TRATAMIENTO Y ANÁLISIS DE LA **INFORMACIÓN DE MERCADOS**

Andrej Nicolás Hillebrand

IDEASPROPIAS
editorial

Tratamiento y análisis de la información de mercados

Tratamiento y análisis de la información de mercados

Recogida e interpretación de datos para informes comerciales

Autor

Andrej Nicolás Hillebrand (Dachau [Alemania], 1983) es licenciado en Sociología por la Universidad de La Laguna. Además, ha realizado el curso de Técnico en Investigación Cuantitativa y Trabajo de Campo en el Ilustre Colegio Nacional de Doctores y Licenciados en Ciencias Políticas y Sociología, así como el curso de Investigación de Mercados en la Universidad Politécnica de Madrid.

Esta formación, junto a su experiencia en el desarrollo de formulaciones para la depuración masiva en base de datos y en el análisis, tabulación e interpretación de datos, se complementa con su labor profesional en servicios de encuestas de campo como grabador, supervisor y redactor de informes en proyectos de investigación.

Ficha de catalogación bibliográfica

Tratamiento y análisis de la información de mercados. Recogida e interpretación de datos para informes comerciales

1.ª edición
Ideaspropias Editorial, Vigo, 2015

ISBN: 978-84-9839-524-2
Formato: 17 x 24 cm • Páginas: 196

TRATAMIENTO Y ANÁLISIS DE LA INFORMACIÓN DE MERCADOS.
RECOGIDA E INTERPRETACIÓN DE DATOS PARA INFORMES COMERCIALES.

ISBN: 978-84-9839-524-2
Depósito legal: VG 210-2015
Autor: Andrej Nicolás Hillebrand
Impreso en España - Printed in Spain

ÍNDICE

INTRODUCCIÓN

En este manual formativo se tratará el proceso del tratamiento de datos y su interpretación y presentación en el contexto de la investigación de mercados. A lo largo de las tres unidades didácticas se profundizará en los diferentes mecanismos, informáticos y humanos, existentes para hacer frente a una tarea que exige agilidad, constancia y estudio permanente debido a la gran competitividad que existe en el sector.

Así, en la primera unidad didáctica, se estudiarán las principales técnicas relacionadas con la codificación y tabulación de los datos después de aplicar la correspondiente limpieza y edición de los mismos. Asimismo, se hará hincapié en las ventajas de algunas aplicaciones informáticas y su funcionamiento a favor de la reducción de errores. Se tendrán en cuenta los distintos tipos de preguntas en encuestas a la hora de codificar y tabular.

En la segunda unidad didáctica se abordarán los mecanismos estadísticos que se usan para el análisis de los datos obtenidos y su interpretación con el objetivo de extraer conclusiones que sean de utilidad para la empresa que lleva a cabo el estudio. Aunque se parte de métodos básicos, también se abordan otros tipos de análisis como el bivariante y el multivariante, así como los programas informáticos utilizados para estas tareas.

En la última unidad didáctica se resaltará la importancia de la realización de informes: su estructura y diseño, su planificación y elaboración, el establecimiento de un objetivo, etc. Además, también se tratarán las herramientas para realizar gráficos y tablas de apoyo y se establecerán las pautas organizativas y las actitudes necesarias para realizar presentaciones atendiendo a los resultados plasmados en el informe.

Con el estudio de este libro, el lector aprenderá a aplicar las técnicas de tratamiento y análisis de datos para construir un informe comercial, así como a realizar los diferentes tipos de encuestas y métodos de investigación de mercados al servicio de clientes y empresas.

Tratamiento y análisis de la información de mercados

1 Codificación y tabulación de datos e información de mercados

Objetivos

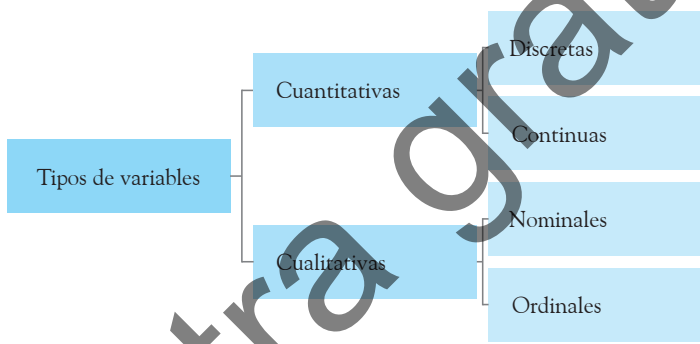
- Aplicar técnicas de codificación y tabulación de datos a la información recogida para la alimentación del SIM (Sistema de Información de mercados).
- Identificar y analizar las diferentes formas de representación de los datos obtenidos en la investigación comercial.
- Explicar las ventajas de la utilización de aplicaciones informáticas, hojas de cálculo y bases de datos en el tratamiento de un SIM empresarial.
- Utilizar herramientas como gráficos y tablas para facilitar la comprensión de los datos incluidos en los informes comerciales.

