Transformaciones de variables

Introducción

La tipificación de variables resulta muy útil para eliminar su dependencia respecto a las unidades de medida empleadas. En realidad, una tipificación equivale a una transformación lineal

$$Z = \frac{X - \bar{x}}{\sigma} = \frac{1}{\sigma}X - \frac{\bar{x}}{\sigma}$$

siendo $Z = aX + b$ donde $a = \frac{1}{\sigma}$ y $b = \frac{\bar{x}}{\sigma}$.

La variable tipificada expresa el número de desviaciones típicas que dista de la media cada observación. Por ello, se puede comparar la posición relativa de los datos de diferentes distribuciones.

Otra situación habitual se presenta cuando se hace un cambio de unidades de medida.

A pesar de las buenas propiedades de las transformaciones lineales, éstas no son suficientes para modificar rasgos más complejos de una distribución como por ejemplo la asimetría.

Para hacer más simétrica una distribución se deben hacer transformaciones no lineales.

Transformaciones no lineales

Supongamos que se trata de estudiar el crecimiento del consumo de energía en diferentes países.

Una opción consiste en en estudiar las diferencias de consumo entre dos instantes de tiempos $C_t - C_{t-1}$, pero en general resulta más conveniente considerar las diferencias relativas: $(C_t - C_{t-1})/C_{t-1}$ o bien $(C_t - C_{t-1})/C_t$.

Una medida más adecuada consiste en tomar logaritmos

$$\ln C_t - \ln C_{t-1} = \ln \frac{C_t}{C_{t-1}} = \ln \left(1 + \frac{C_t - C_{t-1}}{C_{t-1}}\right) \approx \frac{C_t - C_{t-1}}{C_{t-1}}$$

(ya que $\ln(1 + x) \approx x$, para valores de x pequeños).

Así, si se expresa la variable en logaritmos, su crecimiento en dicha escala es una buena medida del crecimiento relativo.

Por otro lado, dado que $C_t \geq C_{t-1}$, entonces

$$\frac{C_t - C_{t-1}}{C_t} \leq \ln \frac{C_t}{C_{t-1}} \leq \frac{C_t - C_{t-1}}{C_{t-1}}$$

de modo que las diferencias de las variables transformadas por un logaritmo, son una medida promedio de las dos formas posibles de medir el crecimiento relativo.

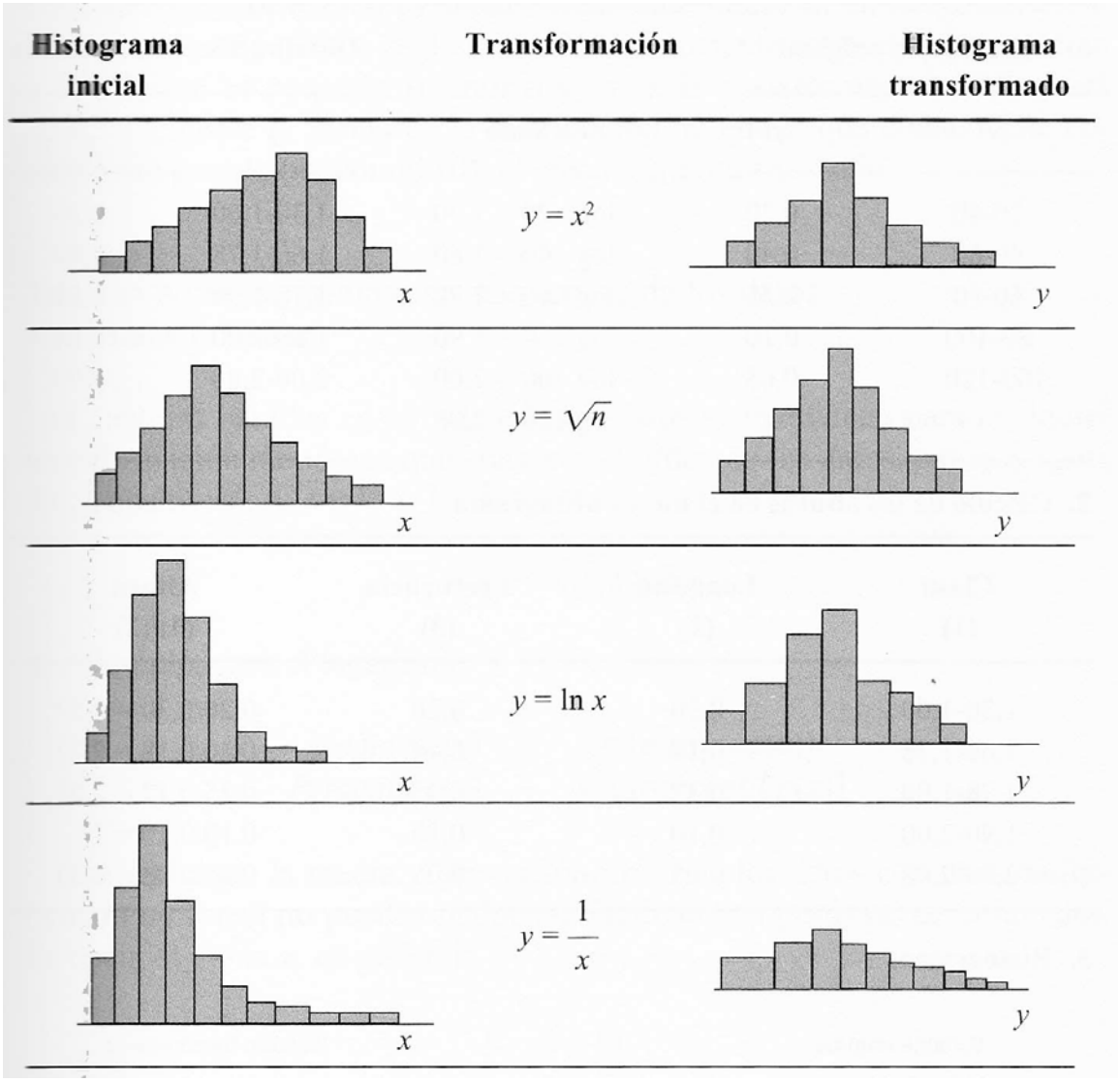
Como regla general, se trata de escoger una transformación que conduzca a una distribución simétrica, y más cercana a la distribución normal. De este modo, se pueden aplicar numerosas técnicas de inferencia estadística.

