

FASES DE LIMPIEZA Y TRANSFORMACIÓN DE DATOS

LIMPIEZA Y TRANSFORMACIÓN DE DATOS EN EL PROCESO DE EXTRACCIÓN DEL CONOCIMIENTO

Después de la fase de exploración, el proceso de extracción del conocimiento contempla la fase de limpieza de datos (*data cleaning*). La información puede contener valores atípicos, valores faltantes y valores erróneos. En esta fase se analiza la influencia de los datos atípicos, se imputa la información faltante y se eliminan o corrigen los datos incorrectos. La presencia de datos atípicos y valores desaparecidos (*datos missing*) puede llevarnos a usar algoritmos robustos a datos atípicos y desaparecidos (p.ej. árboles de decisión), a filtrar la información, a reemplazar valores mediante *técnicas de imputación* y a transformar datos continuos en discretos mediante *técnicas de discretización*.

A continuación, si es necesario, se lleva a cabo la *transformación* de los datos, generalmente mediante técnicas de reducción o aumento de la dimensión y escalado simple y multidimensional, entre otras. Las cuatro primeras fases estudiadas hasta ahora (selección, exploración, limpieza y transformación) se suelen englobar bajo el nombre de *preparación de datos*. Entre las técnicas avanzadas de transformación tenemos las de reducción y aumento de la dimensión.

VALORES ATÍPICOS (*OUTLIERS*)

Un valor *outlier* o atípico es una puntuación extrema dentro de una variable. Este tipo de valores afecta fuertemente a los análisis en que intervenga la citada variable, sobre todo si trabajamos con muestras pequeñas. Por ejemplo, si estamos trabajando con un modelo de regresión lineal en el que interviene la variable, la distorsión producida normalmente es aumentar de forma "espurera" el grado de relación lineal.

Más concretamente, podemos definir los valores atípicos como observaciones aisladas cuyo comportamiento se diferencia claramente del comportamiento medio de resto de las observaciones. Existe una primera categoría de casos atípicos formada por aquellas observaciones que provienen de un error de procedimiento, como por ejemplo, un error de codificación, error de entrada de datos, etc. Estos datos atípicos, si no se detectan mediante filtrado, deben eliminarse o recodificarse como datos ausentes. Esta categoría de casos atípicos contempla aquellas observaciones que ocurren como consecuencia de un acontecimiento extraordinario existiendo una explicación para su presencia en la muestra. Este tipo de casos atípicos normalmente se retienen en la muestra, salvo que su significancia sea sólo anecdótica. Otra categoría adicional de datos atípicos comprende las observaciones extraordinarias para las que el investigador no tiene explicación. Normalmente estos datos atípicos se eliminan del análisis. Una última categoría de datos atípicos la forman las observaciones que se sitúan fuera del rango ordinario de valores de la variable. Suelen denominarse valores extremos y se eliminan del análisis si se observa que no son elementos significativos para la población. Las propias características del caso atípico, así como los objetivos del análisis que se realizan determinan los casos atípicos a eliminar. No obstante, los casos atípicos deben considerarse en el conjunto de todas las variables consideradas. Por lo tanto, hay que analizarlos desde una perspectiva multivariante. Puede ocurrir que una variable tenga valores extremos eliminables, pero al considerar un número suficiente de otras variables en el análisis, el investigador puede decidir no eliminarlos.

Pueden utilizarse *herramientas de análisis exploratorio de datos* para detectar casos atípicos en un contexto univariante. Por ejemplo, en el gráfico de caja y bigotes los valores atípicos se presentan como puntos aislados en los extremos de los bigotes. Los valores extremos suelen aparecer tachados con una *x*. El software habitual indica el número de observación correspondiente a los valores atípicos. En la Figura 10-1 se muestra el gráfico de caja y bigotes para una variable *V1*. Se observan dos valores atípicos anteriores al bigote izquierdo y otros dos posteriores al bigote derecho. El último de ellos es un valor extremo (aparece tachado).

Gráfico de Caja y Bigotes

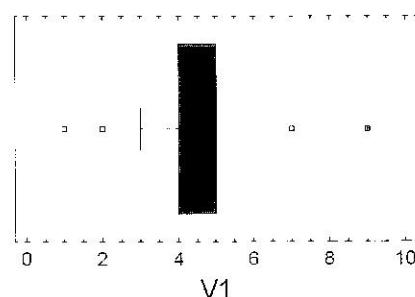


Figura 10-1

Otro camino para detectar valores atípicos consiste en utilizar un *diagrama de control*, consistente en una representación gráfica con una línea central que denota el valor medio de la variable y con otras dos líneas horizontales, llamadas *Límite Superior de Control (LSC)* y *Límite Inferior de Control (LIC)*. Se escogen estos límites de manera que casi la totalidad de los puntos de la variable se halle entre ellos. Mientras los valores de la variable se encuentran entre los límites de control, se considera que no hay valores atípicos. Sin embargo, un punto que se encuentra fuera de los límites de control se interpreta como un valor atípico, y son necesarias acciones de investigación y corrección a fin de encontrar y eliminar la o las causas asignables a este comportamiento. Se acostumbra a unir los diferentes puntos en el diagrama de control mediante segmentos rectilíneos con objeto de visualizar mejor la evolución de la secuencia de los valores de la variable. Sin importar la distribución de la variable, es práctica estándar situar los límites de control como un múltiplo de la desviación típica. Se escoge en general el múltiplo 3, es decir, se acostumbra utilizar los límites de control de tres sigmas en los diagramas de control. A continuación se presenta el gráfico de control tres sigma para una variable con los 25 valores entre 1238 y 1295 siguientes (Figura 10-2). Se constata que la observación número 22 es un valor atípico por caer fuera de los límites de control.