Contenidos

1.1. Objeto de la codificación y tabulación de datos

Cuando se lleva a cabo una investigación comercial, solamente una reducida parte de la información es, por sí misma, de carácter numérico o cuantitativo. Habitualmente, en las encuestas que se implementan para recabar la información que se necesita para una investigación comercial, existen muy pocas variables que dan un dato numérico con el que se puede operar desde un primer momento.

La **codificación** tiene por objeto operar con una variable cuando es posible llevar a cabo cálculos matemáticos a partir de los valores que la variable contiene y cuando es posible calcular métricas básicas como medias, medianas, etc.



Estas variables **cuantitativas** no requieren de una codificación, aunque se les pueda aplicar. Son variables discretas cuando sus valores únicamente pueden ser valores numéricos enteros, ya que un valor intermedio no tendría sentido; por ejemplo, la edad, el número de hijos, la cantidad de personas que viven en un hogar, etc. Se consideran variables continuas cuando admiten valores decimales o cuando hay continuidad entre un intervalo y otro, como en la altura, el peso, etc.

Es habitual que se creen agrupaciones de las variables cuantitativas para facilitar su análisis, lo que se consigue agrupando los valores mediante la construcción de intervalos. Estos intervalos siempre deben ser discriminatorios de modo que ningún valor pueda estar en dos intervalos a la vez, ya que esto generaría problemas a la hora de interpretar los datos.

Las variables **cualitativas,** en oposición a las cuantitativas, son aquellas que hacen referencia a una cualidad o característica. Las diferentes opciones de una variable cualitativa son sus características o categorías.

Hay preguntas cualitativas que únicamente disponen de dos posibles respuestas, estableciendo un resultado dicotómico. Estas preguntas se basan en respuestas de sí o no, como puede ser la variable de sexo que distingue entre hombre y mujer. También existen las variables politómicas, que son aquellas que cuentan con más de dos posibles respuestas.

Sin embargo, la mayoría de la información disponible en las variables cualitativas, es de carácter nominal u ordinal. Las variables nominales son aquellas que no tienen un criterio que permita determinar un orden y no es posible establecer una jerarquía. La mayoría de listados son variables nominales porque no tienen un orden inherente. Algunos ejemplos de este tipo de variables son la nacionalidad, el país de residencia, etc.

Cuando la información que se recoge en las variables se establece atendiendo a un orden, se consideran variables ordinales. Estas poseen una serie de valores que están establecidos en torno a una escala determinada previamente, no siendo necesario que presenten intervalos uniformes.

Ejemplo

Una pregunta típica que da lugar a una variable ordinal es: ¿En qué medida es probable que recomiende el servicio a un familiar o amigo?

(Puntúe del 1 al 5 según su respuesta).

- Nada probable (1).
- Poco probable (2).
- Algo probable (3).
- Bastante probable (4).
- Muy probable (5).

A pesar de esta clasificación, en el contexto de la información de mercados, las variables deben ser procesadas con el fin de extraer de ellas una medición cuantitativa. Los programas informáticos que se utilizan trabajan con números y por eso los valores que deben asignarse a las distintas opciones de una variable nominal u ordinal deben ser numéricas.

Hay otro motivo adicional, aunque este cada vez es menos relevante. Las encuestas que se hacen en papel deben ser posteriormente procesadas y grabadas en formato digital para su análisis. Esta labor de grabación de los cuestionarios puede ser ejecutada a una velocidad mucho mayor, siendo más operativo si las personas encargadas de hacerlo únicamente tienen que introducir un número en cada variable en lugar de escribir las respuestas completas.

Por ello, a la hora de diseñar un cuestionario, este cuenta con preguntas cerradas, que son aquellas en las que a cada opción de respuesta se le asigna un valor numérico, tal y como se puede observar en el ejemplo propuesto anteriormente. Este valor es el que queda registrado en la base de datos que servirá para la tabulación posterior. Estas preguntas habitualmente vienen precodificadas, es decir, los valores numéricos se definen desde un principio.

Es frecuente que los cuestionarios también presenten una serie de preguntas abiertas, es decir, preguntas en las que el encuestado puede contestar libremente y redactar la respuesta sin necesidad de ceñirse a una serie de respuestas dadas. Este tipo de preguntas también deben ser codificadas para que puedan ser analizadas.