En una distribución simétrica unimodal, la media, moda y mediana coinciden; además, el coeficiente de asimetría es cero (así como todos los *momentos* de orden impar).

Transformaciones no lineales más frecuentes

Cuando se tienen distribuciones de frecuencias con asimetría negativa (frecuencias altas hacia el lado derecho de la distribución), es conveniente aplicar la transformación $y = x^2$. Esta transformación comprime la escala para valores pequeños y la expande para valores altos.

Para distribuciones asimétricas positivas se usan las transformaciones $\sqrt{x}$, $\ln(x)$ y $1/x$, que comprimen los valores altos y expanden los pequeños. El efecto de estas transformaciones está en orden creciente: menos efecto $\sqrt{x}$, más $\ln(x)$ y más aún $1/x$.



La transformación más utilizada es la del logaritmo. Muchas distribuciones de datos económicos, o de consumos se convierten en simétricas al tomar la transformación logaritmo.

Las medidas basadas en el orden de los datos, como la mediana o los cuartiles se mantienen iguales cuando se hace una transformación monótona, h, del estilo de las previamente citadas:

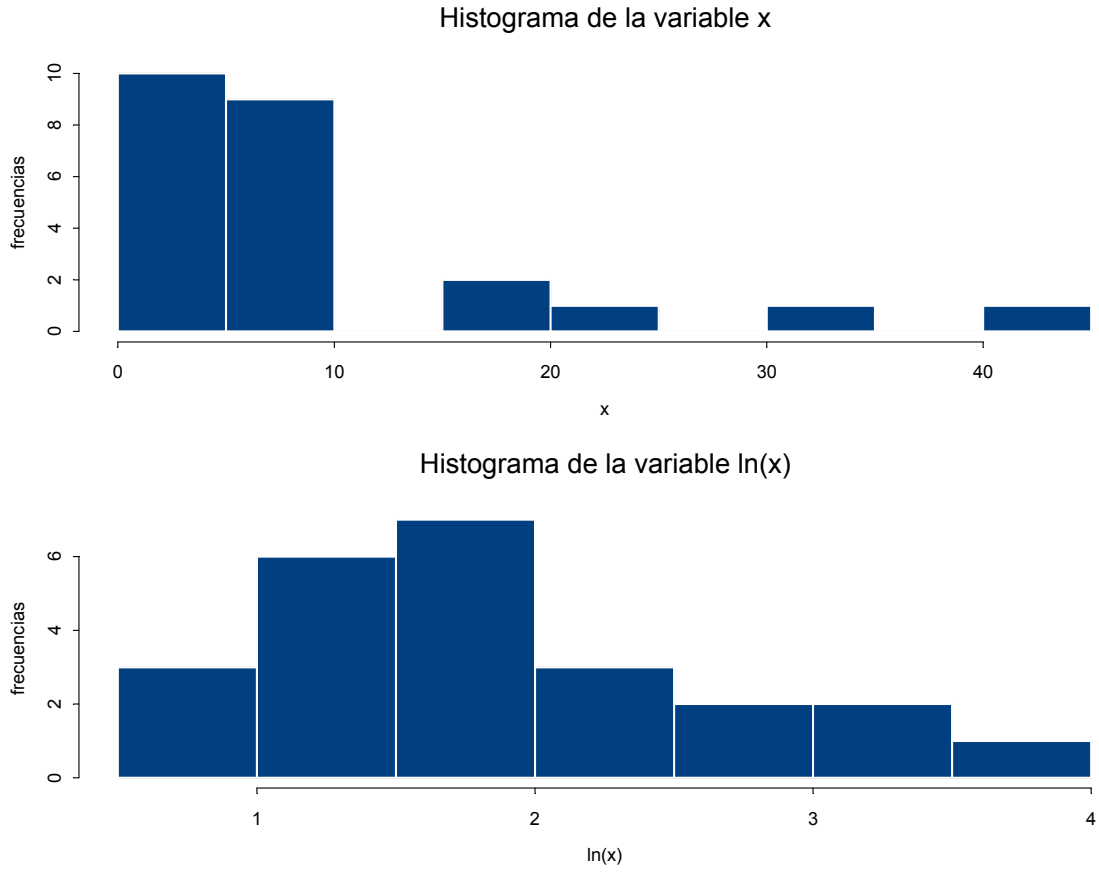
$$x_1 > x_2 \Rightarrow h(x_1) > h(x_2).$$

