Instituto de Investigación Psicológica (IPsi)
Universidad de Puerto Rico
Recinto de Río Piedras

MANUAL DE PREVENCIÓN DE ERRORES Y LIMPIEZA DE DATOS EN SPSS

Preparado por:
Lunimar Curbelo González
Asistente de Investigación
Unidad de Investigación

Febrero de 2011

PREVENCIÓN DE ERRORES Y LIMPIEZA DE DATOS EN SPSS

Objetivos

Este Manual fue preparado con el objetivo principal de ofrecer algunas guías para la revisión de bancos de datos de los proyectos de investigación del IPsi. Otros objetivos del Manual consisten en presentar las situaciones más frecuentes que pueden provocar dificultades durante el análisis de datos; proveer una guía para prevenir errores durante la construcción del banco de datos; y discutir la utilidad y el proceso de limpieza de datos (*data cleaning*). Con esto se busca que los datos obtenidos se encuentren en las condiciones adecuadas al momento de realizar los análisis estadísticos pertinentes a la investigación, y que el impacto de los errores sea minimizado en nuestros resultados.

Proceso de control de calidad de los datos

El proceso de control de calidad de datos nos permite identificar y corregir los errores que ocurren en nuestros datos, o al menos que su impacto sea minimizado en los resultados obtenidos. Dentro de los procesos de control de calidad de los datos, hay varios pasos a seguir:

- *Prevención de errores:* precauciones previas a que cualquier error ocurra
- *Monitoreo de datos:* observación del curso de los datos para detectar anomalías
- *Limpieza de datos:* se realiza para detectar y corregir anomalías
- *Documentación:* Reporte de procedimientos llevados a cabo para identificar errores y editarlos, con su justificación de ser necesario, para llevar un control ético en la investigación

En este Manual se hará énfasis en los pasos de prevención de errores y la limpieza de datos.

Prevención de errores en la construcción de bancos de datos: puntos a considerar

La prevención de errores consiste en, como implica su nombre, evitar que ocurran errores en el proceso de recolección, entrada y análisis de datos, de manera que los resultados que obtengamos nos resulten confiables. El proceso de prevención comienza una vez empezamos a planificar el estudio que nos interesa. Por ello, se considera que la planificación y ejecución del estudio debe ser cuidadosa, ya que las decisiones que tomamos en las distintas etapas (revisión de literatura, formulación de hipótesis, la elección del diseño del estudio, localización de los sujetos, etc.) del estudio pueden afectar la calidad de los resultados.

Es altamente recomendable realizar un **estudio piloto.** Éste nos permite identificar posibles dificultades con los cuestionarios (preguntas que no se entienden, preguntas dejadas en blanco de manera frecuente) y otros aspectos del proceso de la investigación, como por ejemplo, que el equipo tecnológico (cámaras de video, grabadoras, computadoras, etc.) a utilizarse funcione adecuadamente. También es muy importante proveer adiestramiento adecuado a quienes realizan a quienes colaboran en el estudio, realizando entrevistas y/o codificando datos. Deben tener claro lo que se está

buscando en la investigación, a donde se dirigirán sus observaciones, y deben estar familiarizados con los instrumentos que se estarán utilizando, tanto en formato de papel y lápiz como los computadorizados.

