Francisco M. Ocaña Peinado

Técnicas estadísticas en Nutrición y Salud

Tratamiento estadístico de outliers y datos faltantes

1. Datos Atípicos o Outliers

Se denominan casos atípicos u outliers a aquellas observaciones con características diferentes de las demás. Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar.

Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos. Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra.

1.1. Tipos de outliers

Los casos atípicos pueden clasificarse en 4 categorías:

- Casos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificarse como datos ausentes.

- Observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento válido de la población y puede ser eliminado del análisis.

- Observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables. Estas observaciones deberían ser retenidas en el análisis pero estudiando qué influencia ejercen en los procesos de estimación de los modelos considerados.

- Datos extraordinarios para las que el investigador no tiene explicación. En estos casos lo mejor que se puede hacer es replicar el análisis con

y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista debería reportarlo en sus conclusiones y debería averiguar el por qué de dichas observaciones.

1.2. Identificación de outliers

Se debe examinar la distribución de observaciones para cada variable, seleccionando como casos atípicos aquellos casos cuyos valores caigan fuera de los rangos de la distribución. La cuestión principal consiste en el establecimiento de un umbral para la designación de caso atípico. Esto se puede hacer gráficamente mediante histogramas o diagramas de caja o bien numéricamente, mediante el cálculo de puntuaciones tipificadas.

Para muestras pequeñas (de 80 o incluso menos observaciones), las pautas sugeridas identifican como atípicos aquellos casos con valores estándar de 2.5 o superiores. Cuando los tamaños muestrales son mayores, las pautas sugieren que el valor umbral sea 3.

2. Datos ausentes o missing

Los datos ausentes son algo habitual, de hecho, rara es la investigación en la que no aparece este tipo de datos. En estos casos la ocupación primaria del investigador debe ser determinar las razones que subyacen en el dato ausente buscando entender el proceso principal de esta ausencia para seleccionar el curso de acción más apropiado. Para ello se debe determinar cuál es el proceso de datos ausentes, entendido como cualquier evento sistemático externo al encuestado (errores en la introducción de datos) o acción por parte del encuestado (tales como rehusar a contestar) que da lugar a la ausencia de datos.

En particular, el investigador debe analizar si existe algún patrón no aleatorio en dicho proceso que pueda sesgar los resultados obtenidos debido a la pérdida de representatividad de la muestra analizada.

2.1. Tipos de valores missing

Existen 2 tipos de valores missing:

- **Datos ausentes prescindibles**: son resultado de procesos que se encuentran bajo el control del investigador y pueden ser identificados

explícitamente. En estos casos no se necesitan soluciones específicas para la ausencia de datos dado que dicha ausencia es inherente a la técnica usada.
Ejemplos de estas situaciones son aquellas observaciones de una población que no están incluidas en la muestra o los llamados datos censurados que son observaciones incompletas como consecuencia del proceso de obtención de datos seguido en el análisis.

- **Datos ausentes no prescindibles**: son resultado de procesos que no se encuentran bajo el control del investigador y/o no pueden ser identificados explícitamente. Ejemplos de estas situaciones son los errores en la entrada de datos, la renuncia del encuestado a responder a ciertas cuestiones o respuestas inaplicables. En estos casos se debe analizar si existen o no patrones sistemáticos en el proceso que puedan sesgar los resultados obtenidos.

Si los datos ausentes son no prescindibles conviene, por lo tanto, analizar el grado de aleatoriedad presente en los mismos. Según este grado el proceso de datos ausentes se puede clasificar del siguiente modo:

1. **Datos ausentes completamente aleatorios (MCAR)**: este es el mayor grado de aleatoriedad y se da cuando los datos ausentes son una muestra aleatoria simple de la muestra sin un proceso subyacente que tiende a sesgar los datos observados. En este caso se podría solucionar el problema sin tener cuenta el impacto de otras variables

2. **Datos ausentes aleatorios (MAR)**: en este caso el patrón de los datos ausentes en una variable Y no es aleatorio sino que depende de otras variables de la muestra X. Ahora bien, para cada valor de X, los valores observados de Y sí representan una muestra aleatoria de Y. Así, por ejemplo, si X es el sexo del encuestado e Y es su renta, un proceso MAR se tendría si existen más valores ausentes de Y en hombres que en mujeres y, sin embargo, los datos son aleatorios para ambos sexos en el sentido de que, tanto en los hombres como en las mujeres, el patrón de ausentes es completamente aleatorio. Si, además, tampoco existen diferencias por sexos los datos ausentes serían MCAR. Si los datos ausentes son MAR cualquier solución al problema deberá tener en cuenta los valores de X dado que afectan al proceso generador de datos ausentes.