Esta codificación se lleva a cabo una vez que está disponible el fichero de datos creando variables adicionales al lado de la variable original para no perderla y creando códigos que agrupen, en mayor o menor medida, respuestas similares.

Importante

Es conveniente no simplificar en exceso para no perder demasiada riqueza, aunque tampoco es conveniente excederse en la creación de códigos para evitar tener un montón de categorías con un número de respuestas muy reducidas.

Cuando todas las variables están codificadas se puede tabular el cuestionario con el fin de comenzar a analizar los resultados obtenidos.

La **tabulación** de datos es la acción que permite hacer un recuento de los datos que han obtenido mediante la recogida de información a través de la encuesta empleada.

La tabulación permite crear tablas de datos mediante las que se puede analizar toda la información numérica. De este modo se pueden crear gráficos que representen de manera visual los resultados y extraer con mayor facilidad las conclusiones que respondan a los objetivos planteados desde un comienzo. Así, con la tabulación se realizará el análisis estadístico de los datos a partir de unos claros objetivos y unas variables preestablecidas que determinarán el tipo de información que se puede extraer.

La tabulación se lleva a cabo con programas informáticos como el IBM® SPSS, BarbWin o R. Aunque también es posible hacer gran parte de estas tabulaciones con herramientas menos especializadas como Excel®. Obviamente, es mucho más costoso realizar en Excel las tareas disponibles en un paquete estadístico específico; sin embargo, es frecuente que se combine el uso de un paquete estadístico con Excel para alternar tareas entre uno y otro.

Las ventajas de utilizar los programas de análisis estadístico especializado para tratar datos de un SIM empresarial son la posibilidad de agilizar muchos de los cálculos que implica la tabulación y, debido a que los cálculos los lleva a cabo el propio programa, se evitan errores humanos que se pueden dar durante el proceso.

Los tres programas citados permiten aplicar técnicas estadísticas descriptivas, pero también análisis multivariantes de mayor complejidad. IBM SPSS es el programa referente porque es el más conocido y su uso está más extendido. BarbWin es una alternativa muy buena y con una interfaz con mejor usabilidad. R es una aplicación que requiere de un mayor grado de conocimiento y del aprendizaje de código para poder aprovechar su potencial, por lo que es el programa más complejo y menos indicado para los primeros contactos con el mundo de la tabulación de datos.

Excel es la herramienta por excelencia para la construcción y operación más básica con las bases de datos. Es el perfecto aliado para facilitar la revisión de datos y para utilizarlo en todo momento; de hecho, una vez que se tabulan los datos en un programa estadístico, es conveniente pasar los datos a Excel para facilitar su manejo. Sin embargo, conviene hacer uso de uno de los programas anteriores para reducir los errores en la tabulación y los tiempos de trabajo.

Si el fichero de datos proviene de una encuesta *on-line,* es habitual que ya tenga identificadas los tipos de cada variable que se van a tratar. Si se realiza un estudio en papel o por cualquier otro motivo en el fichero no está especificado para cada variable el tipo que es, es fundamental que se cumplimente esta información.

Las variables cualitativas nominales y ordinales pueden contabilizarse mediante la frecuencia absoluta, la frecuencia en porcentaje y, cuando se trata de variables ordinales, tiene sentido también la frecuencia acumulada.

La frecuencia absoluta es simplemente el recuento de las veces que se repite un mismo valor dentro de una distribución muestral. Es la manera de contar cuántos encuestados hacen alusión a cada posibilidad de respuesta de la pregunta que se analiza.

La suma de las frecuencias de todas las categorías da como resultado el conjunto total de la muestra analizada siempre y cuando no se trate de una pregunta que ha sido filtrada por otra previa, en cuyo caso sumará el número total de la muestra que cumpla el requisito establecido para la pregunta en cuestión.

Cuando se trata de una pregunta de respuesta múltiple, la suma del conjunto será mayor a la muestra, ya que cada encuestado podrá dar más de una respuesta elevando por tanto el número de las mismas.

La frecuencia en porcentaje es la forma más habitual de representar los datos estadísticos que se obtienen de un estudio porque es la manera más sencilla de interpretar unos datos y valorar las distribuciones. Pueden calcularse porcentajes horizontales o verticales, aunque lo habitual es analizar las tabulaciones con los segundos, que son los que permiten analizar las distribuciones dentro de cada variable.

La frecuencia acumulada se puede utilizar con las variables ordinales. En este caso, lo único que se hace es sumar al propio valor de una categoría a las frecuencias de todos los valores menores que esta. Esto también se puede aplicar a las variables cuantitativas cuando se codifican y se recodifican en variables ordinales mediante la creación de intervalos.