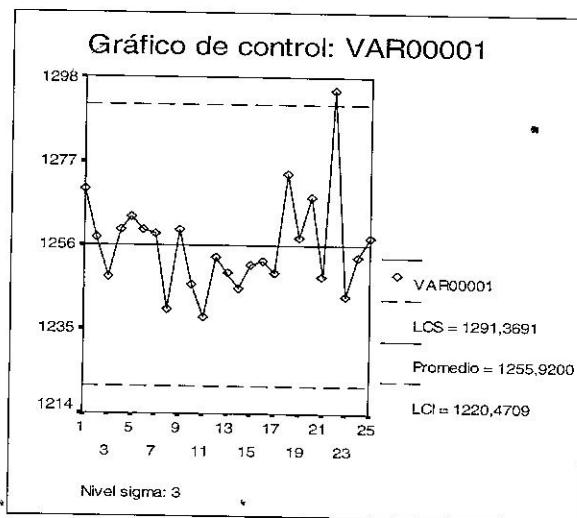


Figura 10-2

Se utilizan límites tres sigma porque la mayoría de las distribuciones con que nos encontramos en la práctica se aproximan a la forma de campana de Gauss (función de densidad de la distribución normal). Como indica la Figura 10-3, la probabilidad de encontrar un valor dentro de $\mu \pm \sigma$ es aproximadamente del 68%. Similarmente, la probabilidad de que los valores caigan fuera de los límites $\mu \pm 2\sigma$ es aproximadamente del 4,5%, mientras que la probabilidad de que los valores caigan fuera de los límites $\mu \pm 3\sigma$ es despreciable (sólo del 0,3% o del tres por mil). Por esta razón se utilizan límites tres sigma.

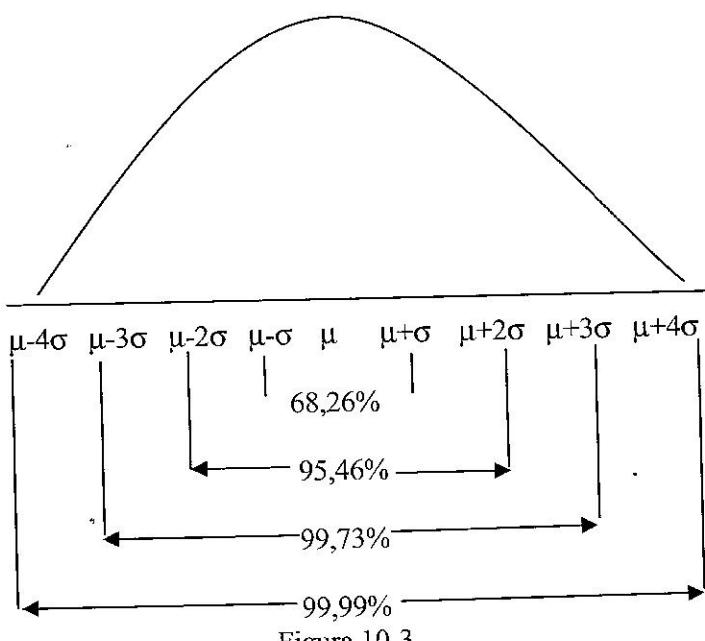


Figura 10-3

También se detectan posibles valores atípicos mediante los estadísticos robustos de la variable y ver su diferencia respecto de los estadísticos no robustos. Suelen considerarse como estadísticos robustos de centralización (localización) la mediana, la media truncada y la media winsorizada. La media truncada prescinde del 15% de los valores de la variable por cada extremo y la media winsorizada sustituye ese 15% de valores por valores del centro de la distribución. Como estadísticos robustos de dispersión (escala) se usan respectivamente la variación media respecto de la mediana, la desviación típica truncada y la desviación típica winsorizada. Cuando no hay valores atípicos, los estadísticos robustos y los estadísticos normales no difieren mucho. También pueden calcularse intervalos de confianza para la media normal y para la media winsorizada. Si su anchura es similar no hay valores atípicos. No obstante, es más efectivo utilizar un contraste formal estadístico para detectar valores atípicos, por ejemplo el test de Dixon o el test de Grubbs, cuyos p -valores detectan los valores atípicos. Para p -valores menores que 0,05 hay valores atípicos al 95% de confianza.

Cuando se trata de *detectar casos atípicos en un contexto bivariante*, pueden utilizarse *herramientas de análisis exploratorio de datos*, por ejemplo, el gráfico de caja y bigotes múltiple (Figura 10-4) que representa distintos gráficos de una variable (potencia de los automóviles) para diferentes niveles de la otra (país de origen). Se observan valores atípicos para los tres orígenes (3 para el origen uno, 4 para el dos y 1 para el tres).

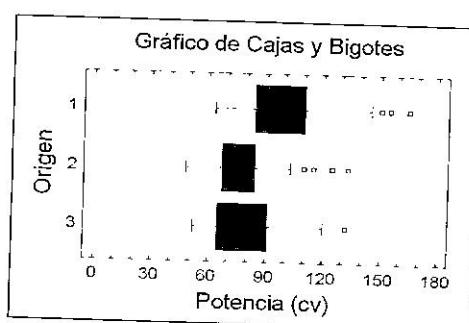


Figura 10-4

Otra forma de detectar casos atípicos en un contexto bivariante consiste en evaluar conjuntamente pares de variables mediante un gráfico de dispersión. En la Figura 10-5, que representa el consumo de los coches en función de su potencia, aparecen 5 valores atípicos por encima de la banda de confianza exterior y 3 por debajo. Casos que caigan manifiestamente fuera del rango del resto de las observaciones pueden identificarse como puntos aislados en el gráfico de dispersión.

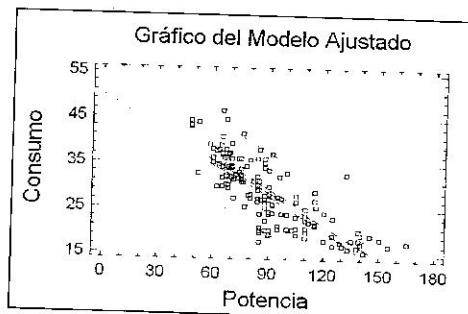


Figura 10-5

Para detectar casos atípicos en un contexto multivariante, pueden utilizarse estadísticos basados en distancias, para detectar los puntos influyentes. La *distancia Mahalanobis* es una medida de la distancia de cada observación en un espacio multidimensional respecto del centro medio de las observaciones. El *estadístico FITS* mide la influencia de cada observación en caso de ser eliminada del análisis. La *Levancia (Leverage)* mide la influencia de cada observación.

INFORMACIÓN FALTANTE (DATOS MISSING)

El tratamiento de la información faltante constituye una de las tareas previas a cualquier análisis. Cuando se aplica un método de análisis multivariante sobre los datos disponibles puede ser que no exista información para determinadas observaciones y variables. Estamos entonces ante valores ausentes o valores *missing*. La presencia de esta información faltante puede deberse a un registro defectuoso de la información, a la naturaleza natural de la información buscada o a una falta de respuesta (total o parcial).

