

Transformaciones de datos numéricos

**Variable de Salida** : Variable Dependiente  
y  
**Variables de Entrada** : Variables Independientes

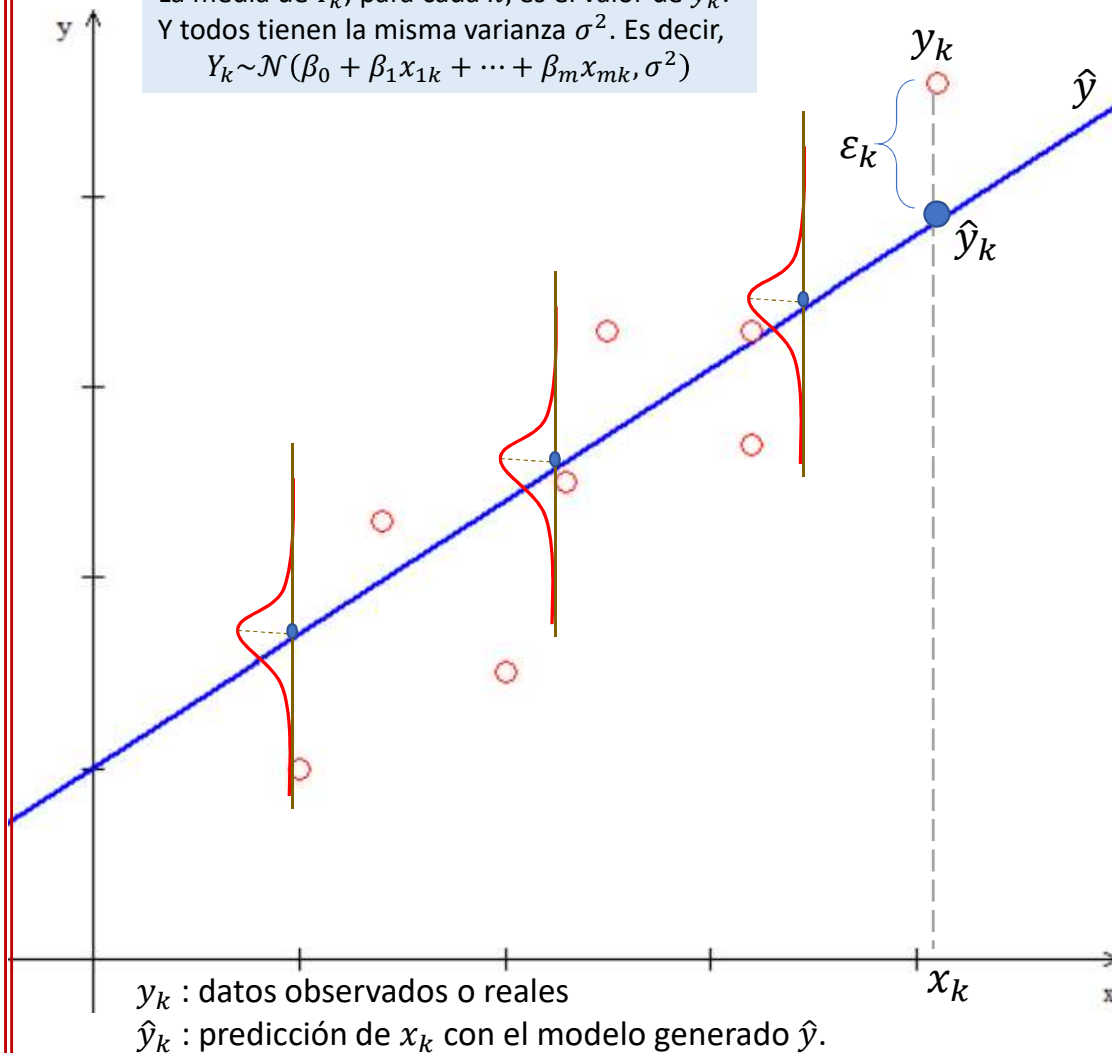
Luis Eduardo Falcón Morales  
ITESM – Campus Guadalajara

## **Nota Aclaratoria:**

Las siguientes diapositivas hablan sobre transformaciones y análisis en datos numéricos, de preferencia continuos o reales. Sin embargo, es usual en ocasiones aplicar estas transformaciones para datos numéricos discretos o enteros, con sus restricciones o consideraciones respectivas.

Los datos categóricos o cualitativos tienen sus propias transformaciones, por lo que no deben considerarse dentro de los casos que estudiaremos a continuación.

