
  Plata
  Bronce

Variables cuantitativas o numéricas

- Son aquellas variables que toman valores numéricos
- Cada valor posible es mayor o menor que otro
- El conjunto de valores forma una escala de intervalo (distancia entre intervalos)
- Se puede calcular la distancia o intervalo entre cualquier par de valores de la variable.

Las variables cuantitativas se clasifican según el número de valores que puede tomar la variable:

Variables cuantitativas discretas

Son aquellas variables que solo toman valores enteros.

Ejemplos:

- ✓ Variable: Número de casas
  1, 2,3,…..
- ✓ Variable: Hijos por familia
  0, 1,2; 3;….

Variables cuantitativas continuas

Son aquellas variables que pueden tomar valor dentro de un intervalo de valores determinado.

Ejemplos:

- ✓ Variable: Peso
  52,64 kg; 90,20 kg;

- ✓ Variable: Talla 152,15 cm; 180,20 cm

2.6.1 Revisión de valores extremos o atípicos (categoría fuera de rango)

Para poder llevar a cabo la revisión de las inconsistencias de las variables de una determinada base de datos, se debe tener en cuenta lo siguiente:

- Revisar el diccionario de datos el cual nos va a permitir apreciar el nombre y el tipo de variable, la etiqueta, así como los valores y/o categorías que tiene cada una de las variables.
- Realizar frecuencias y cruces de variables con la finalidad de ver los valores máximos o mínimos, si las variables cuentan con las categorías correctas, datos vacíos o alguna otra inconsistencia.

Frecuencia

Ejemplo:

Tomando como referencia la base de datos de defunciones, se realizará una frecuencia de las variables categóricas como DA31SEXO, DA33CONYU y DA34CODNIV.

De acuerdo al diccionario de datos se tiene lo siguiente:

Nombre de la variable: DA31SEXO
Tipo: Cadena
Etiqueta: Sexo del fallecido
Valores o categorías:

1. Hombre
2. Mujer
9. Ignorado

DA31SEXO Sexo del fallecido

	Frecuencia	Porcentaje
	9	0,0
1 Hombre	51 976	53,5
2 Mujer	44 911	46,2
9	57	0,1
F	86	0,1
M	99	0,1
Total	**97 138**	**100,0**

Se puede observar que realmente esta variable muestra los valores asignados en el diccionario de datos, sin embargo, existen algunas inconsistencias como:

- 9 casos con categoría vacío
- 86 casos con la letra “F”
- 99 casos con la letra “M”

Nombre de la variable: DA33CONYU

Tipo: Cadena
Etiqueta: Estado conyugal/marital
Valores o categorías:

1. Conviviente
2. Casado/a
3. Viudo/a
4. Divorciado/a
5. Separado/a
6. Soltero/a
9. Ignorado

Obteniendo la frecuencia con SPPS

DA33CONYU Estado conyugal/marital

	Frecuencia	Porcentaje
	7 583	7,8
0	2	0,0
1 Conviviente	6 699	6,9
2 Casado	35 054	36,1
3 Viudo	12 959	13,3
4 Divorciado	909	0,9
5 Separado	1 076	1,1
6 Soltero	24 364	25,1
9 Ignorado	8 492	8,7
Total	**97 138**	**100,0**

Si bien la variable DA33CONYU presenta todas las categorías del diccionario, existen 7 mil 583 casos con categoría “vacío” y dos casos con categoría “0”.

Nombre de la variable: DA34CODNIV

Tipo: Cadena
Etiqueta: Nivel de instrucción del fallecido
Valores o categorías:

0. Ningún nivel/Iletrado
1. Inicial/Pre-escolar
2. Primaria Incompleta
3. Primaria Completa
4. Secundaria Incompleta
5. Secundaria Completa

6. Superior no universitaria incompleta
7. Superior universitaria completa
8. Superior universitaria incompleta
9. Superior universitaria completa
99. Ignorado

DA34CODNIV Nivel de instrucción del fallecido

	Frecuencia	Porcentaje
	1 580	1,6
0	1	,0
2	1	,0
3	6	,0
4	2	,0
5	3	,0
0 Ningun nivel / Iletrado	811	,8
00	10 617	10,9
01	381	,4
02	4 604	4,7
03	6 281	6,5
04	1 947	2,0
05	5 251	5,4
06	360	,4
07	1 083	1,1
08	494	,5
09	1 714	1,8
1 Inicial / Pre-escolar	1 542	1,6
10	1	,0
2 Primaria Incompleta	11 018	11,3
3 Primaria Completa	9 730	10,0
33	1	,0
39	1	,0
4 Secundaria Incompleta	2 692	2,8
5 Secundaria Completa	8 519	8,8
51	1	,0
6 Superior no universitaria incompleta	647	,7
62	2	,0
7 Superior no universitaria completa	1 191	1,2
8 Superior universitaria incomple	1 218	1,3
9 Superior universitaria completa	2 456	2,5
99 ignorado	22 983	23,7
Total	**97 138**	**100,0**

Si

Aparte de las categorías validas según cuestionario, se observan otros valores, es decir, que el sistema de ingreso de datos permite ingresar otros valores extraños o no se definió las reglas de entrada de datos adecuados.

Las inconsistencias presentadas y que se deben corregir son, por ejemplo:

- En vez de “0” se digitó “00” y no se escribió la categoría “Ningún nivel/Iletrado; de la misma manera, en vez de “1” se digitó “01” y tampoco se apuntó la categoría “Inicial/Pre-escolar”.
- Asimismo, se aprecian 1 mil 580 casos “vacíos”, dos casos con categoría “62”, un caso con categorías “10”, “33”, “39” y “51”.

Por lo tanto, se puede decir que un análisis de frecuencias de cada variable es importante para poder encontrar inconsistencias (que impide realizar un buen análisis), al menos que estos errores sean corregidos.

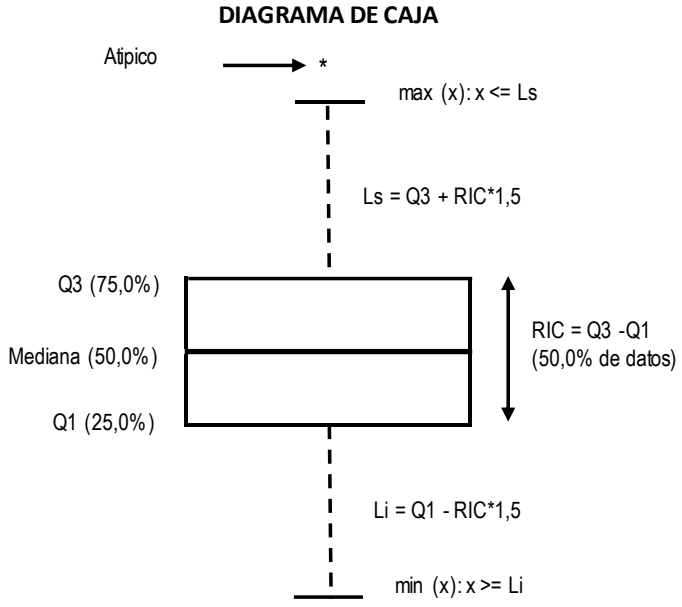
De no corregirse estas inconsistencias, al momento de efectuar cruces de variables podemos obtener resultados incorrectos o perder casos.

Diagramas de caja

Es una presentación visual que describe varias características importantes, al mismo tiempo, tales como la dispersión y simetría. Para su elaboración se representan los tres cuartiles y los valores máximos y mínimos de los datos, sobre un rectángulo el cual está alineado horizontal o verticalmente y en donde los lados más largos muestran el recorrido intercuartilico.

Este rectángulo está dividido por un segmento vertical que indica donde se posiciona la mediana y por lo tanto su relación con los cuartiles primero y tercero (el segundo cuartil coincide con la mediana).

La caja se ubica a escala sobre un segmento que tiene como extremo los valores mínimo y máximo de la variable. Las líneas que sobresalen de la caja se llaman bigotes. Estos bigotes tienen un límite de prolongación, de modo que cualquier dato o caso que no se encuentre dentro de este rango es marcado e identificado individualmente.