La primera prueba a realizar cuando existen datos *missing* es comprobar si se distribuyen aleatoriamente en todo el conjunto de datos. Es vital que el investigador averigüe si el proceso de ausencia de datos tiene lugar de forma aleatoria. Una primera prueba para valorar los datos ausentes para una única variable Y consiste en formar dos grupos de valores para Y , los que tienen datos ausentes y los que no los tienen. A continuación, para cada variable X distinta de Y , se realiza un test para determinar si existen diferencias significativas entre los dos grupos de valores determinados por la variable Y (ausentes y no ausentes) sobre X . Si vamos considerando como Y cada una de las variables del análisis y repitiendo el proceso anterior se encuentra que todas las diferencias son no significativas, se puede concluir que los datos ausentes obedecen a un *proceso completamente aleatorio*; por tanto pueden realizarse análisis estadísticos fiables con nuestras variables *imputando los datos ausentes* por los métodos que se verán más adelante. Si un porcentaje bastante alto de las diferencias son no significativas, puede considerarse que los datos ausentes obedecen a un *proceso aleatorio* (no completamente aleatorio) que también permitirá realizar análisis estadísticos fiables con nuestras variables previa *imputación de la información faltante*, aunque con menos fiabilidad que en el caso anterior.

También es habitual comprobar la distribución aleatoria de los datos *missing* mediante la *prueba de las correlaciones dicotomizadas*. Para realizar esta prueba, para cada variable Y del análisis se construye una variable dicotomizada asignando el valor cero a los valores ausentes y el valor uno a los valores presentes. A continuación se dicotomizan todas las variables del análisis y se halla su matriz de correlaciones acompañada de los contrastes de significatividad de cada coeficiente de correlación de la matriz. Las correlaciones indican el grado de asociación entre los valores perdidos sobre cada par de variables, con lo que se puede concluir que si los elementos de la matriz de correlaciones son no significativos, los datos ausentes son completamente aleatorios. Si existe alguna correlación significativa y la mayor parte son no significativas, los datos ausentes pueden considerarse aleatorios. Una vez comprobada la aleatoriedad de los datos *missing* en el conjunto total de datos ya es posible imputar la información faltante y realizar análisis estadísticos precisos.

Adicionalmente existen pruebas formales de aleatoriedad de los datos *missing* como el *test conjunto de aleatoriedad de Little*, contraste formal basado en la Chi-cuadrado, cuyo p -valor indica si los valores perdidos constituyen o no un conjunto de números aleatorios.

A continuación se ilustran los conceptos anteriores con un ejemplo basado en los datos recogidos en un cuestionario con 6 preguntas sobre comportamientos y actitudes de compra de 20 encuestados. Las respuestas a las 6 preguntas se recogen en 6 variables ($V1, V2, V3, V4, V5$ y $V6$) cuyo rango varía entre 1 y 10 reflejando la valoración que el encuestado da a la característica que refleja la pregunta.

La primera pregunta valora la importancia que el encuestado da a la impresión que los demás tienen sobre él. La segunda pregunta refleja la valoración que el encuestado da a la garantía de las marcas. La tercera pregunta ofrece información sobre la frecuencia con que el encuestado compra sobre la marcha. La cuarta pregunta mide la preferencia que el encuestado da al comprar sobre ahorrar y vivir mejor. La quinta pregunta mide el gusto del encuestado por vestir a la moda y la sexta pregunta mide la tendencia del encuestado a conocer tiendas nuevas. Los datos de las 6 variables en los 20 cuestionarios se recogen en la tabla de la Figura 10-6.

| Cuestionario | V1 | V2 | V3 | V4 | V5 | V6 |
|--------------|----|----|----|----|----|----|
| 1 | 5 | 6 | 2 | 1 | . | 5 |
| 2 | 7 | . | 4 | 5 | 5 | 7 |
| 3 | . | 1 | 5 | 8 | 5 | 8 |
| 4 | 3 | 5 | 1 | . | 7 | 5 |
| 5 | 5 | 5 | 8 | 3 | 7 | 8 |
| 6 | 5 | 1 | . | 1 | 2 | 8 |
| 7 | 4 | . | 2 | 8 | 9 | 8 |
| 8 | 5 | 1 | 9 | 1 | 1 | 9 |
| 9 | 7 | 5 | 1 | 1 | 1 | . |
| 10 | 2 | 2 | 1 | 4 | 6 | 6 |
| 11 | 9 | 1 | 1 | . | 7 | 5 |
| 12 | 5 | 5 | 8 | 9 | 9 | 5 |
| 13 | . | 9 | 1 | 9 | 7 | 9 |
| 14 | 5 | 6 | 2 | 1 | 1 | 5 |
| 15 | 7 | 7 | 4 | 5 | 4 | 7 |
| 16 | 1 | 1 | 5 | 8 | 5 | . |
| 17 | 3 | 5 | 1 | 7 | . | 5 |
| 18 | 5 | 5 | . | 3 | 7 | 8 |
| 19 | 5 | 1 | 1 | 1 | 2 | 8 |
| 20 | 5 | 1 | 9 | 1 | 1 | 9 |

Figura 10-6

Una vez tabulada la información, la primera tarea sería ver la tabla de frecuencias de los valores perdidos por variables para tener una idea de su magnitud. A continuación se presenta dicha información (Figura 10-7), observándose que para todas las variables el porcentaje de valores perdidos es del 10%, mientras que el de valores válidos es el 90%.