La media de  $Y_k$ , para cada  $k$ , es el valor de  $\hat{y}_k$ .  
Y todos tienen la misma varianza  $\sigma^2$ . Es decir,  
 $Y_k \sim \mathcal{N}(\beta_0 + \beta_1 x_{1k} + \dots + \beta_m x_{mk}, \sigma^2)$

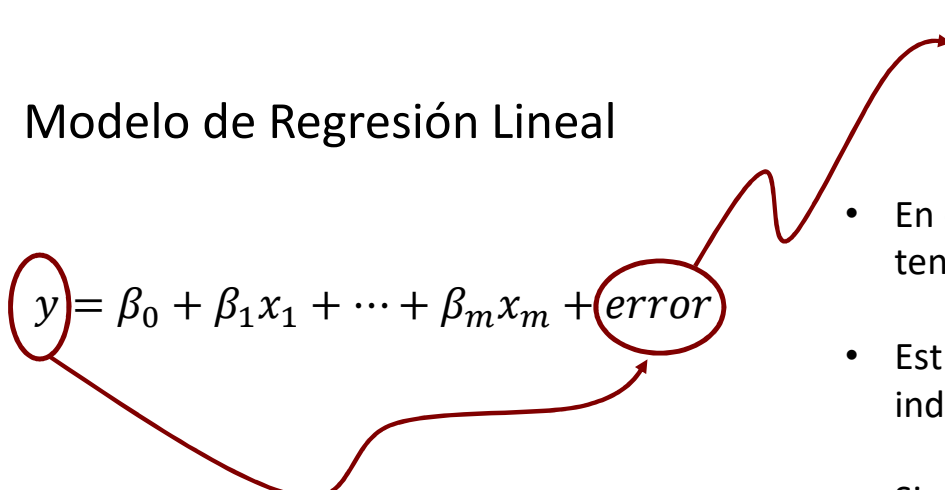


Para aplicar el modelo de regresión lineal entre dos variables aleatorias  $X$  y  $Y$ , se deben cumplir los siguientes supuestos:

- **Linealidad:** La relación entre ambas variables debe ser lineal.
- **Normalidad:** Los valores de  $Y$  están distribuidos normalmente para cada valor de  $X$ .
- **Homocedasticidad:** La variación de los errores alrededor de la línea de regresión tiene un mismo valor constante  $\sigma^2$ .
- **Independencia de los residuos:** Los residuos o errores  $\varepsilon = y - \hat{y}$  deben ser independientes para cada valor de  $X$ .