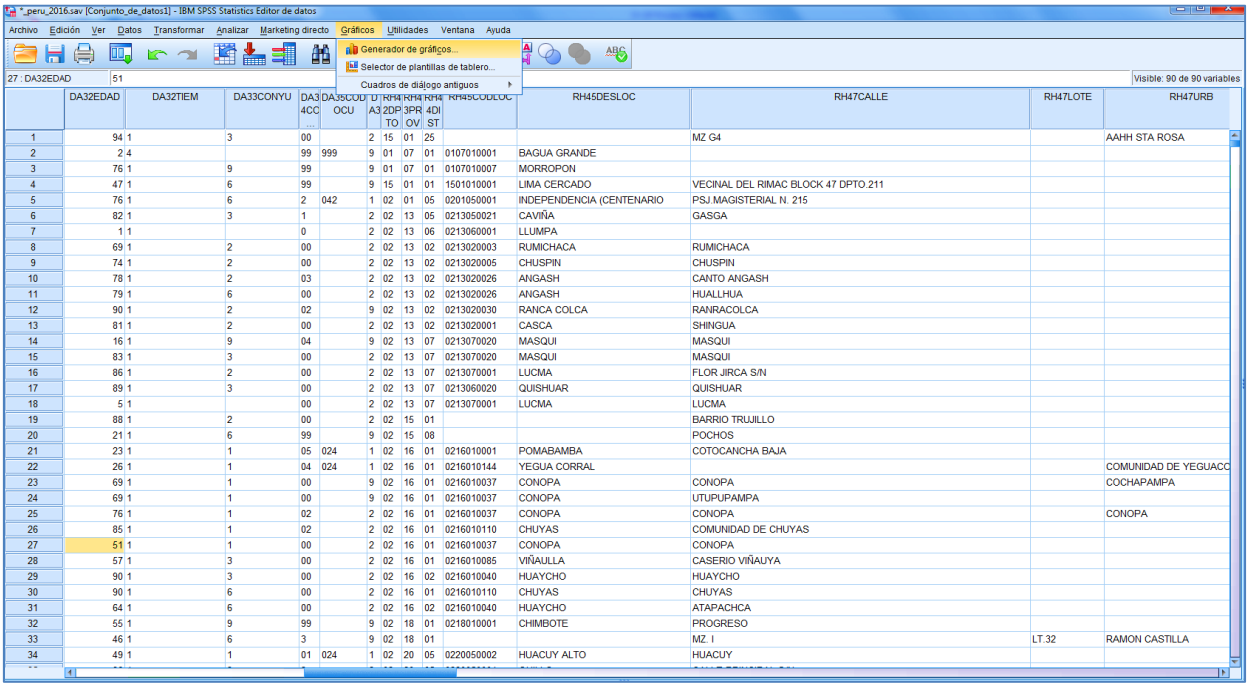
Dónde:

- ✓ Mediana, valor que deja a la mitad de los casos por encima y a la otra mitad por debajo.
- ✓ Primer Cuartil (Q1), indica que el 25,0% de los casos se encuentra por debajo de este valor.
- ✓ Tercer Cuartil (Q3), indica que el 75,0% de los casos se encuentra por encima de este valor.
- ✓ Rango Intercuartilico (RIC), es la diferencia entre el tercer y el primer cuartil.

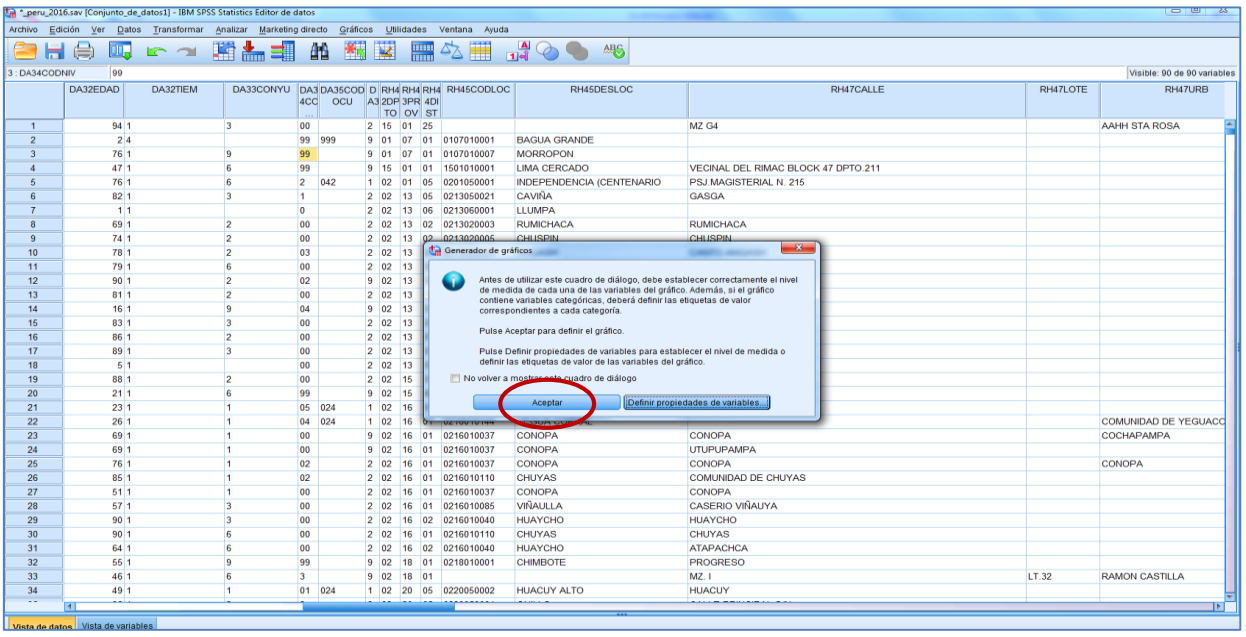
- ✓ Limite Superior o Inferior (Ls o Li), contiene los casos que se encuentran por encima de Q3 + 1,5*RIC o Li por debajo de Q1 – 1,5*RIC.
- ✓ Valores atípicos, son aquellos que están más allá de los limites inferior o superior. Cuando estos valores están más allá de 3 veces el RIC en vez del 1,5 son denominados valores extremos.

Ejemplo:

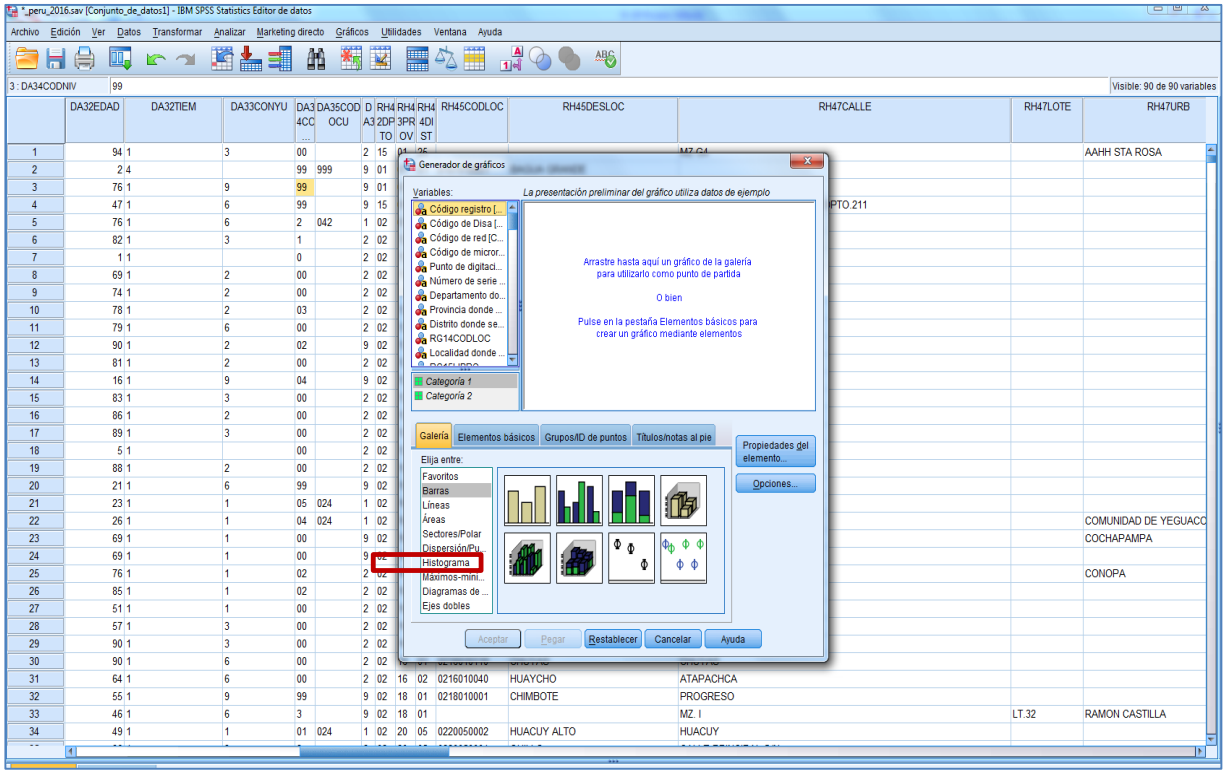
Tomando como referencia la base de datos de defunciones, se ha obtenido un diagrama de caja y bigotes para la variable EDAD DEL FALLECIDO (**DA32EDAD)** con el programa SPSS.
Para ello vamos a:
Gráficos
Generador de gráficos