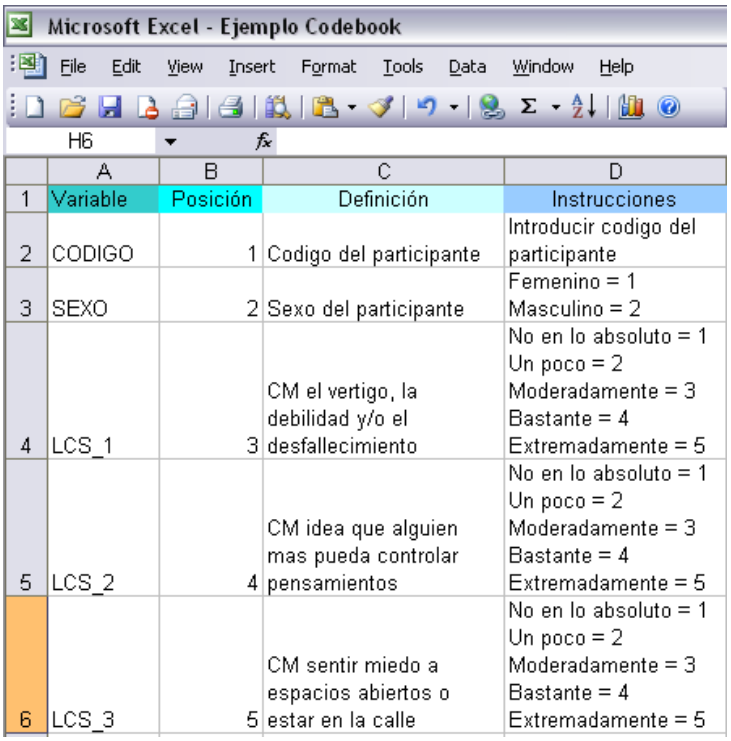
Otro aspecto a considerar son las escalas y medidas o instrumentos que se estarán utilizando en el estudio. Los instrumentos se eligen de acuerdo al propósito y objetivos del estudio. Por ejemplo, si mi objetivo es estudiar la percepción de la violencia doméstica en mujeres casadas, no puedo administrar un instrumento que mida percepción del divorcio, a menos que esto sea parte de otros objetivos del estudio.

También es fundamental la revisión de literatura, particularmente lo que nos informa en torno la confiabilidad y la validez de los instrumentos. Este aspecto es de suma importancia, ya que estas propiedades psicométricas nos aseguran que el error estándar de la medida se mantenga en un mínimo. Por otro lado, si se construye un instrumento, este debe estar bien planificado y diseñado (p. ej. tipo de preguntas, formato de respuestas, qué estadísticas se utilizarán para analizar la información recogida, entre otros), ya que el mismo tendrá un impacto en los tipos de análisis que podremos realizar. Por ejemplo, se puede perder información si elegimos hacer la pregunta cerrada cuando debería ser abierta (i.e. edad).

Otro aspecto a considerar es el desarrollo de una estructura para el estudio, en particular para el banco de datos. Antes de comenzar a construir un banco y eventualmente introducir datos de los cuestionarios, es necesario preparar un *codebook* (libro de codificaciones) que permita ver la operacionalización de las variables. El *codebook* es un resumen de las instrucciones que se utiliza para convertir la información obtenida a un formato que SPSS pueda comprender. Además es una guía que permite identificar los valores que se supone puedan atribuirse a cada variable. En el *codebook* se especifica al menos:

- Nombre (Variable) y la definición (Label) de las variables
- Asignarle los valores a cada una de las repuestas posibles (Values/Instructions)

Una opción para hacerlo es utilizando Excel o en una libreta por escrito; el programa QDS también tiene esa opción. Un ejemplo de un *codebook* se presenta a continuación.



Microsoft Excel - Ejemplo Codebook

File Edit View Insert Format Tools Data Window Help

H6

	A	B	C	D
1	Variable	Posición	Definición	Instrucciones
2	CODIGO	1	Codigo del participante	Introducir codigo del participante
3	SEXO	2	Sexo del participante	Femenino = 1 Masculino = 2
4	LCS_1	3	CM el vertigo, la debilidad y/o el desfallecimiento	No en lo absoluto = 1 Un poco = 2 Moderadamente = 3 Bastante = 4 Extremadamente = 5
5	LCS_2	4	CM idea que alguien mas pueda controlar pensamientos	No en lo absoluto = 1 Un poco = 2 Moderadamente = 3 Bastante = 4 Extremadamente = 5
6	LCS_3	5	CM sentir miedo a espacios abiertos o estar en la calle	No en lo absoluto = 1 Un poco = 2 Moderadamente = 3 Bastante = 4 Extremadamente = 5

Es muy útil verificar que nuestro banco esté bien construido. Una forma de verificarlo es utilizando la opción de *codebook* en SPSS. Para hacerlo, vamos a **Analyze ->Reports->Codebook.** El programa te dará las opciones de aquella información del banco que se desea revisar. La información mostrada en las tablas sobre nuestras variables debe coincidir con el *codebook* que habíamos hecho previamente. Esta función de *codebook* de SPSS no puede realizarse previa a la construcción del banco de datos.

Antes de entrar los datos también es recomendable revisar los cuestionarios tan pronto como son entregados y verificar si las respuestas son **legibles** y si están **completas**. De forma si notamos una irregularidad, podemos hacer que el participante clarifique cualquier error a tiempo.