Resumen del procesamiento de los casos

| | Casos | | | | | |
|----|---------|------------|----------|------------|-------|------------|
| | Válidos | | Perdidos | | Total | |
| | N | Porcentaje | N | Porcentaje | N | Porcentaje |
| V1 | 18 | 90,0% | 2 | 10,0% | 20 | 100,0% |
| V2 | 18 | 90,0% | 2 | 10,0% | 20 | 100,0% |
| V3 | 18 | 90,0% | 2 | 10,0% | 20 | 100,0% |
| V4 | 18 | 90,0% | 2 | 10,0% | 20 | 100,0% |
| V5 | 18 | 90,0% | 2 | 10,0% | 20 | 100,0% |
| V6 | 18 | 90,0% | 2 | 10,0% | 20 | 100,0% |

Figura 10-7

El siguiente paso es determinar si los datos ausentes se distribuyen aleatoriamente. Para ello comparamos las observaciones con y sin datos ausentes para cada una de las funciones de las demás variables. La primera tarea será generar nuevas variables $V31$, $V41$, $V51$ y $V61$ (una para cada variable existente) asignándole el valor uno para datos válidos y el valor cero para datos ausentes. Tendremos la tabla de la Figura 10-8.

| Cuest. | V1 | V2 | V3 | V4 | V5 | V6 | V11 | V21 | V31 | V41 | V51 | V61 |
|--------|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|
| 1 | 5 | 6 | 2 | 1 | . | 5 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 7 | . | 4 | 5 | 5 | 7 | 1 | 0 | 1 | 1 | 1 | 1 |
| 3 | . | 1 | 5 | 8 | 5 | 8 | 0 | 1 | 1 | 1 | 1 | 1 |
| 4 | 3 | 5 | 1 | . | 7 | 5 | 1 | 1 | 1 | 1 | 1 | 0 |
| 5 | 5 | 5 | 8 | 3 | 7 | 8 | 1 | 1 | 1 | 1 | 1 | 1 |
| 6 | 5 | 1 | . | 1 | 2 | 8 | 1 | 1 | 0 | 1 | 1 | 1 |
| 7 | 4 | . | 2 | 8 | 9 | 8 | 1 | 0 | 1 | 1 | 1 | 1 |
| 8 | 5 | 1 | 9 | 1 | 1 | 9 | 1 | 1 | 1 | 1 | 1 | 1 |
| 9 | 7 | 5 | 1 | 1 | 1 | . | 1 | 1 | 1 | 1 | 1 | 1 |
| 10 | 2 | 2 | 1 | 4 | 6 | 6 | 1 | 1 | 1 | 1 | 1 | 1 |
| 11 | 9 | 1 | 1 | . | 7 | 5 | 1 | 1 | 1 | 0 | 1 | 1 |
| 12 | 5 | 5 | 8 | 9 | 9 | 5 | 1 | 1 | 1 | 1 | 1 | 1 |
| 13 | . | 9 | 1 | 9 | 7 | 9 | 0 | 1 | 1 | 1 | 1 | 1 |
| 14 | 5 | 6 | 2 | 1 | 1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 |
| 15 | 7 | 7 | 4 | 5 | 4 | 7 | 1 | 1 | 1 | 1 | 1 | 1 |
| 16 | 1 | 1 | 5 | 8 | 5 | . | 1 | 1 | 1 | 1 | 1 | 1 |
| 17 | 3 | 5 | 1 | 7 | . | 5 | 1 | 1 | 1 | 1 | 0 | 1 |
| 18 | 5 | 5 | . | 3 | 7 | 8 | 1 | 1 | 0 | 1 | 1 | 1 |
| 19 | 5 | 1 | 1 | 1 | 2 | 8 | 1 | 1 | 1 | 1 | 1 | 1 |
| 20 | 5 | 1 | 9 | 1 | 1 | 9 | 1 | 1 | 1 | 1 | 1 | 1 |

Figura 10-8

Ahora consideramos los dos grupos formados en la variable $V1$ (valores válidos y valores ausentes) que vienen definidos por la variable $V11$ y hacemos un contraste de igualdad de medias para los dos grupos de valores definidos en cada una de las restantes variables ($V2$ a $V6$) por los valores de $V11$. Tenemos el resultado de la Figura 10-9.

| V1 | Prueba de Levene (para la igualdad de varianzas) | | Prueba T para la igualdad de medias | | | | | | |
|----|--|------|-------------------------------------|-------|---------------|----------------------|-----------------------------|---|--------|
| | F | Sig. | t | gl | Sig. (bilat.) | Diferencia de medias | Error típ. de la diferencia | 95% Intervalo de confianza para la diferencia | |
| V2 | 14,050 | ,002 | ,668 | 14 | ,515 | 1,36 | 2,033 | -3,002 | 5,711 |
| | | | ,335 | 1,048 | ,792 | 1,36 | 4,047 | -44,817 | 47,551 |
| V3 | ,435 | ,520 | -,321 | 14 | ,753 | -,79 | 2,444 | -6,028 | 4,451 |
| | | | -,360 | 1,412 | ,765 | -,79 | 2,182 | -15,118 | 13,551 |
| V4 | 3,168 | ,097 | 2,370 | 14 | ,033 | 4,93 | 2,079 | ,469 | 9,311 |
| | | | 5,412 | 7,787 | ,001 | 4,93 | ,911 | 2,819 | 7,051 |
| V5 | 2,865 | ,113 | ,521 | 14 | ,610 | 1,14 | 2,192 | -3,558 | 5,841 |
| | | | ,894 | 2,595 | ,447 | 1,14 | 1,279 | -3,311 | 5,541 |
| V6 | 4,359 | ,053 | 1,524 | 16 | ,147 | 1,75 | 1,148 | -,684 | 4,111 |
| | | | 2,753 | 2,549 | ,085 | 1,75 | ,636 | -,492 | 3,911 |

Figura 10-9

Se observa que salvo para la variable $V4$, no hay diferencias significativas entre las medias de los dos grupos definidos por los valores ausentes de $V1$ en las variables $V2$, $V3$, $V5$ y $V6$ (los intervalos de confianza contienen el valor cero). El contraste de igualdad de medias se realiza suponiendo varianzas iguales (primera línea de la tabla para cada variable) y desiguales (segunda línea para cada variable).

Ahora consideramos los dos grupos formados en la variable $V2$ (valores válidos y valores ausentes) que vienen definidos por la variable $V21$ y hacemos un contraste de igualdad de medias para los dos grupos de valores definidos en cada una de las restantes variables ($V1$ y $V3$ a $V6$) por los valores de $V21$. Tenemos el resultado de la Figura 10-10.