El resto de estadísticos cambia.

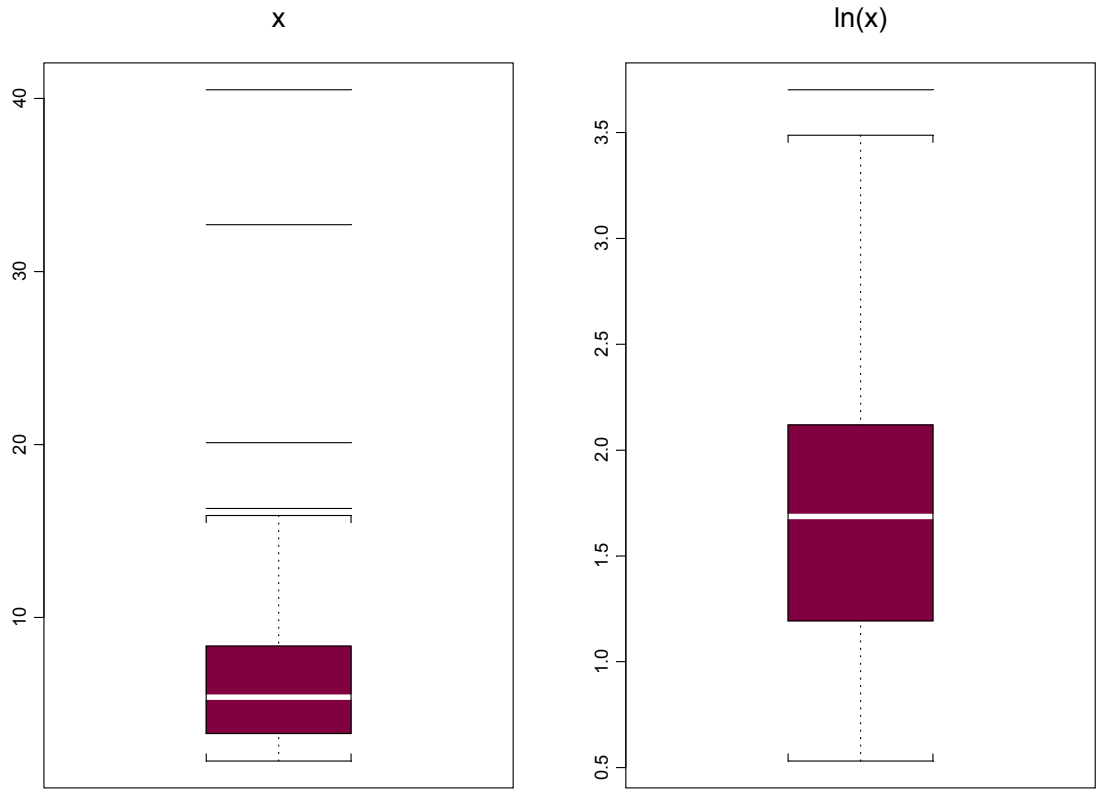
Ejemplo: Se consideran los siguientes datos, correspondientes a la tasa de incrementos de precios al consumo, en 1985, para 25 paises de la *OCDE*:

X =(2.2, 7.6, 2.9, 4.6, 4.1, 3.9, 7.4, 3.2, 5.1, 5.3, 20.1, 2.3, 5.5, 32.7, 9.1, 1.7, 3.2, 5.8, 16.3, 15.9, 5.9, 6.7, 3.4, 40.5).

Si se dibuja el histograma, se observa que la distribución es muy asimétrica: la mayor parte de los países tienen un incremento menor que 10 y unos pocos un incremento mucho mayor. Si se toma la transformación logaritmo, se obtiene una distribución simétrica de los datos.



Respectivamente, si se dibuja el diagrama de cajas, se obtienen numerosos datos atípicos con los datos originales. Si se considera la transformación logaritmo, los atípicos desaparecen.



Si en vez del conjunto de observaciones originales, se tiene sólo la distribución de frecuencias en una tabla, se puede realizar la transformación modificando los extremos de las clases mediante la función elegida. En general, esto hace que cambien las longitudes de las clases. Si a continuación se dibuja el histograma con las nuevas clases, hay que recalcular las alturas, ya que los histogramas representan las frecuencias mediante áreas.