Problemas y errores frecuentes

Podemos clasificar los errores más frecuentes de la siguiente manera:

- Falta o exceso de datos
- Valores extremos incluyendo inconsistencias

- Patrones extraños en las distribuciones (asimetría, kurtosis puntiaguda o plana)
- Resultados inesperados de los análisis y otros tipos de inferencias o abstracciones

En la siguiente tabla se muestran en detalle los errores que nos encontramos comúnmente en las distintas etapas de los datos.

Tabla de problemas y errores frecuentes

Etapa de los datos	Fuentes de problemas: falta o exceso de datos	Fuentes de problemas: valores extremos e inconsistencias
Cuestionario	Cuestionario perdido	El valor correcto fue introducido en el lugar equivocado
	Cuestionarios dobles, recogidos repetidamente	No es legible
	Contestaciones en blanco	Error en la escritura
	Más de una opción es elegida cuando no se permite	La respuesta que dio está fuera del rango esperado
Banco de datos	Falta o exceso de datos que traen los cuestionarios	Valores extremos e inconsistencias que traía el cuestionario
	No se introdujo el dato	Valor introducido erróneamente
	Se introdujo el dato dos veces por error	Valor cambiado incorrectamente durante limpieza de datos previa
	El valor se introdujo en el lugar equivocado	Error de transformación* (programación)
	Valores suprimidos o duplicados durante el manejo del banco	
Análisis	Falta o exceso de datos que trae la base de datos	Valores extremos e inconsistencias que traía el banco de datos
	Error de extracción o transferencia de los datos**	Extracción o error de transferencia**
	Supresiones o duplicaciones por el analista	Errores de clasificación (*sorting errors*)***
		Errores de limpieza de datos

***Error de transformación:** ocurre cuando uno hace una recodificación de una variable (se verifica con un case summary; se hace antes y después de recodificar y se comparan las contestaciones de cada participante)*
***** Error de transferencia:** cuando transfieres de SPSS a SAS, MPlus u otro programa, en las reestructuraciones de los bancos de datos o de QDS a SPSS*
****** Error de clasificación** (sorting errors): posibles errores en la codificación*

Limpieza de datos

La limpieza de datos (*data cleaning*) implica lidiar con los problemas en los datos una vez han ocurrido (la prevención de errores puede reducirlo, pero no eliminarlos). Envuelve ciclos repetidos de cernimiento, diagnóstico y edición de los datos que nos resultan sospechosos. Una forma de detectarlos de manera más eficiente es teniendo predefinidas las fuentes y tipos de errores en cada etapa del flujo de datos: debe basarse en el conocimiento de errores técnicos y valores esperados dentro de un rango de valores normales.

En bancos muy grandes, se recomienda realizar la limpieza de datos con el 10% de los casos. Esto se hace de la siguiente forma: **Data-> Select cases-> Random sample of cases-> Sample-> Approximately 10 %of all cases-> Continue-> Ok.*

Cernimiento (data screening)

Para poder identificar un dato como sospechoso, primero hay que tener predefinidas las expectativas de una distribución normal, formas de la distribución, y la magnitud de las relaciones. Estos criterios pueden tomarse en cuenta mientras se recogen los datos, se entran en el banco y mientras se esté trabajando con ellos. Al compararse los criterios con los datos obtenidos con distribuciones de frecuencia de nuestras variables, podemos identificar datos, patrones o resultados dudosos. Esto es lo que nos permite hacer el cernimiento de datos.

Métodos para realizar cernimiento de datos

Hay varios métodos para realizar el cernimiento de datos, entre los que se encuentran:

- Exploración gráfica de las distribuciones: *boxplots*, histogramas, *scatterplots*
- Distribuciones de frecuencia o tablas de contingencia (*cross-tabulations*)

- Resumen de las estadísticas
- Detección estadística de valores extremos (*outliers*)

Uno de los más utilizados es la distribución de frecuencia para revisar variables categóricas en SPSS. Para realizarla, se deben seguir los siguientes pasos:

- En el menú principal: **Analyze -> Descriptive Statistics -> Frequencies**
- Escoger las variables que deseas verificar (sexo, estado marital, nivel educativo, ¿cuánto le molestó X situación?, etc.).
- Selecciona la flecha para moverlas al espacio de variables
- En el botón de estadísticas elige **Mínimo** y **Máximo** en la sección de **Dispersión**
- Selecciona **Continue -> Ok**