3. **Datos ausentes no aleatorios**: en este caso existen patrones sistemáticos en el proceso de datos ausentes y habría que evaluar la magnitud del problema calibrando, en particular, el tamaño de los sesgos

introducidos por dichos patrones. Si éstos son grandes habría que atacar el problema directamente intentando averiguar cuáles son dichos valores.

2.2. Localización de datos missing

El primer paso en el tratamiento de datos ausentes consiste en evaluar la magnitud del problema. Para ello se comienza analizando el porcentaje de datos ausentes por variables y por casos.

Si existen casos con un alto porcentaje de datos ausentes se deberían excluir del problema. Así mismo si existe una variable con un alto porcentaje de este tipo de casos su exclusión dependerá de la importancia teórica de la misma y la posibilidad de ser reemplazada por variables con un contenido informativo similar.

Como regla general, sin embargo, si dicha variable es dependiente debería ser eliminada ya que cualquier proceso de imputación de valores puede distorsionar la significación estadística y práctica de los modelos estimados para ella.

2.3. Diagnóstico de la aleatoriedad de datos missing

Existen 3 métodos:

- Para cada variable Y formar dos grupos (observaciones ausentes y presentes en Y) y aplicar contrastes de comparación de dos muestras para determinar si existen diferencias significativas entre los dos grupos sobre otras variables de interés. Si se encuentran diferencias significativas el proceso de datos ausentes no es aleatorio.

- Utilizar correlaciones dicotomizadas para evaluar la correlación de los datos ausentes en cualquier par de valores. Estas correlaciones indicarían el grado de asociación entre los valores perdidos sobre cada par de variables. Bajas correlaciones implican aleatoriedad en el par de variables y que los datos ausentes pueden clasificarse como MCAR. En caso contrario son MAR.

- Realizar contrastes conjuntos de aleatoriedad que determinen si los datos ausentes pueden ser clasificados como MCAR. Estos contrastes analizan el patrón de datos ausentes sobre todas las variables y las compara con el patrón esperado para un proceso de datos ausentes aleatorio. Si

no se encuentran diferencias significativas el proceso puede clasificarse como MCAR; en caso contrario deben utilizarse los procedimientos a) y b) anteriores para identificar los procesos específicos de datos ausentes que no son aleatorios.

2.4. Aproximaciones al tratamiento de datos ausentes

Si se encuentran procesos de datos ausentes MAR o no aleatorios, el investigador debería aplicar sólo el método diseñado específicamente para este proceso. Sólo si el investigador determina que el proceso de ausencia de datos puede clasificarse como MCAR pueden utilizarse las siguientes aproximaciones:

- Utilizar sólo los casos completos: conveniente si el tamaño muestral no se reduce demasiado
- Supresión de casos y/o variables con una alta proporción de datos ausentes. Esta supresión deberá basarse en consideraciones teóricas y empíricas. En particular, si algún caso tiene un dato ausente en una variable dependiente, habitualmente excluirlo puesto que cualquier proceso de imputación puede distorsionar los modelos estimados. Así mismo una variable independiente con muchos datos ausentes podrá eliminarse si existen otras variables muy similares con datos observados.
- Imputar valores a los datos ausentes utilizando valores válidos de otras variables y/o casos de la muestra:

    1) Métodos de disponibilidad completa que utilizan toda la información disponible a partir de un subconjunto de casos para generalizar sobre la muestra entera. Se utilizan habitualmente para estimar medias, varianzas y correlaciones.
    2) Métodos de sustitución que estiman valores de reemplazo para los datos ausentes, sobre la base de otra información existente en la muestra. Así se podría sustituir observaciones con datos ausentes por observaciones no maestrales o sustituir dichos datos por la media de los valores observados o mediante regresión sobre otras variables muy relacionadas con aquella a la que le faltan observaciones.
    3) Métodos basados en modelos que construyen explícitamente el mecanismo por el que se producen los datos ausentes y lo estiman por máxima verosimilitud. Entran en esta categoría el algoritmo EM o los procesos de aumento de datos.

3. Referencias sobre Análisis de Outliers y Missing

- HAIR, J., ANDERSON, R., TATHAM, R. y BLACK, W. (1999). Análisis Multivariante. 5ª Edición. Prentice Hall.
- LITTLE, R.J.A. and RUBIN, D. (1987) Statistical Analysis with Missing Data. New York. Wiley.
- RIAL, A.; VARELA, J. y ROJAS, A. (2001). Depuración y Análisis Preliminares de Datos en SPSS. Sistemas Informatizados para la Investigación del Comportamiento. RA-MA.