Cuando se trata con variables cuantitativas, además de las frecuencias ya descritas, se pueden hacer cálculos para obtener medidas de tendencia central como la media, mediana, moda, así como la varianza, desviación típica, coeficiente de variación o cuantiles.

Un ejemplo de tabulación es el que se muestra en la tabla que se muestra a continuación. En ella se obtienen los resultados de una pregunta relacionada con la edad que se codifica mediante la creación de grupos de edad.

	Absolutos	Porcentaje	Porcentaje acumulado
De 18 a 25 años	152	21,7	21,7
De 26 a 35 años	131	18,7	40,4
De 36 a 45 años	104	14,9	55,3

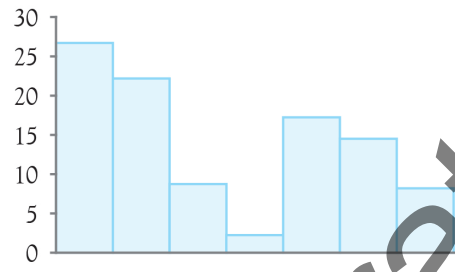
De 46 a 55 años	91	13	68,3
De 56 a 65 años	125	17,9	86,1
Más de 65 años	97	13,9	100
Total	700	100	

La tabulación es el paso previo a la representación de datos. En ocasiones, cuando una variable tiene muchas opciones de respuesta puede tener sentido mantener una tabla con los enunciados de las categorías y sus frecuencias en porcentaje para facilitar la lectura. Pero lo más habitual es que las frecuencias sean la antesala de la construcción de un informe con gráficos que faciliten la lectura de los datos obtenidos en la investigación comercial.

El propio Excel o PowerPoint® permiten crear diversos tipos de gráficos que se ajusten a las necesidades de los usuarios para la representación de los datos. Los informes habitualmente se hacen en PowerPoint, pero los tipos de gráficos que se pueden construir son exactamente los mismos que en Excel.

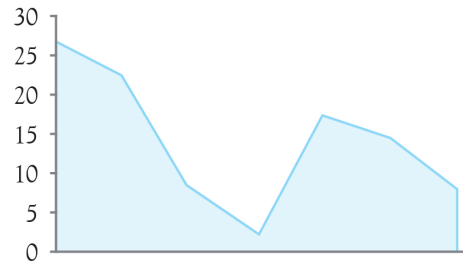
Dentro de las variables **cuantitativas** existen tres formas de representación de datos más habituales: histograma, gráfico de líneas y gráfico de áreas.

El histograma permite conocer fácilmente la evolución de los datos a lo largo de un periodo de tiempo determinado; por ejemplo, la evolución de un producto en el mercado.



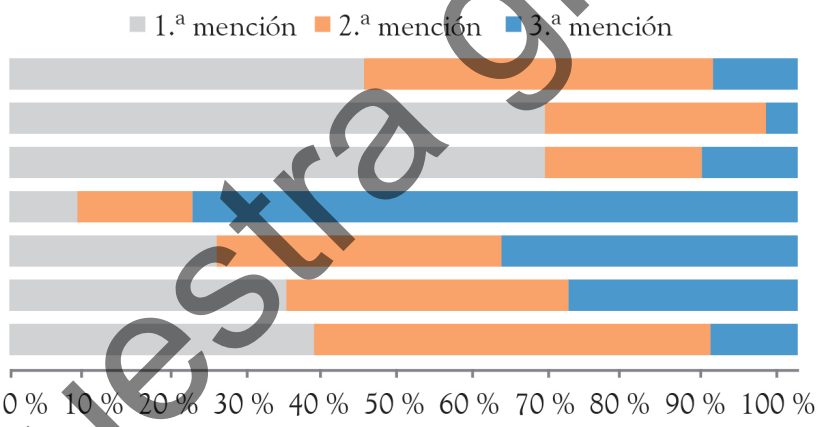
El gráfico de líneas puede usarse tanto para representar variables cuantitativas, en cuyo caso funcionan como un histograma, o con variables cualitativas politómicas. En este último caso realizan la función de un gráfico de barras. En la investigación comercial puede ser útil para considerar evoluciones y saber cómo actuar en el futuro.

El gráfico de áreas es un tipo de gráfico que es preferible evitar, ya que su lectura es compleja y, en líneas generales, no aporta mayor beneficio que un gráfico de líneas.