| V2 | Levene | | Prueba T para la igualdad de medias | | | | | | |
|----|--------|------|-------------------------------------|-------|------------------|----------------------|-----------------------------|---|----------|
| | F | Sig. | t | gl | Sig. (bilateral) | Diferencia de medias | Error típ. de la diferencia | 95% Intervalo de confianza para la diferencia | |
| | | | | | | | | Inferior | Superior |
| V1 | ,290 | ,599 | ,439 | 14 | ,667 | ,57 | 1,300 | -2,218 | 3,360 |
| | | | ,365 | 1,188 | ,769 | ,57 | 1,566 | -13,239 | 14,382 |
| V3 | 3,295 | ,091 | -,321 | 14 | ,753 | -,79 | 2,444 | -6,028 | 4,456 |
| | | | -,587 | 3,067 | ,598 | -,79 | 1,339 | -4,995 | 3,424 |
| V4 | 1,160 | ,300 | 1,121 | 14 | ,281 | 2,64 | 2,358 | -2,414 | 7,700 |
| | | | 1,533 | 1,733 | ,283 | 2,64 | 1,724 | -5,988 | 11,273 |
| V5 | ,309 | ,587 | 1,075 | 14 | ,301 | 2,29 | 2,127 | -2,277 | 6,848 |
| | | | 1,070 | 1,301 | ,444 | 2,29 | 2,137 | -13,729 | 18,300 |
| V6 | 5,873 | ,028 | ,513 | 16 | ,615 | ,63 | 1,219 | -1,959 | 3,209 |
| | | | ,960 | 2,786 | ,413 | ,63 | ,651 | -1,540 | 2,790 |

Figura 10-10

Se observa que para todas las variables, no hay diferencias significativas entre las medias de los dos grupos definidos por los valores ausentes de $V2$ en cada una de ellas (los intervalos de confianza contienen el valor cero). Repitiendo los contrastes de igualdad de medias para los grupos que determinan los valores válidos y ausentes de las variables $V3$, $V4$, $V5$ y $V6$ en el resto de las variables, tenemos los resultados de las Figura 10-11 a 10-14.

| V3 | Levene | | Prueba T para la igualdad de medias | | | | | | |
|----|--------|------|-------------------------------------|--------|------------------|----------------------|-----------------------------|---|----------|
| | F | Sig. | t | gl | Sig. (bilateral) | Diferencia de medias | Error típ. de la diferencia | 95% Intervalo de confianza para la diferencia | |
| | | | | | | | | Inferior | Superior |
| V1 | 1,956 | ,181 | -,085 | 16 | ,933 | -,13 | 1,473 | -3,248 | 2,998 |
| | | | -,246 | 15,000 | ,809 | -,13 | ,507 | -1,206 | ,956 |
| V2 | ,187 | ,671 | ,409 | 16 | ,688 | ,81 | 1,988 | -3,402 | 5,027 |
| | | | ,386 | 1,228 | ,756 | ,81 | 2,106 | -16,631 | 18,256 |
| V4 | 3,604 | ,076 | 1,048 | 16 | ,310 | 2,50 | 2,386 | -2,559 | 7,559 |
| | | | 1,936 | 2,698 | ,158 | 2,50 | 1,291 | -1,882 | 6,882 |
| V5 | ,017 | ,898 | ,143 | 16 | ,888 | ,31 | 2,178 | -4,305 | 4,930 |
| | | | ,120 | 1,169 | ,922 | ,31 | 2,600 | -23,309 | 23,934 |
| V6 | 9,655 | ,007 | -,996 | 16 | ,334 | -1,19 | 1,192 | -3,715 | 1,340 |
| | | | -2,893 | 15,000 | ,011 | -1,19 | ,410 | -2,062 | -,313 |

Figura 10-11

| V4 | Prueba de Levene (varianzas) | | Prueba T para la igualdad de medias | | | | | | |
|----|------------------------------|------|-------------------------------------|--------|------------------|----------------------|-----------------------------|---|--|
| | F | Sig. | t | gl | Sig. (bilateral) | Diferencia de medias | Error típ. de la diferencia | 95% Intervalo de confianza para la diferencia | |
| V1 | 4,819 | ,043 | -,868 | 16 | ,398 | -1,25 | 1,440 | -4,31 | |
| | | | -,413 | 1,038 | ,749 | -1,25 | 3,028 | -36,53 | |
| V2 | ,187 | ,671 | ,409 | 16 | ,688 | ,81 | 1,988 | -3,40 | |
| | | | ,386 | 1,228 | ,756 | ,81 | 2,106 | -16,63 | |
| V3 | 5,206 | ,037 | 1,320 | 16 | ,206 | 2,94 | 2,226 | -1,78 | |
| | | | 3,833 | 15,000 | ,002 | 2,94 | ,766 | 1,3 | |
| V5 | 5,840 | ,028 | -1,197 | 16 | ,249 | -2,50 | 2,088 | -6,92 | |
| | | | -3,478 | 15,000 | ,003 | -2,50 | ,719 | -4,03 | |
| V6 | 6,021 | ,026 | 1,988 | 16 | ,064 | 2,19 | 1,100 | -1,45 | |
| | | | 5,775 | 15,000 | ,000 | 2,19 | ,379 | 1,38 | |

Figura 10-12

| V5 | Prueba de Levene (varianzas) | | Prueba T para la igualdad de medias | | | | | | |
|----|------------------------------|------|-------------------------------------|--------|------------------|----------------------|-----------------------------|---|--|
| | F | Sig. | t | gl | Sig. (bilateral) | Diferencia de medias | Error típ. de la diferencia | 95% Intervalo de confianza para la diferencia | |
| V1 | ,054 | ,819 | ,689 | 16 | ,501 | 1,00 | 1,452 | -2,07 | |
| | | | ,897 | 1,536 | ,488 | 1,00 | 1,114 | -5,49 | |
| V2 | 6,349 | ,023 | -1,034 | 16 | ,317 | -2,00 | 1,935 | -6,10 | |
| | | | -2,405 | 6,336 | ,051 | -2,00 | ,832 | -4,00 | |
| V3 | 3,618 | ,075 | 1,047 | 16 | ,310 | 2,38 | 2,268 | -2,43 | |
| | | | 2,565 | 8,439 | ,032 | 2,38 | ,926 | ,259 | |
| V4 | ,044 | ,837 | ,101 | 16 | ,921 | ,25 | 2,466 | -4,97 | |
| | | | ,080 | 1,148 | ,948 | ,25 | 3,106 | -28,99 | |
| V6 | 6,021 | ,026 | 1,988 | 16 | ,064 | 2,19 | 1,100 | -1,45 | |
| | | | 5,775 | 15,000 | ,000 | 2,19 | ,379 | 1,38 | |

