Transformación de variables numéricas

Dr. Gaddiel Desirena López

Contenido

Transformación

- Es una función que mapea los elementos del conjunto X al conjunto Y.
- En términos estadísticos, se usan para estabilizar la varianza.
- Una vez trabajados los datos, se realiza la transformación inversa para poder interpretar correctamente los datos.

Función logaritmo

- Una forma de visualizar la función logaritmo es como el inverso de la exponencial, es decir

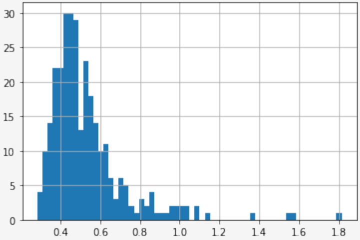
$$a^y = x \quad \leftrightarrow \quad \log_a(x) = y.$$

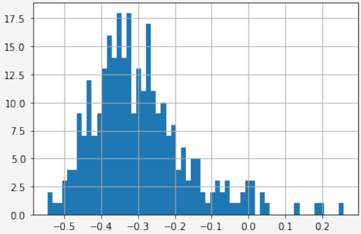
- Comprime el rango numérico alto y expande el rango bajo.
- Los valores pequeños, entre (0,1) los mapea al intervalo ($-\infty$,0).

x	$\log_{10}(x)$
$[1, 10]$	$[0, 1]$
$[10, 100]$	$[1, 2]$
$\dots$	

Transformación logaritmo y recíproco

- Es útil para lidiar con números positivos con una distribución de cola pesada.





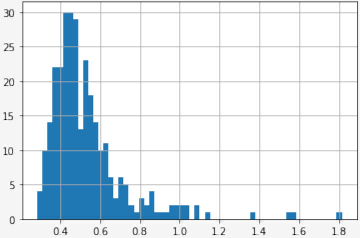
Función recíproco

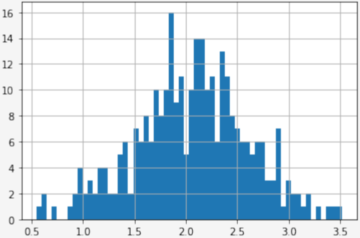
Otra transformación útil para variables con datos con un sesgo muy grande es la función recíproca

$$y = 1/x$$

- Es más potente que la función logarítmica.
- Tiene validez para número negativos.

- Al usar esta transformación con el pasado ejemplo, se nota una mayor simetría.





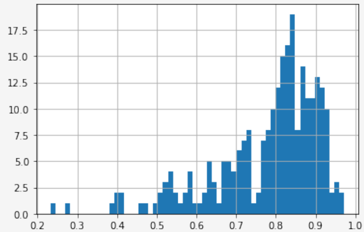
Función cuadrática

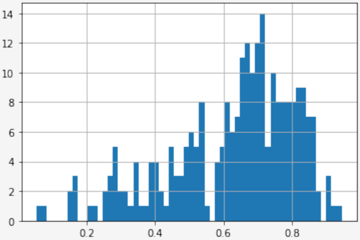
Otra opción si se quiere compensar el sesgo positivo (cola a la derecha) es una transformación de la forma $y = \sqrt{x}$, ya que también amplifica los valores menores que uno y atenúa los mayores a éste. Ahora si se tienen datos cuya distribución presenta una cola cargada al lado izquierdo, se desea hacer lo contrario, y la forma inversa de la raíz cuadrada es la función cuadrática

$$y = x^2.$$

Transformación cuadrática y cúbica

- Los valores pequeños, entre (-1,1) son atenuados, los valores mayores que uno, en magnitud, se amplifican.
- Es útil para lidiar con valores que presentan un sesgo negativo (cola a la izquierda).





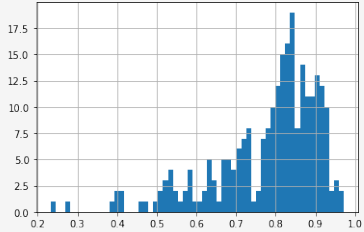
Función cúbica

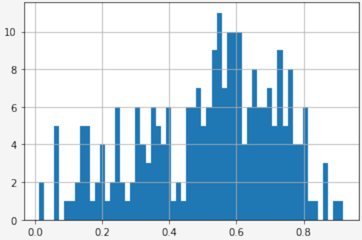
Una transformación más de potencia. Tiene el mismo efecto que la función cuadrática: para variables con densidad de distribución que presentan una cola más alargada del lado izquierdo, las distribuye de forma que los valores pequeños sean más frecuentes. La transformación cúbica se expresa como:

$$y = x^3.$$

Transformación cuadrática y cúbica

- Es más potente que la transformación cuadrática.
- Mantiene el signo de los valores negativos.





Transformación Box–Cox

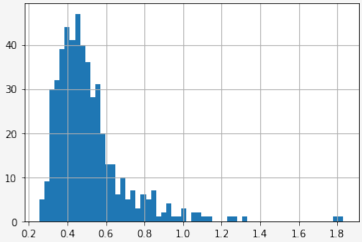
Una generalización simple de las transformaciones de potencia (incluyendo la transformación logarítmo) es la transformación Box–Cox

$$y = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}$$

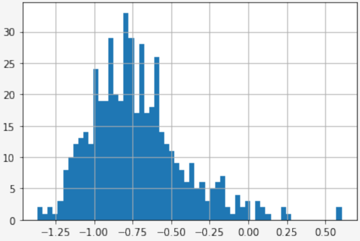
- Se define únicamente para valores de x positivos.
- Es una transformación flexible

λ	Comportamiento
0	$\ln(x)$
$0 < \lambda < 1$	$y = \sqrt{x}$, $y = \sqrt[3]{x}$, etc.
$\lambda < 0$	$y = 1/x^2$, $y = 1/\sqrt{x}$, etc.
$\lambda > 1$	x^2, x^3, etc.

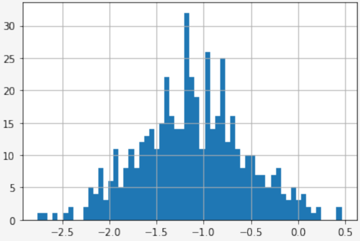
Transformación Box–Cox



Datos originales.



Cox-Box con $\lambda = 0.0$.



Cox-Box con $\lambda = -0.9328$.

Transformación Yeo–Johnson

Es una extensión de la transformación Box–Cox para variables con valores negativos

- Siempre se puede desplazar la variable de forma que presente solo valores positivos para usar Box–Cox.
- La transformación se define como

$$y = \begin{cases} \dfrac{(x+1)^{\lambda}-1}{\lambda} & x \geq 0,\ \lambda \neq 0 \\ \ln(x+1) & x \geq 0,\ \lambda = 0 \\ -\dfrac{(-x+1)^{2-\lambda}-1}{2-\lambda} & x < 0,\ \lambda \neq 2 \\ -\ln(-x+1) & x < 0,\ \lambda = 2 \end{cases}$$

- El primer criterio para identificar la función a ser ejecutada es el signo del elemento x después se usa λ bajo el mismo criterio que Box–Cox para estabilizar la varianza.

Transformación Yeo–Johnson

La función anterior tiene las siguientes propiedades

- Es continua entre valores positivos y negativos.
- La función que se ejecuta cambia dependiendo del signo de x.