Las tablas que se presentan a continuación son un ejemplo de lo que aparecerá en la pantalla del *output*:

Statistics

Cual es el estado civil

N	Valid	139
	Missing	1
Minimum		1
Maximum		6

Cual es el estado civil

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Casada/o	35	25.0	25.2	25.2
	Separada/o	1	.7	.7	25.9
	Divorciada/o	11	7.9	7.9	33.8
	Viuda/o	1	.7	.7	34.5
	Soltera/o	79	56.4	56.8	91.4
	Convivencia/o	12	8.6	8.6	100.0
	Total	139	99.3	100.0	
Missing	99	1	.7		
Total		140	100.0		

Aquí nos podemos preguntar qué vamos a verificar. En primer lugar, los **valores mínimos y máximos:** estos ¿hacen sentido? ¿se encuentran dentro del rango esperado? Aquí el *codebook* es muy útil para corroborar los valores posibles. En este ejemplo, si mi instrumento indica que hay 6 posibles respuestas (del 1 al 6), este valor no debe ser menor al mínimo ni mayor al máximo (por ejemplo, no debe aparecer un 0 o un 7). También vamos a mirar la **cantidad de valores válidos y *missing values***: ¿hay muchos *missing values*? ¿por qué? En el ejemplo, se muestra un 99, valor que se asignó como *missing value* y que el programa reconoce como tal. Si no le indicamos al programa que este valor es un *missing value*, va a contarlo como válido.

En el caso de las variables continuas (edad, total de la escala X, etc.) vamos a realizar un resumen descriptivo de la variable, haciendo lo siguiente:

- En el menú principal: **Analyze -> Descriptive Statistics -> Descriptives**
- Seleccionas las variables que se desean verificar, y se mueven al espacio de variables
- En el botón de **Options**, seleccionar al menos la **Media**, la **Desviación Estándar**, el **Mínimo** y el **Máximo**
- Marca ***Save standarized values as variables***
- Selecciona **Continue -> Ok**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Total LCS 36	140	37	141	70.62	25.519
Valid N (listwise)	140				

¿Qué vamos a verificar? Comenzamos con los **valores máximos y mínimos:** ¿hacen sentido? En el ejemplo, el LCS-36 tiene una puntuación mínima de 36 y una máxima de 180 puntos. En este caso, 37 y 141, que son los valores mínimo y máximo de esta muestra, caen dentro del rango de posibles valores. En el caso de la **media:** ¿hace sentido? Si la variable es la totalidad de una escala, ¿es esta media lo que se esperaría, según investigaciones anteriores? ¿El valor, se encuentra en la mitad del rango de puntuaciones posibles o se acerca a un extremo? En el ejemplo, en una muestra no clínica, esta media hace sentido (está más ubicada en los valores bajos). Si la muestra

fuera clínica, la media debería ubicarse en los niveles de psicopatología, que es lo que mide el LCS-36.

Al marcar ***Save standarized values as variables,*** podemos hacer una detección estadística de valores extremos. Siguiendo el ejemplo, el programa nos va a crear una nueva variable con las puntuaciones crudas obtenidas del LCS-36 convertidas a puntuación z. Se considera un valor verdadero extremo si z > 3.29 (para más detalles sobre valores extremos, ver la sección *Edición de datos: Valores Extremos/Outliers* en la página 12).