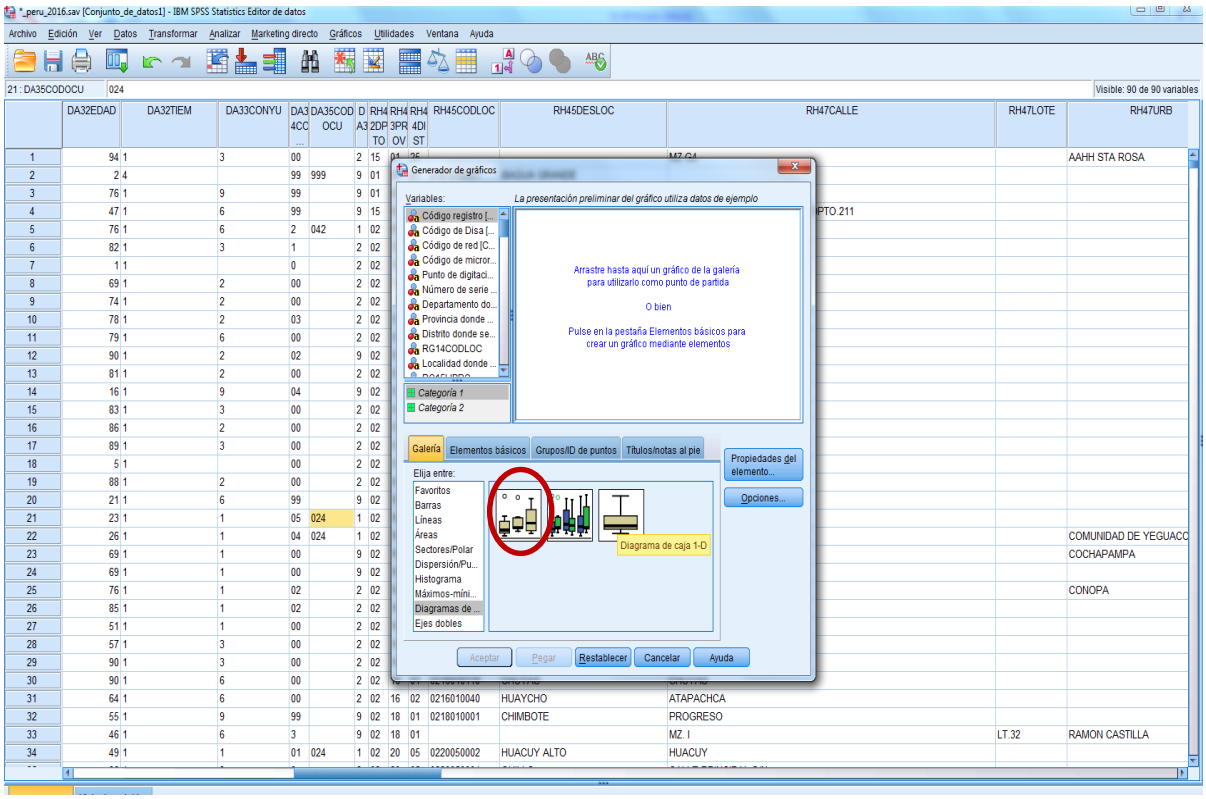
Y dar aceptar



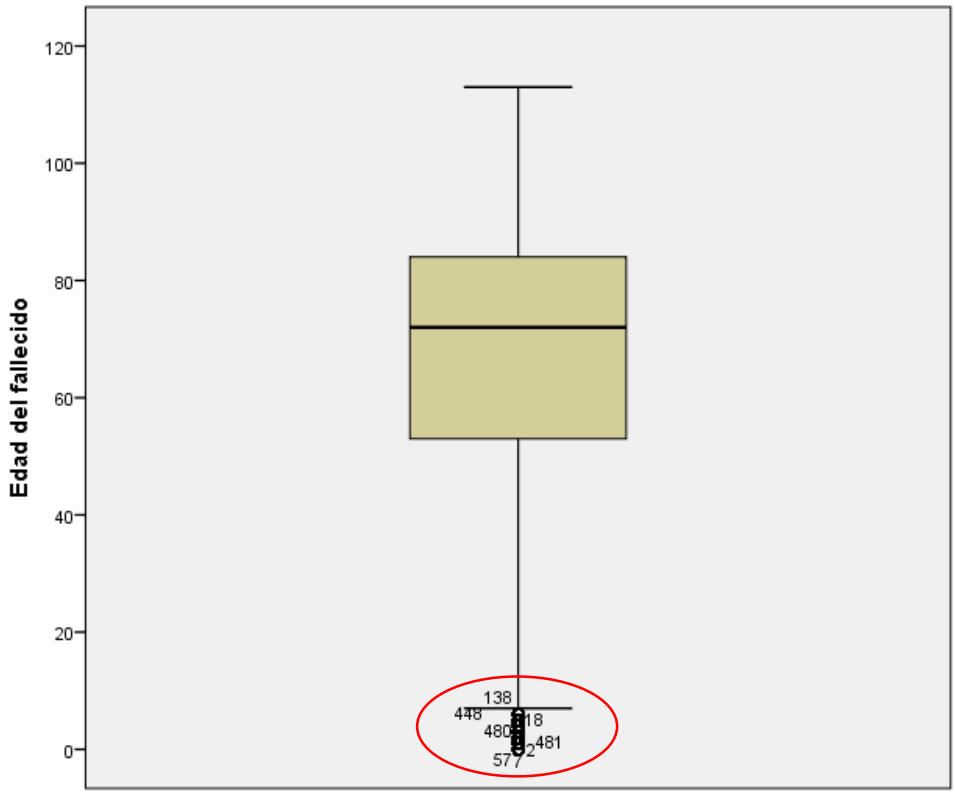
Luego aparece la siguiente ventana



Donde se elige diagramas de caja



Y se obtiene el siguiente gráfico



Como podemos observar, un conjunto de datos discordantes en la parte inferior de la caja, lo cual debe ser revisado para determinar si son errores o validar el dato.

Revisión de cruce de variables

El cruce de variables permite identificar si existe relación entre dos o más de ellas. En este caso se ha considerado las variables edad en años (edfall_anio) y nivel educativo (da34codniv) de la base de defunciones.

Previamente antes de realizar el cruce de dichas variables, se realizó un análisis de frecuencia.

EDAD EN AÑOS (edfall_anio)

edfall_anio

Edad	Frecuencia	Porcentaje
0	3 638	3,8
1	540	0,6
2	307	0,3
3	194	0,2
4	172	0,2
5	163	0,2
.		
.		
.		
10	95	0,1
11	120	0,1
12	127	0,1
13	124	0,1
14	149	0,2
15	173	0,2
.		
.		
.		
30	420	0,4
31	392	0,4
32	375	0,4
33	420	0,4
.		
.		
.		
.		
50	799	0,8
51	711	0,7
52	791	0,8
53	853	0,9
.		
.		
.		
.		
80	2 413	2,5
81	2 406	2,5
82	2 471	2,6
83	2 436	2,5
.		
.		
.		
.		
100	220	0,2
101	157	0,2
102	132	0,1
.		
.		
.		
110	7	0,0
111	6	0,0
112	1	0,0
113	1	0,0
Total	**96 836**	**100,0**

NIVEL DE EDUCACIÓN (da34codniv)

da34codniv

	Frecuencia	Porcentaje
0 Ningun nivel / Iletrado	11 427	11,8
1 Inicial / Pre-escolar	1 920	2,0
2 Primaria Incompleta	15 619	16,1
3 Primaria completa	16 014	16,5
4 Secundaria Incompleta	4 641	4,8
5 Secundaria completa	13 767	14,2
6 Superior No Universitaria	1 007	1,0
7 Superior No Universitario	2 272	2,3
8 Superior No Universitario	1 712	1,8
9 Superior Universitario	4 169	4,3
10	1	,0
33	1	,0
39	1	,0
51	1	,0
62	2	,0
99 Ignorado	22 714	23,5
Total	95 268	98,4
Perdidos	1 568	1,6
TOTAL	**96 836**	**100,0**

Estas frecuencias nos muestran que la edad en años tiene como valor mínimo “0” años y 113 años de valor máximo. Además, existen 3 mil 638 personas fallecidas con “0” años de edad.

En el caso de la variable nivel de educación existen 1 mil 568 valores perdidos, dos casos con categoría “62” y un caso con categoría “10”, “33”, “39” y “51”, respectivamente.

Luego al cruzar estas variables se obtiene:

	da34codniv															
edfall_anio	0 Ningun nivel / Iletrado	1 Inicial / Pre-escolar	2 Primaria Incompleta	3 Primaria completa	4 Secundaria Incompleta	5 Secundaria completa	6 Superior No Universitaria Incompleta	7 Superior No Universitario Completo	8 Superior No Universitario Incompleto	9 Superior Universitario Completo	10	33	39	51	62	99 Ignorado
0	9	2	0	0	0	1	0	0	0	0	0	0	0	0	0	3 565
1	257	33	16	0	1	1	7	0	0	7	0	0	0	0	0	216
2	148	34	6	1	2	1	10	0	1	5	0	0	0	0	0	99
3	84	38	8	1	0	3	4	0	0	5	0	0	0	0	0	50
4	55	44	14	2	0	0	5	0	0	4	0	0	0	0	0	48
5	41	58	16	0	0	0	2	0	0	3	0	0	0	0	0	39
6	23	36	30	2	2	0	3	0	0	1	0	0	0	0	0	25
7	25	17	47	2	1	1	4	0	0	1	0	0	0	0	0	38
8	13	7	63	11	1	1	6	0	1	1	0	0	0	0	0	31
9	16	9	74	4	2	0	0	0	0	1	0	0	0	0	1	18
10	13	6	44	4	1	0	4	0	0	1	0	0	0	0	0	20
11	13	6	46	17	7	0	3	0	0	1	0	0	0	0	0	24
12	9	1	33	12	33	5	5	0	0	0	0	0	0	0	0	28
13	12	0	30	11	49	3	1	0	0	0	0	0	0	0	0	17
14	9	0	26	11	74	5	1	0	0	1	0	0	0	0	0	21
15	10	5	10	8	81	13	0	0	1	1	0	0	0	0	0	41
16	9	6	12	14	76	34	1	0	4	3	0	0	0	0	0	36
17	14	2	21	15	71	58	9	1	9	3	0	0	0	0	0	39
18	10	1	25	19	45	57	11	2	19	2	0	0	0	0	0	55
19	16	2	25	20	47	77	24	7	23	4	0	0	0	0	0	59
20	19	2	31	29	51	81	15	15	28	8	0	0	0	0	0	68
21	10	5	27	19	45	85	13	18	11	8	0	0	0	0	0	67
22	12	4	27	28	38	95	19	17	32	10	0	0	0	0	0	59
23	14	5	24	30	41	103	10	14	17	14	0	0	0	0	0	57
24	12	2	18	31	33	92	12	22	20	9	0	0	0	0	0	73
25	10	2	34	31	41	108	13	22	21	13	0	0	0	0	0	77
26	9	4	37	32	39	98	6	21	12	14	0	0	0	0	0	56
27	2	1	28	35	32	103	10	17	14	16	0	0	0	0	0	78
28	11	2	45	35	43	111	9	18	18	22	0	0	0	0	0	80
29	13	3	36	38	44	111	7	25	16	10	0	0	0	0	0	78
30	11	2	46	36	38	112	12	18	22	27	0	0	0	0	0	88
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
105	11	1	6	8	1	3	0	0	0	1	0	0	0	0	0	9
106	11	0	6	3	1	1	0	0	0	0	0	0	0	0	0	5
107	7	0	0	2	0	1	0	0	0	0	0	0	0	0	0	1
108	3	0	1	4	0	0	0	0	0	0	0	0	0	0	0	0
109	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
110	4	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1
111	1	0	1	2	0	1	0	0	0	0	0	0	0	0	0	1
112	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
113	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Donde se observan algunas inconsistencias:

- ✓ Que un niño fallecido de “0” años de edad tenga secundaria completa.
- ✓ Que siete niños fallecidos de “1” año de edad tengan educación superior no universitaria incompleta o superior universitaria completa.

✓ Que 10 niños de “2” años de edad tengan educación superior no universitaria incompleta.

Para evitar que se presenten este tipo de inconsistencias se debe:

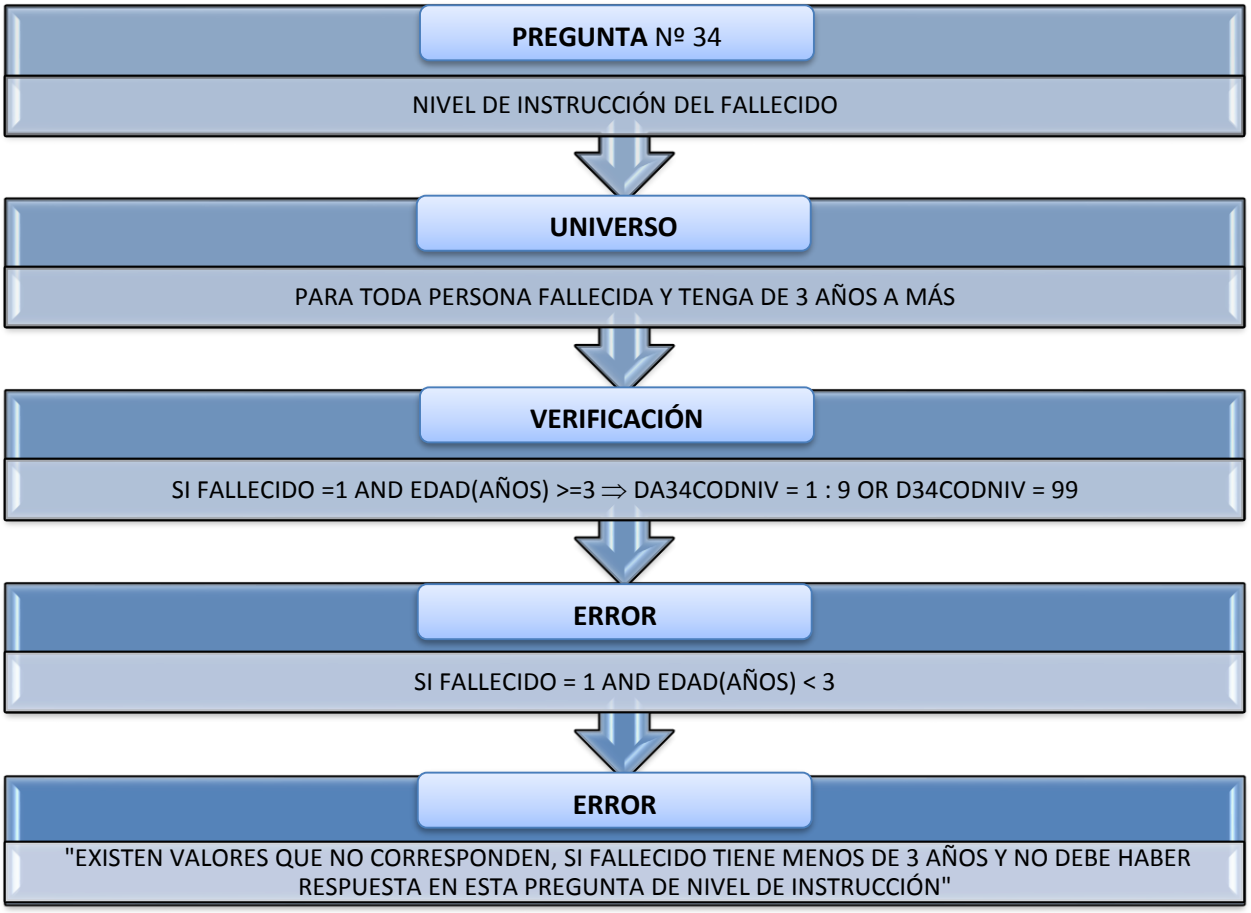
- Determinar las omisiones.
- Establecer reglas de consistencia con la finalidad de poder determinar valores vacíos, fuera de rango, etc.

Reglas de consistencia

✓ **PARA LA VARIABLE DA32EDAD**

PREGUNTA Nº 32	• EDAD DEL FALLECIDO
UNIVERSO:	• PARA TODA PERSONA FALLECIDA
VERIFICACIÓN:	• SI FALLECIDO =1 ⇒ DA32EDAD = 1 : 113
NOTA:	• CONVERTIR VARIABLES A SOLO AÑOS Y TENER OTRA VARIABLE EN MESES PARA LOS MENORES DE 1 AÑO
ERROR:	• "EXISTEN VALORES QUE NO CORRESPONDEN"

✓ **PARA LA VARIABLE DA34CODNIV**



Caso práctico de consistencia de una base de datos

De la base de datos de defunciones se ha considerado las variables tipo de documento de identidad (ID25TIPDOC) y número de documento de identidad (ID25NUMDOC).

Paso 1: realizar análisis de frecuencia de ambas variables tal como se muestra en la base de datos original.

✓ **Frecuencia de la variable Tipo de documento de identidad**

ID25TIPDOC Tipo de documento de identidad

	Frecuencia	Porcentaje
	5 640	5,8
0	3	0,0
1 DNI	88 954	91,6
2 Libreta militar	167	0,2
3 Carné FF.AA/PNP	19	0,0
4 Pasaporte	147	0,2
5 Carné de extranjería	141	0,1
6 Partida de nacimiento	760	0,8
7 Otro	1 307	1,3
Total	**97 138**	**100,0**

En el diccionario de datos las categorías validas son:

1. DNI
2. Libreta militar
3. Carné FF. AA/PNP
4. Pasaporte
5. Carné de extranjería
6. Partida de nacimiento
7. Otro
9. Ignorado

En la tabla de frecuencia existen 5 mil 640 casos con categoría vacíos y 3 casos con categoría “0” y no aparece la categoría “Ignorado”. Si unimos los casos de las categorías vacío y “0” como “Ignorado” se tiene:

ID25TIPDOC

	Frecuencia	Porcentaje
1 DNI	88 954	91,6
2 Libreta militar	167	0,2
3 Carné FF.AA/PNP	19	0,0
4 Pasaporte	147	0,2
5 Carné de extranjería	141	0,1
6 Partida de nacimiento	760	0,8
7 Otro	1 307	1,3
9 Ignorado	5 643	5,8
Total	**97 138**	**100,0**

✓ Frecuencia de la variable Número de documento de identidad (ID25NUMDOC)

ID25NUMDOC Número de documento de identidad

	Frecuencia	Porcentaje
	5 699	69,6
--	1	0,0
25515206	1	0,0
.-----	1	0,0
..	1	0,0
0	2	0,0
000	1	0,0
000000000	1	0,0
00000000324	1	0,0
00000024	1	0,0
0001814962	1	0,0
000263	1	0,0
000266933	1	0,0
00029520-16	1	0,0
000398146-	1	0,0
0009063	1	0,0
00098514-1	1	0,0
001	1	0,0
0010	1	0,0
00103529-1	1	0,0
0010504616	1	0,0
001087097	1	0,0
0011	1	0,0
0011185795	1	0,0
001182128	1	0,0
00120015-16	1	0,0
00173471	1	0,0
002330	1	0,0
0025490956	1	0,0
00259451	1	0,0
002-63	1	0,0
00265898	1	0,0
0051144	1	0,0
00541200 - 16	1	0,0
005589	1	0,0
0057283716	1	0,0
01248	1	0,0
01-38	1	0,0
017358	1	0,0
024	1	0,0
.	.	.
.	.	.
.	.	.
02408318-2	1	0,0
0421806	1	0,0
07295253	1	0,0
075102068-6	1	0,0
09 folio 03	1	0,0
0917	1	0,0
0974748	1	0,0
1101637583	1	0,0
110677973	1	0,0
110783202546	1	0,0
11cy46507	1	0,0
120E2764460	1	0,0
12CV92471	2	0,0
1305488015	1	0,0

Continúa..