Figura 10-13

| V6 | Prueba de Levene (varianzas) | | Prueba T para la igualdad de medias | | | | | | |
|----|------------------------------|------|-------------------------------------|-------|------------------|----------------------|-----------------------------|---|--|
| | F | Sig. | t | gl | Sig. (bilateral) | Diferencia de medias | Error típ. de la diferencia | 95% Intervalo de confianza para la diferencia | |
| V1 | 4,376 | ,053 | ,689 | 16 | ,501 | 1,00 | 1,452 | -2,07 | |
| | | | ,330 | 1,039 | ,795 | 1,00 | 3,029 | -34,218 | |
| V2 | ,187 | ,671 | ,409 | 16 | ,688 | ,81 | 1,988 | -3,402 | |
| | | | ,386 | 1,228 | ,756 | ,81 | 2,106 | -16,631 | |
| V3 | ,340 | ,568 | ,294 | 16 | ,772 | ,69 | 2,338 | -4,268 | |
| | | | ,320 | 1,329 | ,792 | ,69 | 2,148 | -14,850 | |
| V4 | ,574 | ,460 | -,127 | 16 | ,901 | -,31 | 2,466 | -5,540 | |
| | | | -,087 | 1,103 | ,944 | -,31 | 3,587 | -36,945 | |
| V5 | ,134 | ,719 | ,943 | 16 | ,360 | 2,00 | 2,121 | -2,497 | |
| | | | ,943 | 1,264 | ,491 | 2,00 | 2,121 | -14,691 | |

Figura 10-14

Se observa que para prácticamente todas las variables, no hay diferencias significativas entre las medias de los dos grupos definidos por los valores ausentes de cada una de ellas (los intervalos de confianza contienen el valor cero). Por lo tanto se puede concluir con bastante fiabilidad la distribución aleatoria de los datos perdidos, conclusión que permitirá realizar análisis estadísticos con los datos aplicando distintos métodos de imputación de la información faltante.

Para comprobar la aleatoriedad de los datos ausentes también se puede utilizar la matriz de correlaciones dicotomizadas. Se trata de calcular la matriz de correlaciones de las variables resultantes al sustituir los valores perdidos de las variables iniciales por ceros, y los valores válidos por unos. En nuestro caso se trataría de hallar la matriz de correlaciones de las variables V_{12} a V_{62} . Tenemos los resultados de la Figura 10-15.

| | | V_{11} | V_{21} | V_{31} | V_{41} | V_{51} | V_{61} |
|----------|------------------------|----------|----------|----------|----------|----------|----------|
| V_{11} | Correlación de Pearson | 1 | -,111 | -,111 | -,111 | -,111 | -,111 |
| | Sig. (bilateral) | . | ,641 | ,641 | ,641 | ,641 | ,641 |
| V_{21} | Correlación de Pearson | -,111 | 1 | -,111 | -,111 | -,111 | -,111 |
| | Sig. (bilateral) | ,641 | . | ,641 | ,641 | ,641 | ,641 |
| V_{31} | Correlación de Pearson | -,111 | -,111 | 1 | -,111 | -,111 | -,111 |
| | Sig. (bilateral) | ,641 | ,641 | . | ,641 | ,641 | ,641 |
| V_{41} | Correlación de Pearson | -,111 | -,111 | -,111 | 1 | -,111 | -,111 |
| | Sig. (bilateral) | ,641 | ,641 | ,641 | . | ,641 | ,641 |
| V_{51} | Correlación de Pearson | -,111 | -,111 | -,111 | -,111 | 1 | -,111 |
| | Sig. (bilateral) | ,641 | ,641 | ,641 | ,641 | . | ,641 |
| V_{61} | Correlación de Pearson | -,111 | -,111 | -,111 | -,111 | -,111 | 1 |
| | Sig. (bilateral) | ,641 | ,641 | ,641 | ,641 | ,641 | . |

Figura 10-15

Las correlaciones resultantes entre las variables dicotómicas indican la medida en que los datos ausentes están relacionados entre pares de variables. Las correlaciones bajas indican una baja asociación entre los procesos de ausencia de datos para esas dos variables. En nuestro caso todas las correlaciones son bajas y significativas, lo que corrobora la presencia de aleatoriedad de los datos ausentes.

Soluciones para los datos ausentes: Supresión de datos o imputación de la información faltante

Una vez que se ha contrastado la existencia de aleatoriedad en los datos ausentes ya se puede tomar una decisión para dichos datos antes de comenzar cualquier análisis estadístico con ellos.

Podemos comenzar incluyendo sólo en el análisis las observaciones (casos) con datos completos (filas cuyos valores para todas las variables sean válidos), es decir, cualquier fila que tenga algún dato desaparecido se elimina del conjunto de datos antes de realizar el análisis. Este método se denomina **aproximación de casos completos** o **supresión de casos según lista** y suele ser el método por defecto en la mayoría del *software* estadístico. Este método es apropiado cuando no hay demasiados valores perdidos, porque su supresión provocaría una muestra representativa de la información total. En caso contrario se reduciría mucho el tamaño de la muestra a considerar para el análisis y no sería representativa de la información completa.

Otro método consiste en la **supresión de datos según pareja**, es decir, se trabaja con todos los casos (filas) posibles que tengan valores válidos para cada par de variables que se consideren en el análisis independientemente de lo que ocurra en el resto de las variables. Este método elimina menos información y se utiliza siempre en cualquier análisis bivariante o transformable en bivariante.

Otro método adicional consiste en **suprimir los casos (filas) o variables (columnas)** que peor se comportan respecto a los datos ausentes. Nuevamente es necesario sopesar la cantidad de datos a eliminar. Debe siempre considerarse lo que se gana al eliminar una fuente de datos ausentes y lo que se pierde al no contar con una determinada variable o conjunto de casos en el análisis estadístico.

La alternativa a los métodos de supresión de datos es la **imputación de la información faltante**. La imputación es el proceso de estimación de valores ausentes basado en valores válidos de otras variables o casos de la muestra. A continuación se estudian diferentes métodos de imputación.

Un primer método de imputación no reemplaza los datos ausentes sino que imputa las características de la distribución (por ejemplo, la desviación típica) o las relaciones de todos los valores válidos disponibles (por ejemplo, correlaciones).

El proceso de imputación no consiste en reemplazar los datos ausentes por el resto de los casos, sino en utilizar las características de la distribución o las relaciones de todos los valores válidos posibles, como representantes para toda la muestra entera. Este método se denomina **enfoque de disponibilidad completa**.