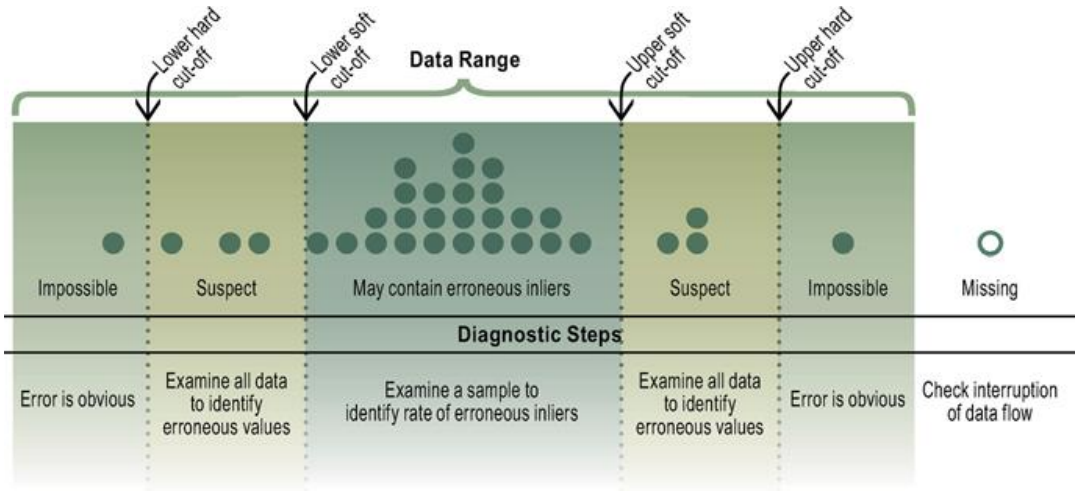
Diagnóstico: ¿Qué tipo de datos problemáticos encontramos?

Los datos problemáticos que podemos encontrar en nuestros datos pueden clasificarse de la siguiente forma:

- Erróneo (*inlier*)
- Verdadero extremo
- De causa desconocida
- Exceso de valores perdidos

En el diagrama de se muestra como se pueden localizar estos datos erróneos en una distribución normal.

Diagrama de datos problemáticos (Tomado de Van den Broeck, J., Cunningham, S. A., Eekels, R. & Herbst, K., 2005).



Dependiendo del tipo de error encontrado se llevará a cabo la corrección. Pero una vez encontramos un dato que nos produce sospecha, lo indicado sería verificar si este

dato ha sido consistente durante el flujo de datos que estamos manejando. Claro que lo primero que vamos a mirar es el cuestionario del participante.

Encontrar el error específico

Cuando ya sabemos que hay un error, el próximo paso es localizar donde se encuentra el mismo. Hay dos métodos para realizar este proceso. El primero es por medio del **Find and Replace**:

- En el *Data View* (la pantalla debe mostrar los valores numéricos que corresponden a cada variable), selecciona la columna (un solo click en donde la identifica, p. ej. género); debe quedar sombreado tanto el nombre de la variable como sus datos
- En el menú principal, selecciona **Edit -> Find** (en la versión SPSS 18.0, buscar los binoculares o CTRL+F)
- El valor que debe ser reemplazado se introduce en el campo **Find** del cuadro de diálogo
- Selecciona **Find Next** para identificar el caso donde ocurrió el problema y buscar el cuestionario que le corresponde
- El nuevo valor se introduce en el campo **Replace with** y el programa de manera automática lo corrige**;** hay que tener mucho cuidado con NO oprimir **Replace All**, ya que en ocasiones podemos tener más de un dato incorrecto con el mismo valor y que el valor correcto sea distinto
- Una vez se reemplace, tira las frecuencias nuevamente para asegurarte de que el error ha sido corregido

El otro método es mediante una **tabla de valores extremos para localizar errores en los datos:**

- En el menú principal: **Analyze -> Descriptive Statistics -> Explore**
- En la sección de ***Display***, marca ***Both*** (o ***Statistics*** si no se desean las gráficas)
- Selecciona las variables que te interesen y las mueves a **Dependent list**

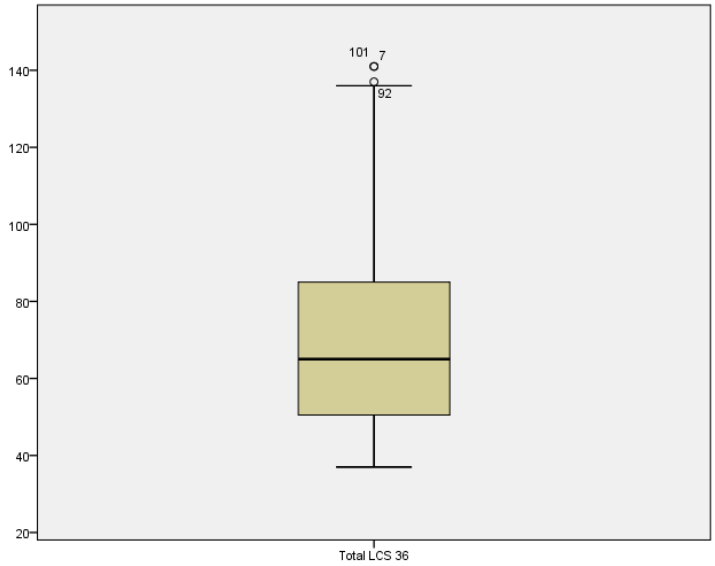
- En la sección **Label cases**, selecciona **ID** de tu lista de variables; esto va a permitir que puedas identificar el caso, y puedes revisar el cuestionario que tiene el error
- En la sección de **Statistics** elige **Outliers.** Si no te interesan**,** puedes quitar la opción de **Descriptives**, que viene marcada automáticamente. Selecciona **Continue**
- En la sección de **Options**, marca **Exclude cases pairwise** (tiene que ver con el manejo de los *missing values*, para mayor información pasar a la página 14) **-> Continue -> Ok**
- La tabla muestra: los valores más altos y los más bajos y te da la identificación de las personas con esa puntuación (2da columna). Ojo: NO es la columna de Case Number