Conclusión.

	Frecuencia	Porcentaje
131217	1	0,0
133-1938	1	0,0
13AR18582	1	0,0
140-E-27732446	1	0,0
14306097204	1	0,0
152293	1	0,0
2324060769	1	0,0
2657-2016	1	0,0
30024490940	1	0,0
303002443383	1	0,0
3408	1	0,0
3411005	1	0,0
344- 1928	1	0,0
375	1	0,0
38+05047	1	0,0
46744309-2	1	0,0
5362	1	0,0
5557247-A	1	0,0
6002627000	1	0,0
637 -2016-MP-FP	1	0,0
6412035281	1	0,0
865416	1	0,0
89520832	1	0,0
964	1	0,0
96885741	1	0,0
97071420446	1	0,0
9719408	1	0,0
979773888	1	0,0
ACTA DE NACIMIE	1	0,0
AT070774	1	0,0
b- 000916	1	0,0
CH1HZ6H1P	1	0,0
E-27589188	1	0,0
FB581762	1	0,0
HK3613361	1	0,0
j494996	1	0,0
LIBRO:111 FOLIO	1	0,0
M74909763	1	0,0
N1567980	1	0,0
NO TIENE	1	0,0
.	.	.
.	.	.
.	.	.
PAA081816	1	0,0
PART.BAUT.965	1	0,0
PARTIDA DE BAUT	1	0,0
S	1	0,0
S/N	19	0,2
SIN DNI	2	0,0
sin documentos	1	0,0
Total	8184	100,0
X	4	0,0
X006956	1	0,0
XX	92	1,1
xxx	1	0,0
XXXX	1	0,0
YA7120226	1	0,0

En la variable número de documento de identidad se encontraron los siguientes caracteres no numéricos, lo cual implica que debemos limpiar dichas variables y eliminar esos caracteres.

-	*
.	'
/	Ý
+	Ç
}	|
e	

Paso 2: analizar el tipo de documento de identidad con la longitud de caracteres de la variable número de DNI (solo tiene 8 dígitos).

Previamente a ello utilizamos la siguiente sintaxis para encontrar la longitud de caracteres del Documento Nacional de Identidad.

Sintaxis:

Si Tipo de Documento es vacío y Longitud de cadena del Documento de Identidad=8, debemos recodificar el Tipo de Documento de Identidad con 1 (DNI).

```
IF (ID25TIPDOC=9 and length(rtrim(ID25NUMDOC_new)) = 8 and
ID25NUMDOC_new~="99999999") ID25TIPDOC=1.
execute.
```

Cuando se tiene DNI y el número de caracteres del Número de DNI es menor de 8 caracteres. compute larg_ini=length(rtrim(ID25NUMDOC)), se obtendría el siguiente cuadro.

ID25TIPDOC Tipo de documento de identidad	Total	larg_fin LONGITUD DE CARACTERES DE LA VARIABLE NUMERO DE DNI													
		2,00	3,00	4,00	5,00	6,00	7,00	8,00	9,00	10,00	11,00	12,00	13,00	14,00	15,00
	5 640	0	0	1	0	1	4	5 597	7	30	0	0	0	0	0
0	**3**	0	0	0	0	0	1	2	0	0	0	0	0	0	0
1 DNI	**88 954**	0	0	0	0	0	0	87 866	885	168	23	1	0	1	10
2 Libreta militar	**167**	3	0	0	1	12	29	107	6	8	1	0	0	0	0
3 Carné FF.AA/PNP	**19**	0	0	1	1	3	8	5	0	0	0	0	0	1	0
4 Pasaporte	**147**	0	0	1	1	2	31	51	58	2	1	0	0	0	0
5 Carné de extranjería	**141**	0	0	0	5	7	3	12	109	3	1	1	0	0	0
6 Partida de nacimiento	**760**	8	66	25	2	16	8	233	23	368	9	0	1	1	0
7 Otro	**1 307**	1	20	11	4	19	34	896	18	285	12	1	4	2	0
9 Ignorado	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	**97 138**	**12**	**86**	**39**	**14**	**60**	**118**	**94 769**	**1 106**	**864**	**47**	**3**	**5**	**5**	**10**

Se puede observar que todos los tipos de documentos de identidad tienen entre 2 a 15 caracteres. Uniendo en la categoría "Ignorado" los casos vacíos y "0"se tiene

ID25TIPDOC Tipo de documento de identidad	Total	larg_fin LONGITUD DE CARACTERES DE LA VARIABLE NUMERO DE DNI													
		2,00	3,00	4,00	5,00	6,00	7,00	8,00	9,00	10,00	11,00	12,00	13,00	14,00	15,00
1 DNI	**88 954**	0	0	0	0	0	0	87 866	885	168	23	1	0	1	10
2 Libreta militar	**167**	3	0	0	1	12	29	107	6	8	1	0	0	0	0
3 Carné FF.AA/PNP	**19**	0	0	1	1	3	8	5	0	0	0	0	0	1	0
4 Pasaporte	**147**	0	0	1	1	2	31	51	58	2	1	0	0	0	0
5 Carné de extranjería	**141**	0	0	0	5	7	3	12	109	3	1	1	0	0	0
6 Partida de nacimiento	**760**	8	66	25	2	16	8	233	23	368	9	0	1	1	0
7 Otro	**1 307**	1	20	11	4	19	34	896	18	285	12	1	4	2	0
9 Ignorado	**5 643**	0	0	1	0	1	5	5 599	7	30	0	0	0	0	0
Total	**97 138**	**12**	**86**	**39**	**14**	**60**	**118**	**94 769**	**1 106**	**864**	**47**	**3**	**5**	**5**	**10**

Paso 3: realizar un filtrado para los casos que tienen un carácter no numérico en el número de documento y considerando que solo interesa obtener información de los que tienen DNI

ID25TIPDOC	Total	larg_fin LONGITUD DE CARACTERES DE LA VARIABLE NUMERO DE DNI													
		2,00	3,00	4,00	5,00	6,00	7,00	8,00	9,00	10,00	11,00	12,00	13,00	14,00	15,00
Total	**90 754**	**12**	**86**	**39**	**14**	**60**	**118**	**88 385**	**1106**	**864**	**47**	**3**	**5**	**5**	**10**
1 DNI	**87 797**	0	0	0	0	0	0	**87 797**	0	0	0	0	0	0	0
2 Libreta militar	**160**	3	0	0	1	12	29	100	6	8	1	0	0	0	0
3 Carné FF.AA/PNP	**19**	0	0	1	1	3	8	5	0	0	0	0	0	1	0
4 Pasaporte	**146**	0	0	1	1	2	31	50	58	2	1	0	0	0	0
5 Carné de extranjería	**141**	0	0	0	5	7	3	12	109	3	1	1	0	0	0
6 Partida de nacimiento	**705**	8	66	25	2	16	8	178	23	368	9	0	1	1	0
7 Otro	**1 786**	1	20	12	4	20	39	243	910	483	35	2	4	3	10

OBS: Se debe seguir realizando filtros con todos los caracteres encontrados

Paso 4: resumiendo, los casos de personas fallecidas que tienen como documento de identidad DNI con los 8 dígitos requeridos.

RESUMEN

	Absoluto	%
Total	**97 138**	**100,0**
Con DNI	87 797	90,4
Otros	2 957	3,0
Sin DNI	6 384	6,6
Tienen nombres y apellidos	4 954	5,1
No tienen nombres ni apellidos	647	0,7
Al menos un apellido o nombre	783	0,8

Se encontraron 6 mil 384 casos sin DNI, de los cuales 4 mil 954 figuran con nombres y apellidos, 783 tienen al menos un apellido y nombre; mientras que, 647 no tienen nombres y apellidos.

III. DEFINICIONES BÁSICAS

III. DEFINICIONES BÁSICAS

1. **Atributo de una entidad,** característica o rasgos de interés de la entidad que la identifica.

2. **Base de datos,** es una colección de información organizada, ordenada de forma que un ordenador pueda seleccionar los datos. Es un sistema de archivos electrónico, se organizan por campos, registros y archivos.

3. **Calidad estadística,** es el cumplimiento de las propiedades que debe tener el proceso y el producto estadístico, para satisfacer las necesidades de información de los usuarios.

4. **Campo,** es la unidad básica de entrada de datos de un registro.

5. **Captura de datos,** procedimiento para transformar la información del cuestionario en un archivo electrónico de datos.

6. **Categoría,** conjunto objeto de cuantificación y caracterización.

7. **Censo,** es un conteo y recuento de la población de un determinado país, en un momento dado. Se realizan generalmente cada 10 años. Este estudio demográfico, arroja datos importantes para constatar la cantidad de personas por región que hay y que necesidades o características específicas tienen las viviendas en las que habitan.