Un segundo grupo de métodos de imputación ya son métodos de sustitución de datos ausentes por valores estimados sobre la base de otra información existente en la muestra. Consideraremos en este grupo el método de sustitución del caso, el método de sustitución por la media o la mediana, el método de sustitución por un valor constante, el método de imputación por interpolación lineal, el método de imputación por regresión y el método de imputación múltiple.

En el **método de imputación por sustitución del caso** las observaciones (casos) con datos ausentes se sustituyen con otras observaciones no maestrales. Por ejemplo, en una encuesta sobre hogares a veces se sustituye un hogar de la muestra que no contesta por otro hogar que no está en la muestra y que probablemente contestará. Este método de imputación suele utilizarse cuando existen casos con todas sus observaciones ausentes o con la mayoría de ellas.

En el **método de imputación de sustitución por la media** los datos ausentes se sustituyen por la media de todos los valores válidos de su variable correspondiente. Este método tiene la ventaja de que se implementa fácilmente y proporciona información completa para todos los casos, pero tiene la desventaja de que modifica las correlaciones e invalida las estimaciones de la varianza derivadas de las fórmulas estándar de la varianza para conocer la verdadera varianza de los datos.

Cuando hay valores extremos en las variables, se sustituyen los valores ausentes por la mediana (en vez de por la media), ya que la mediana es un estadístico resumen de los datos más robusto. De esta forma se tiene el **método de imputación de sustitución por la mediana**.

A veces, cuando hay demasiada variabilidad en los datos, suele sustituirse cada valor ausente por la media o mediana de un cierto número de observaciones adyacentes a él. En este tipo de imputación suele incluirse también el **método de imputación por interpolación** en el cual se sustituye cada valor ausente de una variable por el valor resultante de realizar una interpolación con los valores adyacentes.

En el **método de imputación de sustitución por valor constante** los datos ausentes se sustituyen por un valor constante apropiado derivado de fuentes externas o de una investigación previa. En este caso el investigador debe asegurarse de que la sustitución de los valores ausentes por el valor constante proveniente de una fuente externa es más válido que la sustitución por la media (valor generado internamente).

En el **método de imputación por regresión** se utiliza el análisis de la regresión para predecir los valores ausentes de una variable basándose en su relación con otras variables del conjunto de datos a partir de la ecuación de regresión que las une. Como desventaja de este método destacaríamos que refuerza las relaciones ya existentes en los datos de modo que conforme aumente su uso los datos resultantes sean más característicos de la muestra y menos generalizables. Además, con este método se subestima la varianza de la distribución. Y no olvidemos como desventaja que este método supone que la variable con datos ausentes tiene correlaciones estanciales con otras variables.

El **método de imputación múltiple** es una combinación de varios métodos de entre los ya citados.

TRANSFORMACIÓN DE DATOS

Cuando el análisis exploratorio lo indique, los datos originales (no los estandarizados ni los previamente modificados) pueden necesitar ser transformados. Suelen considerarse cuatro tipos de transformaciones:

Transformaciones lógicas: Se unen categorías del campo de definición de las variables para reducir así su amplitud. De esta forma pueden eliminarse categorías sin respuestas. También pueden convertirse variables de intervalo en ordinales o nominales y crear variables ficticias (*dummy*).

Transformaciones lineales: Se obtienen al sumar, restar, multiplicar o dividir las observaciones originales por una constante para mejorar su interpretación. Estas transformaciones no cambian la forma de la distribución, ni las distancias entre los valores ni el orden, y por tanto no provocan cambios considerables en las variables.

Transformaciones algebraicas: Se obtienen al aplicar transformaciones no lineales monotónicas a las observaciones originales (raíz cuadrada, logaritmos, etc.) por una constante para mejorar su interpretación. Estas transformaciones cambian la forma de la distribución al cambiar las distancias entre los valores, pero mantienen el orden.

Transformaciones no lineales no monotónicas: Cambian las distancias y el orden entre los valores. Pueden cambiar demasiado la información original.

Con estas transformaciones se arreglan problemas en los datos. Por ejemplo: una asimetría negativa puede minorarse con una transformación parabólica o cúbica, una asimetría positiva fuerte puede suavizarse mediante una transformación hiperbólica o hiperbólica cuadrática (con signo negativo) y una asimetría positiva débil puede suavizarse mediante una transformación de raíz cuadrada, logarítmica o recíproca de la raíz cuadrada (con signo negativo). La transformación logarítmica puede conseguir estacionalidad en media y en varianza para los datos. Suele elegirse como transformación aquélla que arregla mejor el problema, una vez realizada. Si ninguna arregla el problema, realizamos el análisis sobre los datos originales sin transformar. Combinando transformaciones lineales y algebraicas pueden modificarse los valores extremos de la distribución.

Transponer, fusionar, agregar, segmentar y ordenar archivos

Transponer crea un archivo de datos nuevo en el que se transponen las filas y las columnas del archivo de datos original de manera que los casos (las filas) se convierten en variables, y las variables (las columnas) se convierten en casos.

Normalmente, si el archivo de datos de trabajo contiene una variable de identificación o de nombre con valores únicos, podrá utilizarla como variable de nombre: sus valores se emplearán como nombres de variable en el archivo de datos transpuestado.

La fusión de archivos consiste en la formación de un nuevo archivo *con las mismas variables y casos diferentes*. Se trata de *Añadir casos (Append)* fusionando el archivo de datos de trabajo con otro archivo de datos que contiene las mismas variables pero diferentes casos.

También es posible *fundir archivos con los mismos casos pero variables diferentes*. En este caso es necesario que existan variables clave tanto en el archivo de trabajo como en el archivo externo que se funde con él. Ambos archivos deben estar ordenados según el orden ascendente de las variables clave.

Agregar datos combina grupos de casos en casos de resumen únicos y crea un nuevo archivo de datos agregado. Los casos se agregan en función del valor de una o más variables de agrupación. El nuevo archivo de datos contiene un caso para cada grupo. Por ejemplo, se pueden agregar datos de regiones por estado y crear un nuevo archivo en el que el estado sea la unidad de análisis.

Segmentar un archivo es dividir el archivo de datos en distintos grupos para el análisis basándose en los valores de una o más variables de agrupación. Si selecciona varias variables de agrupación, los casos se agruparán por variable dentro de las categorías de la variable anterior de la lista.