Extreme Values

			Case Number	Codigo del participante	Value
Total LCS 36	Highest	1	7	7	141
		2	100	101	141
		3	91	92	137
		4	131	132	136
		5	76	77	127
	Lowest	1	122	123	37
		2	109	110	37
		3	10	10	37
		4	5	5	38
		5	103	104	40[a]

a. Only a partial list of cases with the value 40 are shown in the table of lower extremes.

En la tabla podemos observar los cinco valores más altos y los cinco valores más bajos de las puntuaciones obtenidas de la muestra. Estos valores están identificados, de manera que si hubiese un error (un valor extremo) podemos saber en cuál caso (o casos) ocurrió y por qué.



La gráfica que se muestra es un *boxplot*. El *boxplot* es útil cuando se desea comparar la distribución de las puntuaciones en las variables continuas solas o con el efecto de una variable categórica. Por lo pronto, la línea central nos indica la mediana de la distribución. La caja contiene el 50% de los casos. La línea que sale a cada extremo de la caja representa la extensión del rango de puntuaciones de la muestra. Los pequeños círculos, por otro lado, muestran los valores que SPSS considera que son *outliers*. El número en el círculo es la identificación del caso o los casos que resultan ser valores extremos. Por ello, el *boxplot* nos permite observar gráficamente estos valores extremos de la distribución.

Edición de datos: Valores Extremos/Outliers

Los *outliers* o valores extremos son puntuaciones que están distanciadas del resto de los datos y por lo tanto pueden tener un impacto en los análisis estadísticos. Básicamente, hay dos tipos: 1) ***univariados*** se refiere a aquellos que ocurren en una sola variable y 2) ***multivariados*** son valores extremos que ocurren en una combinación de variables. Las razones para tener un *outlier* son:

- Error en la entrada de datos
- No se especificó el código de *missing values,* y estos se están leyendo como datos verdaderos
- El valor extremo no forma parte de la población de la cual se conforma la muestra

- El *outlier* es parte de la población que se desea para la distribución, pero se ve como un caso extremo

¿Cómo podemos manejar los valores extremos? Hay varias opciones, que dependerán de la razón por la cual se obtuvo el valor extremo. **Find and Replace** es una opción si fue un error en la entrada de datos. Otra opción es eliminar los casos extremos, si no forman parte de la población de la cual se conforma la muestra. Cuando forma parte de la población, se puede hacer una transformación, si es que es parte de una distribución no normal (para esto hay que revisar la normalidad). Transformación es el proceso de aplicar una función matemática a todas las observaciones en un grupo de datos, usualmente para corregir anormalidades en la distribución en su simetría o kurtosis. Las transformaciones más comunes son:

- Raíz cuadrada (con una distribución moderadamente asimétrica)/ asimetría positiva, varianzas desiguales
- Logaritmos (con una distribución sustancialmente asimétrica)/ asimetría positiva, varianzas desiguales
- Transformación recíproca (asimetría positiva/ varianzas desiguales)
- Invertir (con una distribución extremadamente asimétrica)/ asimetría negativa

El problema con las transformaciones es la dificultad de realizar interpretaciones de una variable transformada, por ello son el último recurso. Además, no hay consenso en cuanto a considerar que transformar los datos es la mejor opción.