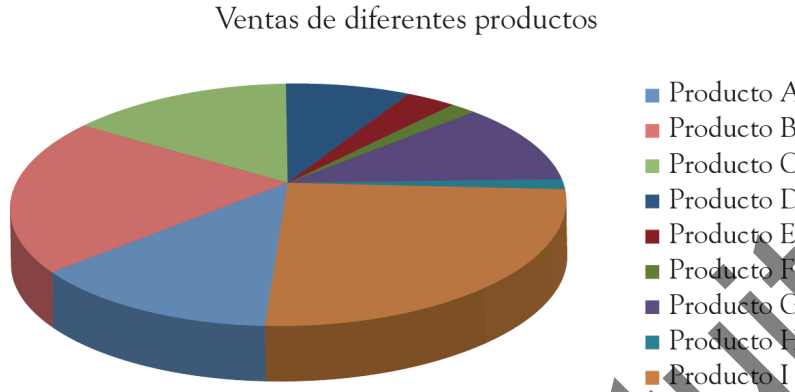
En cuanto a las variables **cualitativas,** las variables dicotómicas, generalmente, se representan con gráficos 100 % apilados o con gráficos de tarta porque facilitan la lectura. Este tipo de gráficos siempre muestran la totalidad de los casos y, por tanto, cuando solamente hay dos opciones de respuesta es muy fácil interpretar cuánta incidencia de casos hay en cada una de ellas.

El gráfico de barras 100 % apiladas es muy útil cuando una misma categoría tiene varias respuestas; por ejemplo, cuando se pide a los encuestados que seleccionen los tres problemas principales del momento e indiquen su orden de importancia. En este caso, de una serie de opciones el encuestado podrá elegir un total de tres y valorarlas con un 1, un 2 o un 3 en función del lugar que el encuestado les otorgue. Posteriormente, se puede realizar un gráfico que permite desglosar el porcentaje de menciones de una opción de respuesta en cada una de las tres opciones posibles. Este tipo de gráfico también es práctico cuando se tienen variables dicotómicas.



El gráfico de tarta o circular es muy utilizado, pero con frecuencia se hace un uso indebido de este, ya que no facilita la interpretación de datos cuanto se pretende representar mucha información con ellos.

El uso de gráficos con formato tridimensional tampoco es conveniente. Son tipos de gráficos que los paquetes ofimáticos incluyen por defecto, pero la lectura de estos dista mucho de ser óptima.



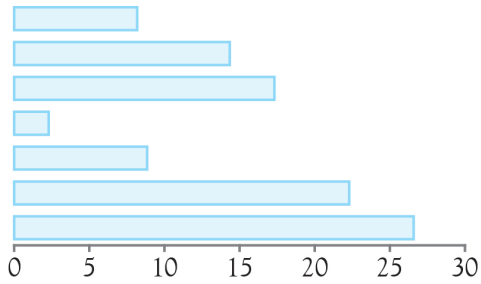
Fuente (modificado y adaptado): http://bit.ly/1yGJU0S

Cuando se quieren representar más de dos variables en un mismo gráfico se suele usar la tabla de frecuencia, el gráfico de barras, el gráfico de líneas o el gráfico de araña porque permiten comparar muy bien las tendencias de los diferentes valores representados.

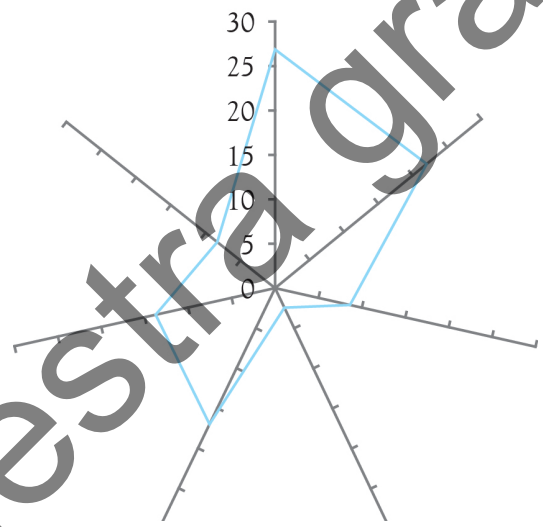
La tabla de frecuencias se utiliza en el procedimiento de representación de las variables cualitativas politómicas. Si se tiene en cuenta que «k» es el número de valores de la variable y «n» el tamaño muestral, la manera de realizarlas es la siguiente:

Valores de la variable	Frecuencia absoluta	Frecuencia relativa en porcentaje	Frecuencia acumulada en porcentaje
xi	ni	Fi = ni / n*100 %	Fi = sum(fi)
x1	n1	f1 = n1 / n*100 %	F1 = f1
x2	n2	f2 = n2 / n*100 %	F2 = f1 + f2
x3	n3	f3 = n3/n*100 %	F2 = f1 + f2 + f3
…	…	…	…
xk	nk	Fk = nk / n*100 %	Fk = f1 + f2 + f3 +…+ fk = 100 %
Total	n	100 %	

Para representar las variables de respuesta múltiple o de respuesta simple que son politómicas se utiliza el gráfico de barras, ya sean horizontales o verticales.