8. **Clase,** es un conjunto de objetos que comparten una estructura y un comportamiento común. Una clase contiene la especificación de los datos que describen un objeto.

9. **Clasificación,** ordenamiento de todas las modalidades nominales o intervalos numéricos admitidos por una variable.

10. **Cobertura geográfica,** territorio al que se refiere la captación de datos en un proyecto estadístico.

11. **Codificación,** procedimiento para asignar identificadores numéricos y/o alfanuméricos a conceptos en un orden establecido.

12. **Claves o identificadores,** son variables claves que permiten identificar de manera unívoca cada objeto, elemento o registro de una tabla o entidad. Por este motivo se debe tener en cuenta que su control interno a efectos del sistema de información propiamente dicho, deba ser lo más sencillo posible.

13. **Código Nacional de Buenas Prácticas para las Estadísticas Oficiales,** instrumento técnico y regulador, aprobado mediante Decreto Supremo Nº 072-2012-PCM, representa un conjunto de principios y buenas prácticas que tienen como finalidad mejorar la calidad de las estadísticas oficiales y fortalecer la credibilidad y confianza de los usuarios.

14. **Control de calidad en el llenado de cuestionarios,** es la medición y análisis de los resultados obtenidos en los cuestionarios aplicados, con base en los criterios de validación establecidos para el trabajo de campo.

15. **Criterios de validación,** conjunto de reglas de naturaleza conceptual, que sirven de base para la identificación y solución de los problemas que se presentan en los datos estadísticos.

16. **Cruce de variables,** combinación de cada una de las clases de una clasificación con cada una de las clases de otra, respecto a las variables involucradas.

17. **Cuestionario,** instrumento de captación, que presenta, bajo un orden determinado, las preguntas e indicaciones necesarias para el registro de los datos correspondientes a las unidades de observación, en un proyecto de generación de estadística básica.

18. **Cuestionario electrónico,** formato que se presenta por medio de programas en equipos informáticos.

19. **Dato estadístico,** valor cuantitativo de un conjunto específico respecto a una variable, con referencia de tiempo y de espacio.

20. **Desagregación geográfica,** nivel de detalle de una división territorial.

21. **Diccionario de datos o variables,** es la metainformación o información sobre los datos almacenados, en la que se describe cada objeto en términos de su definición, sus variables y dominios de valores, allí se detallan los esquemas de la Base de datos y sus respectivas correspondencias.

22. **Diseño conceptual,** etapa de estudio, análisis y diseño de una base de datos que obtiene una estructura de la información de la futura BD independiente de la tecnología que se quiere utilizar.

23. **Diseño físico,** etapa del diseño de una base de datos que transforma la estructura obtenida en la etapa del diseño lógico con el objetivo de conseguir una mayor eficiencia y que, además, la completa con aspectos de implementación física que dependerán del SGBD que se debe utilizar.

24. **Diseño lógico,** etapa del diseño de una base de datos que parte del resultado del diseño conceptual y lo transforma, de modo que se adapte al modelo del SGBD con el que se desea implementar la base de datos.

25. **Diseño de instrumentos de captación**, serie de actividades para operacionalizar el marco conceptual, donde se adecúan los conceptos para fines de captación del dato, en un contexto específico y bajo determinadas características del ámbito geográfico, perfil del informante, perfil del entrevistador y de los procedimientos de captación.

26. **Encuesta por muestreo,** método para generar información estadística mediante la captación de datos para un subconjunto de unidades seleccionadas de la población objeto de estudio.

27. **Entidad,** objeto del mundo real que podemos distinguir del resto de los objetos y del cual nos interesan algunas propiedades.

28. **Entidad asociativa,** entidad resultante de considerar una interrelación entre entidades como una nueva entidad.

29. **Entidad débil,** entidad cuyos atributos no la identifican completamente, sino que sólo la identifican de forma parcial.

30. **Entidad obligatoria en una interrelación binaria,** entidad tal que una ocurrencia de la otra entidad que interviene en la interrelación sólo puede existir si se da como mínimo una ocurrencia de la entidad obligatoria a la que está asociada.

31. **Entidad opcional en una interrelación binaria,** entidad tal que una ocurrencia de la otra entidad que interviene en la interrelación puede existir aunque no haya ninguna ocurrencia de la entidad opcional a la que está asociada.

32. **Escala,** es el conjunto de datos que puede tomar una variable.

33. **Estadística básica,** información generada a partir de un conjunto de datos obtenidos de un proyecto censal, de una encuesta por muestreo o del aprovechamiento de registros administrativos.

34. **Estadística oficial,** estadísticas producidas y difundidas por las entidades integrantes del Sistema Estadístico Nacional que permiten conocer la situación económica, demográfica, ambiental y social a nivel nacional y territorial para la toma de decisiones.

35. **Estándar estadístico,** conjunto completo de directrices para las encuestas y fuentes administrativas recogiendo información sobre un tema en particular. El uso de estándares estadísticos permite repetir la recolección de estadísticas sobre una base constante. También permiten la integración de datos a lo largo del tiempo y entre diferentes fuentes de datos, lo que permite el uso de datos más allá del objetivo inmediato para el que se haya producido.

36. **Estándar estadístico internacional**, conjunto de guías estadísticas y recomendaciones internacionales que han sido desarrolladas por organizaciones internacionales con el trabajo de agencias nacionales. Los estándares cubren casi cada campo del ámbito estadístico; desde el diseño, la recolección de datos, el procesamiento y la difusión. Tales estándares también incluyen clasificaciones estadísticas internacionales (OCDE).

37. **Estrategia operativa**, conjunto integrado y ordenado de procedimientos para determinar la estructura operativa y plantilla de personal, el programa general de actividades y para la cobertura de las áreas seleccionadas, y la organización administrativa del proyecto para gestionar la estimación y adquisición de los requerimientos, flujo de materiales, elaboración de presupuesto y los controles para su eficiente aplicación.

38. **Un evento o suceso,** es un subconjunto de un espacio muestral, es decir, un conjunto de posibles resultados que se pueden dar en un experimento aleatorio.

39. **Fase de captación**, serie de actividades para obtener los datos a nivel de las unidades de observación, conforme a determinado método de generación de estadísticas.

40. **Fase de diseño conceptual**, serie de actividades para identificar las necesidades de información y determinar, el marco conceptual, los instrumentos de captación, los criterios de validación y la presentación de resultados.

41. **Fase de diseño de la captación y el procesamiento**

    Diseño de la captación, serie de actividades para determinar, desarrollar y probar las estrategias para el levantamiento de los datos, así como los procedimientos y sistemas para su seguimiento y control.

    Diseño del procesamiento, serie de actividades para determinar, desarrollar y probar estrategias y procedimientos que habrán de aplicarse para la validación de los datos captados y la generación de resultados estadísticos.

42. **Fase de planeación,** proceso para determinar los objetivos y estrategia de un proyecto, así como la secuencia de actividades y su calendarización, los recursos y la organización requeridos para su realización.

43. **Fase de presentación de resultados,** serie de actividades para la elaboración de productos a partir de la información estadística generada en un proyecto determinado.

44. **Fase de procesamiento,** serie de actividades mediante las cuales se ordenan, almacenan y preparan los archivos con la información captada, asegurando su congruencia a fin de proceder a su explotación para la presentación de resultados estadísticos.

45. **Fases del proceso de generación de estadística básica,** cada una de las series de actividades que se distinguen por su naturaleza técnica específica y los momentos de realización, dado un programa y calendario del proyecto estadístico.

46. **Fuente administrativa,** es la unidad de organización responsable de implementar una regulación administrativa (o grupo de regulaciones), cuyo registro correspondiente de unidades y transacciones se ven como fuente de datos estadísticos.

47. **Generalización /especialización**, construcción que permite reflejar que existe una entidad general, denominada entidad superclase, que se puede especializar en entidades subclase. La entidad superclase nos permite modelizar las características comunes de la entidad vista a un nivel genérico, y con las entidades subclase podemos modelizar las características propias de sus especializaciones.

48. **Inconsistencia,** incompatibilidad numérica o lógica en los valores de dos o más datos.

49. **Información estadística**. conjunto de datos estadísticos referentes a un objeto de conocimiento.

50. **Instrumento de captación,** formato en medio impreso o electrónico, diseñado para el registro de los datos que han de obtenerse de las unidades de observación, en un proyecto de generación de estadística básica.

51. **Macro actividad,** grupo de actividades propias de una fase del proceso de generación de estadística básica.

52. **Macro dato,** dato estadístico obtenido a partir de un conjunto de micro datos.

53. **Macro validación,** conjunto de actividades que se realizan para identificar comportamientos improbables de estructura o de valor en estadísticas generadas con base en un archivo de micro datos.

54. M**anual operativo,** documento con fines didácticos, y de apoyo durante el operativo, en donde se especifican, entre otros aspectos, las responsabilidades y actividades de las diferentes figuras operativas que participarán en el proyecto y cómo interactúan entre sí, se describe con el suficiente detalle los procedimientos que han de seguirse en el desarrollo de tales actividades.

55. **Marco conceptual,** esquema bajo el cual se presenta, en forma ordenada y con los vínculos correspondientes, el conjunto de conceptos referentes a temas, categorías, variables y clasificaciones con sus respectivas definiciones, aplicados en un proyecto de generación de estadísticas.