Otra opción para lidiar con valores extremos es cambiar la puntuación.[*] Hay varias alternativas para hacer este proceso:

- Cambiarla por el próximo valor más alto más uno
- Convertirlo desde una puntuación z (se considera un *outlier* si z > 3.29), cambiar por el valor de la media más tres desviaciones
    - Buscamos el valor z que corresponde a 3
        - $X = (z \times s) + \overline{X}$ es la fórmula que nos permite calcular cuál puntuación cruda corresponde a z = 3 (o 3.29)

[*] Hay opciones para editar outliers y missing values que deben considerarse con mucho cuidado, ya que implican alterar datos y esto levanta cuestionamientos; si van a utilizarse queda a discreción del investigador, pero esto debe ser documentado y justificado

- En la fórmula, se calcula la media y la desviación estándar de los datos, recordando que esto se hace en la sección de ***Options*** del resumen de las estadísticas descriptivas de la variable
- Le sumamos tres veces la desviación estándar a la media
- Reemplazamos los valores de los *outliers* con esa puntuación

También podemos cambiar por el valor de la media más dos desviaciones estándar, que es una variación del procedimiento anterior.

Una buena opción, y que se ha realizado en proyectos anteriores del IPsi, ha sido hacer el análisis de los datos con y sin *outliers* y se informan los resultados.

Edición de datos: Inliers erróneos

Se debe tener mucha precaución con los ***inliers*** **erróneos:** estos datos son generados por error pero pueden caer dentro del rango esperado, como introducir un 2 en vez de un 3 en una variable que va del 1 al 4. Algunos pueden ser detectados cuando son relacionados con otras variables. Por ejemplo, puede que un varón (codificado como 2) por error conteste que sí (codificado como 1) a la pregunta de si está embarazada (esto puede ocurrir por el formato en el que se presentan las respuestas en el instrumento). Al realizar el cernimiento esta situación puede pasar por desapercibida, ya que tanto el 2 (sexo) como el 1 (está embarazada) son valores legítimos para esas variables respectivamente. Hay varios métodos para lidiar con los ***inliers*** **erróneos**, desde tirar un *scatterplot*, un análisis de regresión o indicándole al programa combinaciones inválidas entre variables mediante un *syntax.* También es muy importante conocer las características de la muestra (i. e. cuántas mujeres y cuántos hombre hay).

Edición de datos: Missing Values (MV)

Como indica el nombre, son valores que faltan. Cabe señalar que no es raro obtener datos faltantes en investigaciones con humanos, al contrario, es lo más usual. Lo importante es identificar con qué frecuencia ocurren y si hay algún patrón. SPSS automáticamente considera *missing values* tanto a las celdas en blanco (*system* MV) y los valores designados como tal por el investigador (*discrete* MV). Hay dos tipos de MV:

- ***Aleatorios*****:** son aquellos que ocurren al azar, sin un patrón, usualmente cuando son pocos los valores perdidos

- ***No aleatorios*:** cuando podemos identificar un patrón o hay muchos valores perdidos

Existes varias opciones para lidiar con MV. Una de ellas es dejarlos como están. SPSS tiene dos opciones para ello antes de realizar el análisis de datos. En *exclude cases listwise* se incluyen los casos en el análisis sólo si tienen ***todos los datos en todas las variables*** que están incluidas en el *variable box* para ese caso. El problema con esta opción es que puede producir la eliminación excesiva e innecesaria de casos. En el caso de *exclude cases pairwise* se excluye el caso solo si faltan datos requeridos para el análisis específico, mientras que el resto seguirá siendo incluido en cualquiera de los análisis. Esta es la opción más recomendada por algunos autores. Otra posible, pero muy debatible, alternativa es el reemplazo de los valores por media. Es una opción disponible en algunos procedimientos de SPSS como regresión múltiple, en la cual se calcula la media para una variable y se le da este valor al *missing value*. La literatura no apoya esta opción, ya que puede producir estimados altamente parcializados. Hay que tener en cuenta que en SPSS las opciones para los MV varían según el análisis que va a realizarse.

El Missing Value Analysis es un procedimiento que tiene SPSS (es un *add-on*, programa aparte) que permite describir los patrones de los MV. Existen dos patrones de los MV. El ***Item nonresponse*** ocurre cuando la persona omite *k* reactivos en una parte del cuestionario por alguna razón. El ***Attrition/ wave nonresponse*** ocurre cuando la persona no se encuentra al recoger nuevamente datos de la medida en un estudio longitudinal (i. e. hay una medida pre intervención pero no de post intervención).