El gráfico de araña es especialmente útil en variables politómicas que hacen referencia, por ejemplo, a las características de un producto. Es una manera fácil de observar, de manera general, a qué características asocian los encuestados el producto que se vende o se quiere vender.



1.2. Trabajos previos a la codificación y tabulación de datos

Antes de iniciar la codificación y tabulación, es preciso llevar a cabo algunos trabajos para dejar el fichero con el que se va a trabajar en condiciones óptimas para evitar errores y problemas posteriores.

Para empezar es fundamental cerciorarse de que todas las variables contempladas en el cuestionario contienen datos. Esta es la revisión más básica e inicial que se debe realizar siempre.

Si se parte de un cuestionario en formato papel también es conveniente hacer una selección de casos aleatorios, en torno a un 10 %, y comprobar la grabación de datos de los cuestionarios que resulten seleccionados. Se debe evitar un porcentaje alto de errores; en caso de haberlo, habría que tomar una decisión respecto al trabajo de grabación que puede llegar incluso al punto de retroceder hasta el reinicio del proceso.

Este proceso no puede avanzar hasta que el fichero contenga todas las variables con sus respectivos datos introducidos de forma correcta, para lo cual es necesario introducir todas las etiquetas con todas las opciones de respuesta codificadas.

Además, es conveniente introducir el texto de la pregunta para explicar las variables analizadas, ya que de lo contrario es más complicado interpretar la información cuando se tabulen los datos. Hay que tener en cuenta que las tablas que se van a generar son herramientas muy útiles para las personas que quieran realizar otros estudios para el SIM mediante una investigación comercial.

Una vez introducidos los datos de las categorías de respuestas y tras configurar las distintas variables indicando su tipo ya se pueden realizar algunas explotaciones y tabulaciones previas a la tabulación final.

Ejemplo

Una vez que se han introducido todos los datos se pueden comprobar con una tabulación si se obtiene alguna media que puede indicar la existencia de datos incorrectos en la base de datos. De este modo, tabular algunas variables facilita la tarea de edición y limpieza de datos.

También es positivo preparar el fichero para la codificación de las preguntas abiertas, dejando creadas las variables donde se codificarán posteriormente las variables abiertas.

Un último aspecto fundamental que se debe tener en cuenta antes de pasar a depurar, codificar y tabular un cuestionario, es disponer de una variable que permita identificar cada encuesta de manera unívoca. Normalmente, esto se

consigue asignando un valor numérico o alfanumérico único a cada una de las encuestas, de modo que haya un dígito que identifique a cada una de ellas para que sean fácilmente identificables cuando se requiera su revisión.

Asimismo, esta variable permitirá asegurar, a la hora de depurar, que no se cometen errores al pasar datos de un programa a otro cuando se reordena una base de datos, proceso en el cual es frecuente alternar entre un programa de análisis estadístico especializado y Excel.

1.2.1. Edición de datos

La edición de datos es uno de los procedimientos que forman parte de la depuración de datos. Se debe tener en cuenta que la depuración de un fichero es una tarea compleja pero también lógica: cuando se descubren valores anómalos o incorrectos en la minoría de casos, se puede averiguar el dato real que sirve de base.

Es un proceso que en ocasiones puede ser laborioso y que implica una serie de decisiones (en ocasiones difíciles) que hay que reflexionar y abordar. Es recomendable documentar y guardar todos los cambios que se realicen en un fichero nuevo para tener la oportunidad de volver atrás si existiese un arrepentimiento en algunos de los pasos.

Tener siempre una documentación acerca de los procesos que se llevan a cabo y un fichero en el que puedan ser revisados los pasos que se han seguido, es muy útil para la realización de análisis adicionales posteriores.

Si en un momento determinado se ha simplificado una variable, unido variables, etc., y no se conocen las variables originales, es posible que estos análisis no sean realizables. Por tanto, siempre que sea posible, a la hora de codificar una variable abierta, de recodificar variables, etc., es conveniente hacerlo en variables nuevas que permitan mantener las originales, de manera que sea posible volver a ellas.

Cuando se trabaja con datos recogidos mediante una encuesta on-line una buena parte de los errores que se pueden cometer en la creación de una base de datos que almacena la información desaparecen. Este tipo de encuestas tienen precodificadas las variables cualitativas, de modo que en las preguntas cerradas solo existirán valores dentro del rango aceptado en cada pregunta.