56. **Metadatos,** información necesaria para el uso e interpretación de las estadísticas. Los metadatos describen la conceptualización, calidad, generación, cálculo y características de un conjunto de datos estadísticos (DANE, 2012e).

57. **Micro datos,** son los datos sobre las características de las unidades de estudio de una población (individuos, hogares, establecimientos, entre otros), que constituyen una unidad de información en una base de datos y que son recogidos por medio de una operación estadística.

58. **Micro validación,** conjunto de actividades que se realizan para identificar las inconsistencias en la información a nivel registro, las cuales tienen como fundamento la confronta de los criterios de validación con cada uno de los registros del archivo de micro datos.

59. **Muestra de observaciones,** es el conjunto de valores que toma una variable estadística sobre una muestra de individuos; es decir, es un subconjunto de la población de observaciones. Universidad de Buenos Aires. Glosario de Conceptos de Estadística.

60. **Multirespuesta,** dos o más respuestas para preguntas de una sola opción.

61. **Omisión de respuesta,** ausencia de respuestas en preguntas donde debería haberla.

62. **Operación estadística**, conjunto de procesos y actividades que, partiendo de la recolección sistemática de datos, conduce a la producción de resultados agregados (DANE, 2012e).

63. **Personal operativo.** Personas contratadas para realizar actividades relativas al levantamiento de datos.

64. **Procesamiento de los datos,** es el conjunto de procesos para el tratamiento de datos, necesarios para producir información estadística; comprende la edición y la ponderación de los datos.

65. **Proceso de producción estadística,** se refiere la secuencia de pasos y procedimientos que se realizan para producir estadísticas; comprende la planeación, captación, procesamiento, análisis y difusión de la información, aplicando una metodología elaborada para tal fin.

66. **Registro administrativo,** son los datos individuales sobre un evento o hecho correspondiente a una unidad (Persona, establecimiento o entidad), el cual puede ocurrir en distintos momentos, está sujeto a regulación o control y es recogido sistemáticamente por alguna oficina o dependencia perteneciente a una entidad del sector público a fin de cumplir con su función para la que fue creada.

67. **Regulación administrativa,** son las formalidades administrativas con que los gobiernos recogen información e intervienen en decisiones económicas individuales.

68. **Respuestas a preguntas no aplicables,** registro de valores en preguntas que no corresponden a un determinado grupo.

69. **Servicio en línea,** acceso electrónico, vía Internet, que se brinda a los usuarios por parte de las unidades productoras e integradoras de información estadística, para que consulten las bases de datos (si es el caso), los productos, tabulados e indicadores, en el momento en que lo decidan los usuarios. Este servicio implica ofrecer asistencia al usuario en el momento de la consulta o diferida en el tiempo si así lo requiere el tipo de consulta.

70. **Sistema de Gestión de Base de Datos (SGBD),** en inglés Data Base Management System, es un tipo de software muy específico, dedicado a servir de interfaz entre la base de datos, el usuario y las aplicaciones que la utilizan. Se compone de un lenguaje de definición de datos, de un lenguaje de manipulación de datos y de un lenguaje de consulta.

71. **Sistema Estadístico Nacional (SEN),** conjunto articulado de componentes que, de manera organizada y sistemática, garantiza la producción y difusión de las estadísticas oficiales a nivel nacional que requiera el país. Tiene como objetivo asegurar que las actividades estadísticas que efectúan las entidades del Estado en los tres niveles de gobierno se desarrollen en forma integrada, coordinada, racionalizada y bajo una normatividad técnica común, contando para ello con autonomía técnica y de gestión. El ente rector de este sistema es el Instituto Nacional de Estadística e Informática (INEI).

72. **Software estadístico,** es un software especializado en el análisis y explotación de la información estadística, para lo cual existen una variedad de aplicaciones informáticas disponibles.

73. **Tablas,** la creación de tablas debe estar supeditada a la representación del dominio o el ámbito de aplicación de las mismas. Esto es en función de los procesos y actividades que lleva a cabo una institución, la información y documentación que produce, su tratamiento y los servicios que ofrece. Las tablas o entidades deben tener una denominación que identifique el contenido de la información

que almacenará, el nombre del proceso que se desempeña con la información almacenada o la finalidad en el tratamiento de la misma.

74. **Tema,** enunciado genérico referente a un campo de conocimiento.

75. **Unidad,** concepto primario relacionado con los componentes elementales de la muestra estadística. Sinónimo, pero no esencialmente idéntico, de caso, observación, registro o individuo.

76. **Unidad de análisis,** se refiere al elemento, el que o quien, es objeto de investigación, que debe ser la entidad mayor, primaria o representativa.

77. **Unidad de observación,** conocido también como unidad elemental. Es el elemento o unidad base de la población o de la muestra que permite obtener información o datos referidos a ciertas características o variables, que nos interesan para explicar un determinado fenómeno.

78. **Unidad de registro,** constituye el elemento básico de un sistema de registro. Su definición permite identificar las unidades de análisis para analizar las características registradas.

79. **Validación,** conjunto de actividades para identificar, en la información captada, los datos que cumplen con los requisitos de congruencia lógica y aritmética, completez e integridad, a fin de aplicar a los que no los cumplen, una solución bajo criterios específicos, que aseguren la eliminación de inconsistencias sin afectar los datos válidos originales.

80. **Valores atípicos,** se trata de observaciones cuyos valores se encuentran numéricamente distantes del resto de los datos, pueden generarse por errores de procedimiento, hechos extraordinarios o causas no conocidas. Tienden a distorsionar los resultados y el análisis, por ello deben ser identificados y tratados adecuadamente.

81. **Variable,** es una característica medible de elemento o un conjunto de la población o de la muestra, puede ser cualitativa y cuantitativa.

82. **Variable aleatoria,** conocida también como variable estocástica o probabilística. Es la característica considerada en un experimento aleatorio cuyo valor de ocurrencia se conocerá una vez observado.

83. **Variable bidimensional,** es aquella que proporciona información sobre dos características de la población (por ejemplo: edad y altura de los alumnos de una clase).

84. **Variable continúa**, es una variable cuantitativa. Es la característica de la población, cuyos valores están representados mediante el conjunto de los números reales. Puede tomar cualquier valor real dentro de un intervalo. Por ejemplo, la velocidad de un vehículo puede ser 80,3 km/h, 94,57 km/h.

85. **Variable cualitativa,** es aquella que representa cualidades, atributos o características no numéricas y estas pueden ser nominales y ordinales.

86. **Variable cuantitativa,** es aquella característica de la población o de la muestra que es posible representar numéricamente. Éstas pueden ser continua y discreta.

87. **Variable determinística,** es aquella cuyo valor puede ser predicho con exactitud.

88. **Variable discreta,** es una variable cuantitativa. Es la característica de la población, cuyos valores están representados mediante el conjunto de los números naturales. Por ejemplo, el número de alumnos de un aula.

89. **Variable nominal,** es una variable cualitativa la cual sólo permite asignar nombres a los datos y no implica ningún orden. Ej. el idioma de los habitantes de la tierra.

90. **Variable ordinal,** es una variable cualitativa cuyos valores solamente pueden ser ordenados con algún criterio

ANEXOS

- DICCIONARIO DE DATOS (ENAHO)

- DICCIONARIO DE DATOS - NACIMIENTOS (Registros administrativos)

ANEXOS

1. DICCIONARIO DE DATOS

Encuesta Nacional de Hogares 2017 – Encuesta Continua

INEI

1. ARCHIVOS DEL CUESTIONARIO ENAHO.01.

1.1. ENAHO01-2017-100.SAV: Características de la Vivienda y del Hogar (Módulo 100).

Archivo: ENAHO01-2017-100.SAV

NOMBRE VARIABLE	TAMAÑO	DECIMALES	FORMATO	ETIQUETA
AÑO	4	0	C	Año de la Encuesta
MES	2	0	C	Mes de procesamiento
CONGLOME	6	0	C	Número de conglomerado
VIVIENDA	3	0	C	Número de selección de vivienda
HOGAR	2	0	C	Número secuencial del hogar
UBIGEO	6	0	C	Ubicación geográfica
DOMINIO	1	0	N	Dominio Geográfico 1 Costa Norte 2 Costa Centro 3 Costa Sur 4 Sierra Norte 5 Sierra Centro 6 Sierra Sur 7 Selva 8 Lima Metropolitana Rango: 1 – 8
ESTRATO	1	0	N	Estrato Geográfico 1 De 500 000 a más habitantes. 2 De 100 000 a 499 999 habitantes. 3 De 50 000 a 99 999 habitantes. 4 De 20 000 a 49 999 habitantes. 5 De 2 000 a 19 999 habitantes. 6 De 500 a 1 999 habitantes. 7 Área de Empadronamiento Rural (AER) Compuesto 8 Área de Empadronamiento Rural (AER) Simple Rango: 1 – 8
PERIODO	1	0	N	Período de Ejecución de la Encuesta 1 Primer Período 2 Segundo Período 3 Tercer Período 4 Cuarto Período 5 Quinto Período Rango: 1- 5

NOMBRE VARIABLE	TAMAÑO	DECIMALES	FORMATO	ETIQUETA
TIPENC	1	0	N	Tipo de Selección del Conglomerado 1 Selección automática por computadora 3 Selección de la muestra panel 4 Selección por computadora en el Área Rural 5 Selección por conteo en el Área Rural Rango: 1, 3 – 5
FECENT	6	0	N	Fecha de Resultado final de la encuesta MMDDAA) RESULT
1	0	N		Resultado Final de la Encuesta 1 Completa