## Escalamiento de la Variable Independiente

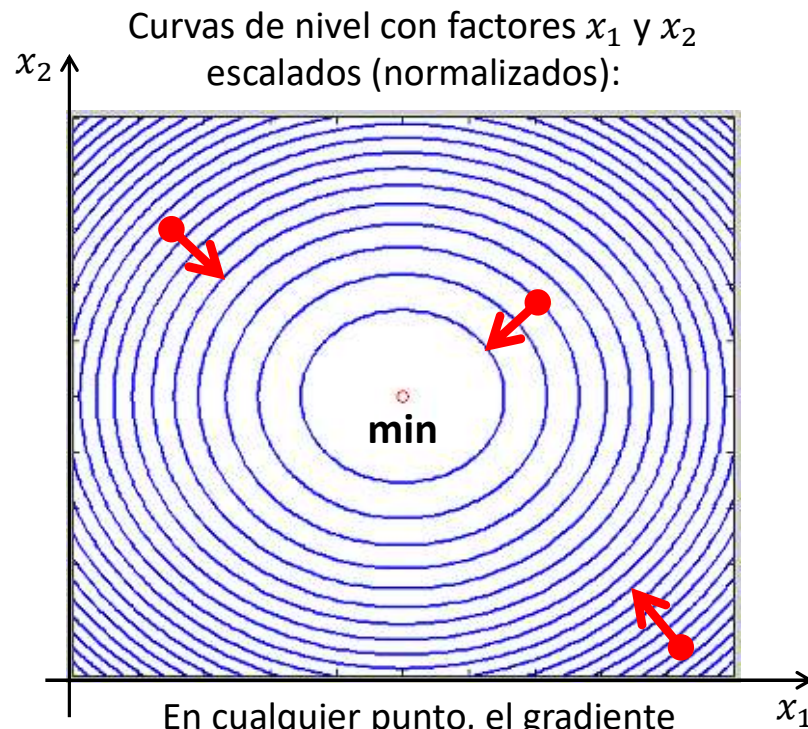
### Modelo de Regresión Lineal

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m + \text{error}$$


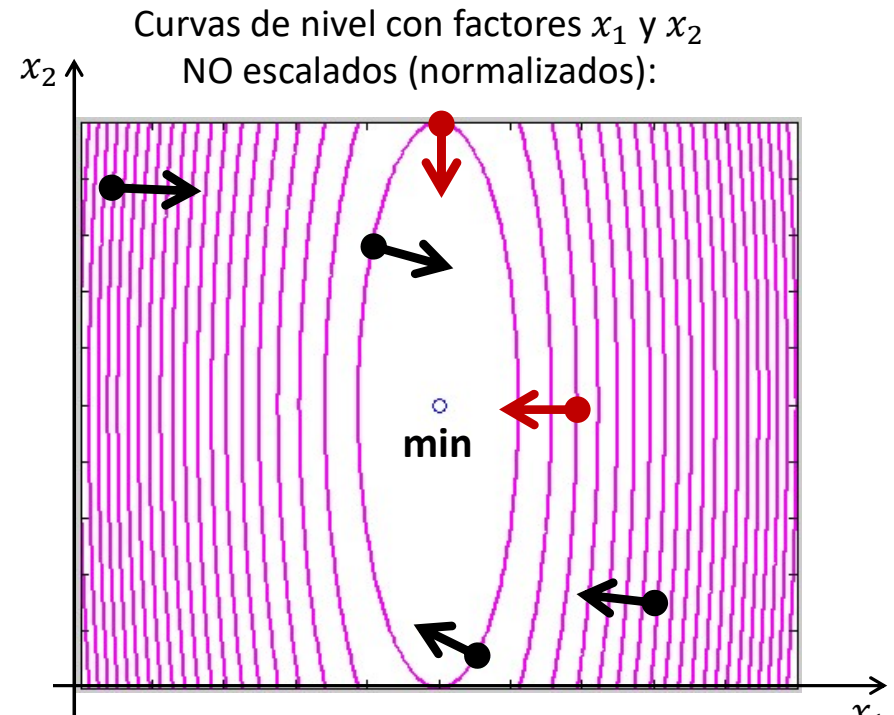
Distribución gaussiana de los errores:  
Modelo Lineal Estándar

- En este caso el modelo requiere que los residuos  $\varepsilon_k = y_k - \hat{y}_k$  tengan una distribución gaussiana con media 0 y varianza 1.
- Estrictamente esta condición gaussiana no aplica a las variables independientes.
- Sin embargo, como veremos a continuación, a las variables independientes también se les suelen aplicar algunas transformaciones para identificar *outliers* o ayudar en la convergencia numérica de métodos iterativos como el gradiente descendente, en varios de los modelos de Aprendizaje Automático (AA). Sin embargo, no aplica en todos, como el caso de árboles de decisión, donde el escalamiento de las variables independientes no es requisito; pero tampoco afecta, por lo que en ocasiones se aplica al comparar los datos con varios modelos de AA.

## Escalamiento o Normalización de Variables Independientes (Factores)



En cualquier punto, el gradiente (ortogonal a las curvas de nivel) ya apunta *casi* al valor mínimo buscado, por lo que la convergencia es más rápida.



En este caso, *casi* ningún gradiente apuntará cerca del óptimo buscado, lo cual implica una mayor cantidad de iteraciones y una convergencia más lenta. De hecho, solo en los vértices de las curvas de nivel (en el caso elíptico) el gradiente apunta al mínimo.

## Transformaciones con respecto a la media

### Transformación de escalamiento

Sea  $X$  un conjunto de datos de entrada. Cada factor o vector columna de  $X$  se puede transformar o escalar como sigue (*feature scaling*):

$$x_k \leftarrow \frac{x_k - \bar{x}}{\text{Rango}}$$

donde

$\bar{x}$  : es el promedio de los datos del factor (o columna) correspondiente.

$\text{Rango} = \text{max} - \text{min}$

$\text{max}$  : valor máximo del factor (columna) correspondiente.

$\text{min}$  : valor mínimo del factor (columna) correspondiente.

### Transformación gaussiana o estándar (StandardScaler)

Otro tipo de escalamiento, usualmente llamada **gaussiana** o **estandarización** se define como sigue:

$$x_k \leftarrow \frac{x_k - \bar{x}}{\text{std}}$$

donde

$\text{std}$ : es la desviación estándar de los datos del factor correspondiente.

Los nuevos datos con esta transformación gaussiana tendrán media 0 y desviación estándar 1.