Fuente: http://www.survio.com/

Con los cuestionarios on-line se evitan otros problemas típicos de las encuestas en papel como que el encuestador no realice algunas de las preguntas, que no cumpla los filtros de manera adecuada y que no recoja de manera fidedigna las respuestas que el encuestado da a las preguntas abiertas.

Este tipo de revisiones es conveniente hacerlas mientras la encuesta aún está activa, de manera que si es una encuesta on-line y se detecta que alguna variable no se está filtrando correctamente o que los datos no se están grabando de forma adecuada, hay tiempo de revisar la programación y solventar el problema.

Si es una encuesta en papel y se detecta que un encuestador en concreto o el conjunto de los encuestadores incurren en un error recurrente, se puede proceder a su reinstrucción o introducir algún tipo de instrucción adicional para cerciorarse de que no se mantiene esta dinámica y se obtienen todos los datos necesarios.

Cuando se revisan los datos grabados manualmente, existe la posibilidad de que se introduzca por descuido algún valor incorrecto. Si esto sucediese, habría que editar el dato cumplimentándolo si es posible y revisando la encuesta en la que se cometió el error.

Aunque en las encuestas on-line se evita, en gran medida, la introducción de datos incorrectos, se mantienen variables en las que se puede introducir algún tipo de error. Lo que sí se puede realizar es una minimización de los fallos mediante la utilización de reglas en la programación del cuestionario.

Es posible que en un campo numérico abierto alguien introduzca un dato que suscite sospechas sobre su incorrección.

Ejemplo

En una encuesta on-line hay un caso en el que el encuestado indica la edad de noventa y cino años. Cuando se tabulan los datos de la encuesta se ve que este dato eleva la media y, en comparación con el resto de casos, parece poco probable que este sea correcto. Además, teniendo en cuenta que se trata de un cuestionario on-line, la probabilidad de que sea cierto es menos creíble. Esto crea la situación de no saber la edad real de la persona encuestada.

En este contexto se puede optar por establecer este dato como desconocido, designándole el valor que representa «NS/NC (No Sabe/No Contesta)», o por editar el dato de la edad por la media, algo que también es habitual. También se podría considerar que la persona que introdujo el dato quiso poner cincuenta y nueve y, sin querer, alteró el orden de los valores (aunque es menos ortodoxo). La mejor opción sería tener la posibilidad de volver a contactar con la persona encuestada, aunque puede resultar mucho más complicado y costoso.

También puede ocurrir que en una encuesta se recojan medidas que generen respuestas inconsistentes, ya sea por error del encuestado, por un error en la programación o redacción del cuestionario o, incluso, puede ser algo que en un momento determinado se haga de manera deliberada para dar más facilidades al encuestado.

En cualquier caso, si se le pregunta al encuestado cuánto tiempo tarda en desplazarse de su casa al trabajo, puede haber personas que respondan en minutos y otras en horas. Habrá que identificar los casos, decidir la medida que se va a elegir (en este caso lo más lógico sería tomar minutos) y convertir aquellos datos que no están registrados a la medida escogida. De este modo, ya se podrá operar con esta variable y analizarla debidamente.

El manejo de los valores perdidos también es uno de los aspectos que hay que tener en cuenta en la edición de datos. Los valores perdidos aparecen en aquellos casos en los que la persona no ha respondido a una pregunta. Esto puede deberse a que la persona no haya contestado deliberadamente, a que haya habido algún tipo de error, etc. En cualquier caso, deben establecerse previamente unas reglas sobre el tratamiento de los valores perdidos.

Habitualmente, se les asigna el valor más alto dentro de un rango categórico; es decir, si una variable admite respuestas de un carácter, por ejemplo una pregunta cerrada de cinco respuestas que irán del 1 al 5, el código 9 se reserva para el valor perdido o NS/NC. Si la variable es de una cadena que requiere dos caracteres, los valores perdidos ocuparán el 99 y así sucesivamente.

Por contraposición, si las preguntas son semicerradas, es decir, si son preguntas con una codificación previa pero que admiten un apartado que, normalmente, se denomina «otros» (en el cual se puede especificar la categoría con mayor detalle), a este valor se le asigna el 98. Si la distribución de «otros» es muy elevada, una vez se tabulan los datos, es conveniente reducir ese apartado codificando sus respuestas.

Una vez que el fichero está preparado se procede a la limpieza de datos que permita trabajar con un fichero en el estado más optimo posible y evitar así incurrir en errores y en enfrentarse a datos falseados o erróneos.