2. DICCIONARIO DE DATOS DE NACIMIENTOS

N°	Variable	Tipo	Descripción	Valores
1	Fuente	n(1)	Tipo de fuente procesada	1 – CNV 2 – HV
2	ID	n(6)	ID por fuente	
3	Nu_cnv	c(10)	Número de certificado de nacido vivo	
4	Cod_disa	c(2)	Código de DIRESA	98 – SIN DEFINIR (Usualmente EESS privados) 99 – IGNORADO (Usualmente EESS que no están registrados en RENAES. Fuente CNV)
5	Cod_red	c(2)	Código de Red	
6	Cod_mred	c(2)	Código de Microred	
7	Cod_estab	n(5)	Código RENAES	
8	Sexo	n(1)	Sexo del nacido vivo	1 – Masculino 2 – Femenino
9	Fechanac	date	Fecha del nacido vivo	
10	Ubi_nac	c(6)	Ubigeo del lugar de nacimiento	
11	Peso	n(4)	Peso del nacido vivo (gramos)	
12	S_ocurren	n(1)	Sitio de ocurrencia del parto	1 – Hospital o Clínica 2 – Centro de Salud 3 – Puesto de Salud 4 – Consultorio 5 – Domicilio 6 – Otro 9 – Ignorado
13	Atendiopart	n(2)	Profesional que atendió el parto	1 – Médico 2 – Obstetriz 3 – Enfermera(o) 4 – Interno(a) 5 – Técnico Salud 6 – Promotor Salud 7 – Partera/Comadrona 8 – Familiar 9 – Otro 10 – Nadie (autoayuda) 99 – Ignorado
N°	**Variable**	**Tipo**	**Descripción**	**Valores**
14	Tipopart	n(1)	Tipo de parto	1 – Unico 2 – Doble 3 – Triple 4 – Más de tres 9 – Ignorado
15	Condipart	n(1)	Condición del parto	1 – Espontáneo 2 – Instrumentado 3 – Cesárea 4 – Otro 5 – Ignorado

16	Durac_emb	n(2)	Duración del embarazo (semanas)	
17	Edad_mama	n(2)	Edad de madre	
18	Leer	n(1)	Condición de analfabetismo	1 – Si 2 – No 9 – Ignorado
19	N_educac	n(2)	Nivel de educación de la madre	0 – Ningún nivel / iletrado 1 – Inicial / pre-escolar 2 – Primaria incompleta 3 – Primaria completa 4 – Secundaria incompleta 5 – Secundaria completa 6 – Superior no universitaria incompleta 7 – Superior no universitaria completa 8 – Superior universitaria incompleta 9 – Superior universitaria completa 99 – Ignorado
20	Ubi_res_ma	c(6)	Residencia habitual de la madre	
21	Ubi_ln_ma	c(6)	Lugar de nacimiento de la madre	
22	Tip_doc	n(1)	Tipo de documento de identidad de la madre	1 – DNI / LE 2 – LM / BOL 3 – Carnet extranjería 4 – Acta nacimiento 6 – Pasaporte 7 – DI del extranjero 9 – Ignorado
23	Num_doc	c(20)	N° de documento de identidad de la madre	
24	Ap_paterno	c(20)	Apellido paterno de la madre	
N°	**Variable**	**Tipo**	**Descripción**	**Valores**
25	Ap_materno	c(20)	Apellido materno de la madre	
26	Nombres	c(20)	Nombres de la madre	
27	Est_civil	n(1)	Estado civil de la madre	1 – Conviviente 2 – Casada 3 – Divorciada 4 – Separada 5 – Soltera 6 – Viuda 7 – Ignorado

28	Ocupación	n(2)	Ocupación de la madre	1 1 - Profesionales, técnicos y trabajadores asimilados 2 - Funcionarios públicos, gerentes adm. empresas no agrícolas 3 - Personal administrativo y trabajadores asimilados 4 - Comerciantes, vendedores y personas en ocupaciones afines 5 - Trabajadores De Los Servicios 6 - Trabajadores Agrícolas, Forestales, Pescadores Y Cazadores 7 - Trabajadores No Agrícolas, Cond. De Maquinas Y Vehic. 8 - No Pea (Estudiantes, Amas De Casa) 9 - Fuerzas armadas y policiales 10 - Miembros del poder ejecutivo y de los cuerpos legislativos 11 - Profesionales, científicos e intelectuales 12 - Técnicos de nivel medio y trabajadores asimilados 13 - Jefes y empleados de oficina 14 - Trabajadores calificados de los servicios personales, protección, seguridad y vendedores del comercio y mercado 15 - Agricultores (explotadores); trabajadores calificados agropecuarios ,pesqueros 16 - Obreros, operadores de las actividades de minas, canteras, petróleo, industrias manufact y otros 17 - Obreros de la construcción, confeccionadores de productos de papel y cartón, trab.del caucho y plástico, de las artes gráficas, fabr. de instrument. de música, pintores, conductores de maq. y medios de transporte y otros afines 99 - No especificados - trabajadores no calificados de los servicios; peones agropecuarios, forestales de la pesca, de las minas y canteras, ind. manufactureras, construcción, peones de carga y vendedores ambulantes y otros afines
N°	**Variable**	**Tipo**	**Descripción**	**Valores**
29	Tot_embar	n(2)	Total de embarazos de la madre	
30	H_vivos	n(2)	N° hijos actualmente vivos (incluído el recién nacido)	
31	H_vivos_f	n(2)	N° hijos nacidos vivos que fallecieron	
32	H_aborto	n(2)	N° abortos y de nacidos muertos	
33	Etnia	n(2)	Etnia del recién nacido	

34	Finan	n(2)	Tipo financiador de salud	1 - Usuario 2 - Sis 3 - Essalud 4 - Soat 5 - Sanidad Fap 6 - Sanidad Naval 7 - Sanidad Ep 8 - Sanidad Pnp 9 - Privados 10 - Otros 11 – Exonerado 99 - Ignorado

REFERENCIAS BIBLIOGRÁFICAS

REFERENCIAS BIBLIOGRÁFICAS

1. Departamento Administrativo Nacional de Estadística (DANE). Metodología para el Fortalecimiento de Registros Administrativos, 2010.
2. Cazau, Pablo (2006). Fundamentos de Estadística, Buenos Aires, enero 2006.
3. Gonzáles Ramírez Byron (2004). Escala de Medición en Estadística, enero 2004.
4. Palomo Sánchez, José Gabriel (2011). Estadística Descriptiva, julio 2011.
5. López Roldán, Pedro y Fachelli, Sandra. Metodología de la Investigación Social Cuantitativa, 2015.
6. Departamento Administrativo Nacional de Estadística (DANE). Lineamientos para la Documentación de Metadatos a partir de los Estándares DDI y Dublin Core, 2014.
7. Departamento Administrativo Nacional de Estadística (DANE). Lineamientos Básicos para el Desarrollo de una Operación Estadística, 2013.
8. Federico Segui Stagno. Marco Conceptual y Metodológico que Sustenta el Diseño. Desarrollo e Implementación de un Sistema Integrado de Registros Estadísticos de Población e Inmuebles, 2016.
9. Instituto Nacional de Estadística y Geografía (INEGI). Operaciones de Base de Datos en Excel, 2008.
10. Instituto Nacional de Estadística Geografía e Informática (INEGI). Proceso Estándar para Encuestas por Muestreo, 2010.
11. Universidad de Buenos Aires. Glosario de Conceptos de Estadística, 2017
12. Departamento Administrativo Nacional de Estadística (DANE). Glosario de Términos de Calidad Estadística, 2018
13. Instituto Nacional de Estadística, Geografía e Informática (INEGI). Norma Técnica para la Generación de Estadística Básica, 2010.
14. Fondo Monetario Internacional (FMI). Glosario de Términos Estadísticos, 2017
15. Instituto Nacional de Estadística Geografía e Informática (INEGI). Presentación de Datos Estadísticos en Cuadros y Gráficas, 2011.
16. Comunidad Andina, Glosario de Términos Estadísticos, 2007.
17. Instituto Nacional de Estadística, Geografía e Informática (INEGI). Diseño Conceptual para la Generación de Estadística Básica, 2010.
18. Organización para la Cooperación y el Desarrollo Económicos (OCDE). Glosario de Términos Estadísticos, 2018.

GUÍA DE PROCEDIMIENTOS PARA EVALUAR BASE DE DATOS DE REGISTROS ADMINISTRATIVOS

Dirección Técnica de Demografía e Indicadores Sociales

Dirección y revisión
Nancy Hidalgo Calle
Dilcia Durand Carrión

Elaboración
Alex Mamani Córdova
Ana Naupari Rivas
Balvina Merino Saldaña
Elvis Manayay Guillermo
Gaby Quispe Huamani
Oscar Kuroiwa Quispe
Orlando Alarcón Medina
Verónica Hilario Campos

Apoyo en edición
Jenniffer Garboza Erazo
Victor Cabello Herencia

w w w. i n e i . g o b . p e

Visita el Portal del Estado Peruano: www.peru.gob.pe

"Produciendo estadísticas para el desarrollo del Perú"