Los algoritmos de tipo regresión se benefician en general de esta transformación, sobre todo si no son muchos datos los que se tienen.

El uso de la media no permite aminorar el efecto negativo que puedan generar la presencia de valores extremos (outliers).

### Transformación min-max:

O también, con una transformación al intervalo  $[0, 1]$  de la forma siguiente:

$$x_k \leftarrow \frac{x_k - \min}{\text{Rango}}$$

donde

$$\text{Rango} = \max - \min$$

Algunos comentarios en relación a esta transformación:

- También es llamada transformación de escalamiento min-max (**Min-Max-Scaler**).
- No ayuda a reducir el efecto de los outliers o a que la gráfica sea más acampanada, pero todas las variables transformadas de esta manera estarán ahora en una escala en que podrán ser más competitivas.
- Los nuevos datos quedan en el intervalo  $[0, 1]$ .
- Es una buena transformación que se puede usar para empezar a mejorar el análisis.
- Es la que menos distorsiona los datos originales y hace lo mínimo para que sean competitivas las variables entre sí.

## Transformación robusta

Una transformación basada en el rango intercuartil IQR se define como sigue:

$$x_k \leftarrow \frac{x_k - \tilde{x}}{IQR}$$

donde

$\tilde{x}$  : es la mediana de los datos del factor (o columna) correspondiente.

Rango Intercuartil :  $IQR = Q_3 - Q_1$

$Q_3$ : tercer cuartil del factor (columna) correspondiente.

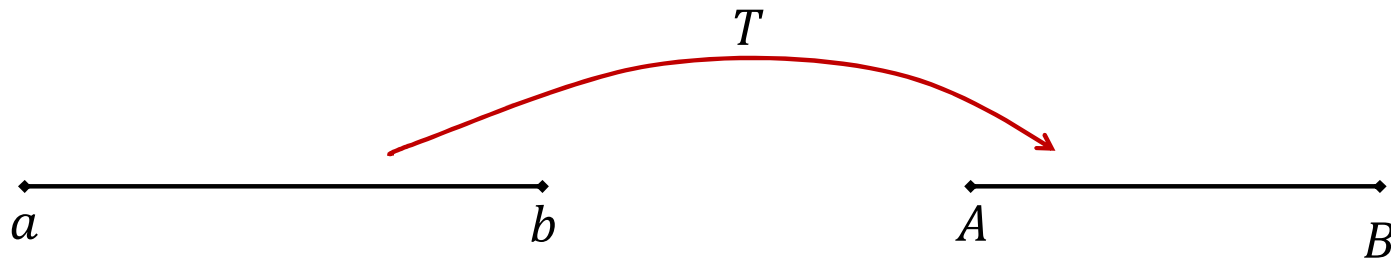
$Q_1$ : primer cuartil del factor (columna) correspondiente.

- Suele llamársele transformación robusta (**Robust-Scaler**), aunque dicho nombre no está del todo generalizado.
- Los datos quedan acotados alrededor del cero.
- El uso de la mediana y el rango intercuartil ayuda a reducir el efecto de outliers sobre todo con datos futuros que contengan valores extremos; pero no significa que ya no existirán los outliers.



## Otro tipo de Escalamiento

Cada factor también puede ser escalado mediante la transformación lineal siguiente:



$$T(x) = \frac{B - A}{b - a} (x - a) + A$$

Esta función transforma los datos del intervalo  $[a, b]$  original, al intervalo  $[A, B]$  deseado.

### **NOTA ACLARATORIA:**

- Aunque las anteriores definiciones de las transformaciones son usualmente las maneras más genéricas en las que usualmente se definen, cada autor y más aún las librerías y paquetes, pueden introducir sus propias definiciones particulares que resulten en valores diferentes al momento de compararlas. Sin embargo, se espera que en general se apliquen y utilicen con propósitos similares.
- Igualmente existen otra gran cantidad de transformaciones que pueden ser útiles en muchos casos particulares. No dudes en seguir investigando y probando otras de ellas, con el tiempo irás teniendo una cartera de ellas que te permitan ir enfrentando mejor los problemas futuros.

## Valores Extremos (outliers)

- Los valores extremos o outliers en general plantean problemas no fáciles de resolver al analista de datos.
- No existe en la literatura un consenso sobre cómo enfrentarlos, aunque existe mucha investigación y técnicas al respecto.
- En ocasiones se usan algunas de las transformaciones anteriores (por ejemplo, la transformación robusta) para aminorar un poco su impacto, pero no siempre con buenos resultados.
- Existen una buena cantidad de estudios y análisis sobre dicho tema que quedan fuera de los alcances de este curso.
- Por el momento mencionamos solamente que es usual que en ocasiones se use algún tipo de transformación o técnica que ayude a aminorar el impacto de dichos datos en el modelo generado.
- A continuación veamos algunas recomendaciones que puedes empezar a considerar.