1.2.2. Limpieza de datos

La línea que separa la edición de la limpieza de datos es algo difusa porque, realmente, sirven al mismo propósito: eliminar toda la información errónea de la base de datos para que se disponga de un fichero limpio y operativo que evite posteriores errores de lectura e interpretación.

La limpieza de datos es un proceso que debe llevarse a cabo porque siempre hay puntos y aspectos que se escapan y en los que el programa o el usuario pueden cometer una serie de acciones que concluyan con la obtención de un fichero de datos con discrepancias y una falta importante de consistencia en los datos.

Incluso se puede dar el caso de que, por un error informático o algún tipo de filtro programado con algún error, se esté obteniendo un fichero de datos incompletos. Por este motivo, cuando se trabaja con encuestas on-line, es conveniente

lanzar el cuestionario a campo de manera controlada y hacer una primera revisión del fichero antes de continuar. De este modo se tiene la seguridad de que el fichero se graba correctamente y de que los errores que pueda haber puedan ser subsanados mediante una depuración adecuada que finalice en una correcta limpieza del fichero.

Los procesos de depuración de ficheros, en general, no pueden ser totalmente computarizados, porque los programas pueden tratar como un error algo que no lo es y, del mismo modo, datos que se mueven en un rango adecuado pueden ser erróneos. Por este motivo, es importante que la máquina no posea todo el poder de decisión a la hora de eliminar o modificar valores de la base de datos, ya que se podría estar incurriendo en un error.

En la **revisión digital,** las comprobaciones se pueden realizar fácilmente empleando el propio Excel, aunque también se puede hacer el mismo tipo de tarea con otro paquete estadístico. Una de las comprobaciones más sencillas es ordenar los datos por columnas, creando una ordenación por variable y así establecer una **verificación de rangos** y realizar una comprobación de que los valores introducidos son correctos. Es una manera muy sencilla de descubrir esos datos en los que se ha introducido algún valor adicional, de manera accidental, en los valores aceptados dentro de un rango.

Ejemplo

Una variable de valoración de un servicio que puede contener valores de 0 a 10.

ID (Identificación)	Valoración
15009	9
15010	8
15011	9
15012	10
15013	9
15014	90
15015	1

En el ejemplo se puede observar claramente que en la encuesta 15014 se ha introducido un error de grabación o introducción de datos. Lo más probable es que la valoración correcta sea un 9 y que el 0 sea el dígito incorrecto. Sin embargo, es conveniente verificar esta información revisando la encuesta original si es posible, consultando al encuestado en cuestión, etc.

Si bien es cierto que este tipo de errores, como ya se ha comentado con anterioridad, se reducen o incluso quedan eliminados del todo si se emplean encuestas on-line o programadas en soportes digitales que introduzcan una serie de reglas y controles que impiden la introducción errónea de datos, hay otro tipo de errores que se mantienen independientemente de que se utilice esta tecnología.

Ejemplo

Ocurre con aquellos errores que se deben a la incorrecta introducción de datos por parte del encuestado, aun estando dentro del rango de valores aceptados. Si se le pide al encuestado que valore un producto de 0 a 10 y este introduce por error un 1 cuando quería introducir un 10, las aplicaciones programadas u on-line no lo solventarán. Sin embargo, se pueden detectar si en el mismo cuestionario se realizan otras preguntas y se constata que la valoración es mayoritariamente positiva.

Si se observa nuevamente las valoraciones del ejemplo anterior, queda en evidencia que la 15015 cuenta con una valoración significativamente inferior al resto y rompe en buena medida con la dinámica y tónica general de las puntuaciones. El valor que se recoge es correcto, por tanto, si se realiza un análisis computarizado sencillo, no se detectaría ningún error. Pero si la comprobación es manual, se observa claramente una anomalía.

Es probable, teniendo en cuenta el resto de puntuaciones, que la persona que grabó los datos o el propio encuestado quisiera dar una valoración de 10 y, por un error humano, no haya introducido el 0. Es algo que se comprueba fácilmente verificando el cuestionario original.

En el caso de que no se pudiese volver a contactar con el encuestado ni recurrir al cuestionario original, (porque es on-line y los datos disponibles provienen directamente del encuestado) se podrían revisar algunas de las otras variables. Con total seguridad, si se trata de una encuesta de valoración de un servicio, existen variables adicionales que pueden dar una pista del grado de satisfacción del mismo encuestado, como la pregunta sobre por qué ha dado tales valoraciones. De este modo, se puede descubrir si realmente se trata de una persona que ha tenido una mala experiencia y está descontenta o, por el contrario, se trata de alguien que por error omitió un 0 a la hora de introducir los datos.