Ponderar casos y categorizar y numerizar variables

Es habitual también utilizar ponderaciones. *Ponderar casos* proporciona a los casos diferentes ponderaciones (mediante una réplica simulada) para el análisis estadístico. Los valores de la variable de ponderación deben indicar el número de observaciones representadas por casos únicos en el archivo de datos. Los casos con valores perdidos, negativos o cero para la variable de ponderación se excluyen del análisis. Los valores fraccionarios son válidos y se usan exactamente donde adquieren sentido y, con mayor probabilidad, donde se tabulan los casos.

Categorizar variables consiste en crear una variable categórica a partir de una variable de escala, es decir, se trata de convertir datos numéricos continuos en un número discreto de categorías. Este procedimiento crea nuevas variables que contienen los datos categóricos. También es posible crear una variable numérica a partir de una categórica asignando valores numéricos a las categorías (*Numerización*).

Pareamiento o matching

Las técnicas de pareamiento o *matching* persiguen la comparabilidad de grupos utilizando características comunes de todos ellos. Aunque los grupos difieran respecto a algunas de sus variables, es posible compararlos mediante un procedimiento de ajuste o estandarización. Este procedimiento consiste en igualar ambos grupos con relación a alguna(s) característica(s), haciéndola homogénea en ambos grupos (como ser por ejemplo sexo, edad, su lugar de vivienda o el número de hijos). Un efecto importante del *matching* es el aumento en la eficiencia del estudio, ya que permite circunscribir la población a estudiar a aquella en la cual la exposición es más representativa. Por ejemplo, en el estudio de accidentes vasculares y uso de anticonceptivos orales, el *matching* por edad podría restringir el ingreso de un control de edad avanzada (65 años), en el cual la probabilidad de exposición a anticonceptivos orales es baja o cero.

Conceptualmente el *matching* corresponde a un procedimiento empleado a priori, en la fase de diseño del estudio. Ocasionalmente se puede efectuar pareamiento a posteriori, cuando el investigador decide parear observaciones una vez recogidos los datos, a partir de un conjunto de individuos controles que previamente no fueron sometidos a *matching*. Sin embargo se prefiere reservar el término *matching* para aquellos casos en que el procedimiento se emplea a priori.

El *matching* se usa también cuando se trabaja con variables confusas de difícil definición o medición, como por ejemplo, las de tipo genético, psicosocial o relacionadas a comportamientos humanos. En estos casos, los investigadores suelen utilizar "pares" de sujetos (hermanos, gemelos, miembros de una familia o grupo social específico), con la finalidad de poder estudiar aisladamente el efecto de la variable de interés habiendo controlado la influencia de las variables sometidas a pareamiento, las que se asumen comunes. Los tipos de variables sometidas a *matching* pueden ser variados, y dependerán lógicamente del problema a investigar.

Existen varias modalidades de pareamiento o *matching*. Dos de las más utilizadas, dependiendo si este procedimiento se aplica colectivamente o a observaciones específicas, son el *matching* de grupos o de frecuencia y el *matching* individual (Figura 10-16).

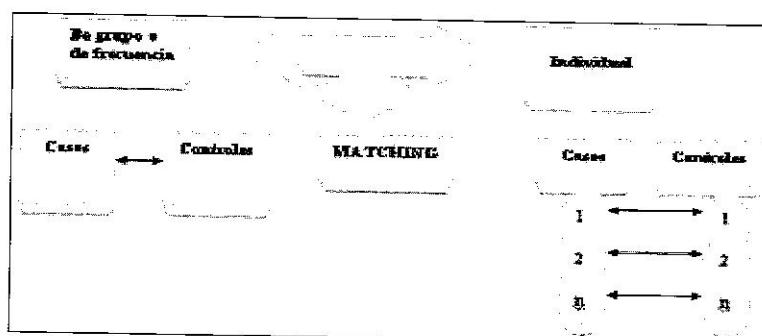


Figura 10-16

En la modalidad de *matching* de grupo o de frecuencia se restringe a priori el ingreso de sujetos en ambos grupos buscando estudiar a sujetos que representen adecuadamente los criterios de inclusión. Así, el ingreso al estudio puede estar regulado por características tales como sexo, grupo de edad, ocupación, lugar de residencia o modalidad de cuidados médicos. La contribución de los grupos en cuanto a eventuales factores confusos tiende a ser homogénea en casos y controles, lo que incrementa la potencia del estudio.

En el *matching* individual, la(s) característica(s) a parear se definen específicamente para cada caso y cada control simultáneamente. Se podrá apreciar que el efecto de este procedimiento tiene implicancias directas en la modalidad de análisis de la información: en este caso el análisis se efectúa por "pares" o "tríos" de observaciones, a diferencia de la modalidad de *matching* por grupos o de frecuencia, en la que se comparan grupos. También tiene implicaciones en la factibilidad de encontrar adecuados sujetos controles que ajusten a los requerimientos exigidos en el *matching*. A mayor cantidad de variables a "parear", mayor dificultad de encontrar controles adecuados. En ambos casos, el *matching* puede considerar más de un control por cada caso.

El *matching* o pareamiento también presenta desventajas. Este procedimiento involucra dificultades técnicas y teóricas en el desarrollo del estudio. El investigador se expone a encontrar dificultades para encontrar controles adecuados y en muchos casos debe descartar controles con el consiguiente riesgo de sesgar las mediciones en el caso de que la(s) variable(s) a parear no sean de valor epidemiológico. El investigador puede verse enfrentado a la realidad de encontrar en su base de una alta frecuencia de valores *missing*, debiendo descartar dichas observaciones o aplicar procedimientos de estimación de ellos usando procedimientos de poca aceptación epidemiológica. El estudio se hace también más largo y por ende, de mayor costo. El término de *overmatching* o *matching* innecesario (sobrepareamiento) se refiere al uso de esta técnica incluyendo innecesariamente variables que pueden no ser necesariamente variables confusas.

TRANSFORMACIÓN DE DATOS MEDIANTE TÉCNICAS DE REDUCCIÓN DE LA DIMENSIÓN

En el mundo de la información de hoy en día es habitual disponer de gran cantidad de variables medidas u observadas en una colección de individuos y pretender estudiarlas conjuntamente. Al observar muchas variables sobre una muestra es presumible que una parte de la información recogida pueda ser redundante o que sea excesiva, en cuyo caso los **métodos multivariantes de reducción de la dimensión** (análisis en componentes principales, factorial, escalamiento óptimo, etc.) tratan de eliminarla. Estos métodos combinan muchas variables observadas para obtener pocas variables ficticias que las representen con la mínima pérdida de información.