También hay mecanismos de datos perdidos o explicaciones que se dan para cada patrón de MV. Los datos son *Missing Completely at Random (MCAR)* si el mecanismo de datos perdidos es totalmente un proceso al azar, como lanzar al aire una moneda; también si la causa de la pérdida no está correlacionada con la variable que contiene los datos perdidos. Los MCAR tienen como consecuencia que no hay un sesgo de la estimación si la causa de la pérdida de los datos es omitida del modelo de MV. Los MV son *Missing at Random (MAR)* (también conocidos como *ignorable missingness*) si la causa de este mecanismo está correlacionada con la variable que contiene los datos perdidos, pero las variables que representan esta causa se han medido y así están disponibles para la inclusión en el modelo de MV. Al ser incluidas en el modelo se corrige el sesgo que está asociado con ellas. Los MAR NO ocurren al azar. Finalmente, en los *Missing not at*

Random (MNAR), la causa de los MV está correlacionada con la variable que contiene los MV, pero la causa no ha sido medida o no está disponible para incluirlo en el modelo de MV. Aquí la pérdida se relaciona con los datos perdidos, aún después de condicionar todos los datos disponibles, por ello NO es al azar.

En el Missing Value Analysis, la función EM (expectation-maximization) es la que nos permite estimar la puntuación perdida de los sujetos, y provee distintas herramientas descriptivas para analizar el *missing data*. El EM es un proceso de valores perdidos basado en datos, en el cual mediante ecuaciones de regresión se manejan los datos perdidos en un primer paso, y luego el análisis principal de los datos se hace en un segundo paso.

Una de las funciones más importantes para analizar los MV es la imputación de datos. Esto se refiere al reemplazo de valores que faltan mediante la regresión o métodos de EM. La imputación de datos puede ser *simple* si genera valores para los MV en una sola variable, mediante EM o regresión. También puede ser *múltiple,* si genera posibles valores para los MV creando varios conjuntos "completos" de datos. Al igual que el EM, la imputación de datos es un procedimiento de valores perdidos basado en datos. En los análisis, no se deben realizar más de un 15% de imputaciones de datos por cada variable.

Además del EM y la imputación de datos, otros procedimientos altamente recomendables son el Multiple Group Structural Equation Modeling y Full Information Maximum Likelihood for SEM (Structural Equation Modeling)- FIML (el programa Amos es uno de los más populares para trabajar con SEM, ya que provee una estimación del parámetro y errores estándar razonables para los MV). Estos dos procedimientos de valores perdidos son basados en modelos, lo que significa que trabajan con los MV al mismo tiempo en que se realiza la estimación del parámetro que estamos buscando. Aunque no hay un consenso sobre cual es la mejor forma de lidiar con los MV en la literatura, algunos consideran que estas son las formas más aceptables para trabajar con ellos.[*]

[*] Para mayor información ver Graham, J. W., Cumsville, P. E. y Elek-Fisk, E. (2003). Methods for Handling Missing Data. En J. A. Schinka y W. F. Velicer (Eds.) *Comprehensive handbook of psychology, Vol 2: Research methods in psychology* (pp. 87-114). NY: Wiley & Sons, Inc.

Bibliografía

Data Screening Check List (s.f.). Recuperado el 18 de octubre, 2010, en http://www.csun.edu/~ata20315/psy524/docs/Data%20Screening%20Check%20List.doc

Field, A. (2009). *Discovering Statistics Using SPSS* (3ra. ed.).UK: Sage.

Graham, J. W., Cumsville, P. E. y Elek-Fisk, E. (2003). Methods for Handling Missig Data. En J. A. Schinka y W. F. Velicer (Eds.) *Comprehensive handbook of psychology, Vol 2: Research methods in psychology* (pp. 87-114). NY: Wiley & Sons, Inc.

Pallant, J. (2005). *SPSS Survival Manual (2nd ed.)*.UK: Open University Press.

SPSS Inc. (s.f.). *SPSS Missing Values 17.0*. Chicago, IL: Author.

Van den Broeck, J., Cunningham, S. A., Eekels, R. & Herbst, K. (2005). Data Cleaning: Detecting, Diagnosing and Editing Data Anormalities. *PLoS Medicine, 2*(10), 966-970.