## Tipo de Valores Extremos (outliers):

En general podemos tener tres tipos de outliers:

- Datos que son parte de la distribución, pero que son parte del 5%, aproximadamente, de datos que están en los extremos (colas).
- **Anomalía/Error:** un registro que fue mal capturado o registrado y que no pertenece a la distribución de los datos principales. **Está bien borrar estos datos ANTES de cualquier transformación.**
- Datos que están muy alejados de la media o mediana y no eres capaz de determinar si son parte o no de la distribución o se deben a un error.

No existe una decisión general entre la comunidad científica sobre cómo proceder con los outliers y actualmente sigue habiendo mucha investigación al respecto.

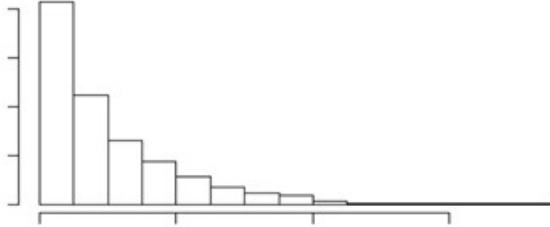
Sin embargo, enunciaremos algunas sugerencias en la siguiente diapositiva.

Toma en cuenta que estas son simplemente recomendaciones, pero la decisión final la debe tomar el analista con su equipo de trabajo, quienes son los que conocen mejor el problema y los datos con los cuales están trabajando.

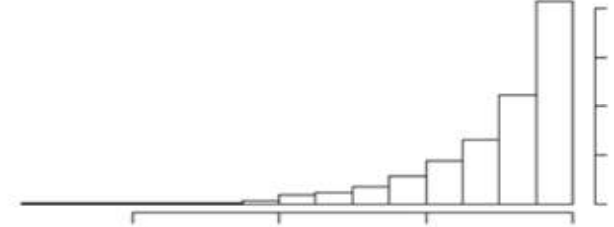
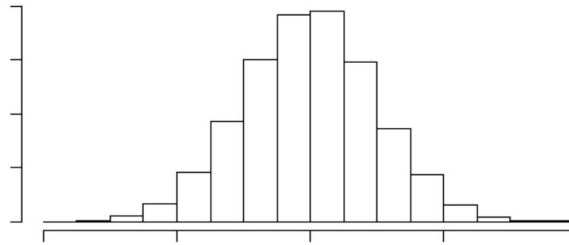
## ¿PROCESAR o BORRAR LOS VALORES EXTREMOS (OUTLIERS)?

- Si identificas que los outliers se deben a errores de captura, de medición, que tienen información imprecisa o bien que tienen poco peso en el análisis, la decisión de borrarlos antes de transformarlos se considera una decisión adecuada.
- Si los outliers son casos “muy raros” pero aún pueden considerarse dentro de la distribución del resto de los datos, una recomendación sería borrar dichos outliers después de su transformación; pero en este caso el analista debería seguir preguntándose si debieron ser borrados. En dado caso, otra recomendación sería generar dos modelos, uno con outliers y el otro sin outliers y comparar cuál pudiera generar mejores predicciones.
- Si el conjunto de outliers es “relativamente grande” (aunque sabemos que deben andar alrededor de un 5% del conjunto total de datos), realizar una investigación particular sobre ellos que permita tomar mejores decisiones.
- Recuerda que borrar outliers o datos, en general puede llevarte a obtener análisis más bonitos, pero fuera de la realidad.
- Qué hacer con los outliers también dependerá del tamaño de datos que tengas: si tienes cientos de miles, probablemente no afecte mucho al modelo el borrarlos (igualmente puede ser algo cuestionable, sobre todo éticamente); pero si tienes muy pocos, el impacto de borrarlos sería muy notorio, por lo que tendrías que considerar bien tu decisión.
- En general si tienes el tiempo y los recursos, genera tres modelos y compara los resultados para elegir el mejor:
  - Primer modelo: realizar tu análisis con todos los datos.
  - Segundo modelo: Primero borrar los outliers, luego transformarlos y luego generar tu modelo.
  - Tercer modelo: Primero transformarlos con todo y outliers, luego borrar los que queden y luego generar tu modelo.

## Transformaciones: Histogramas con Sesgo



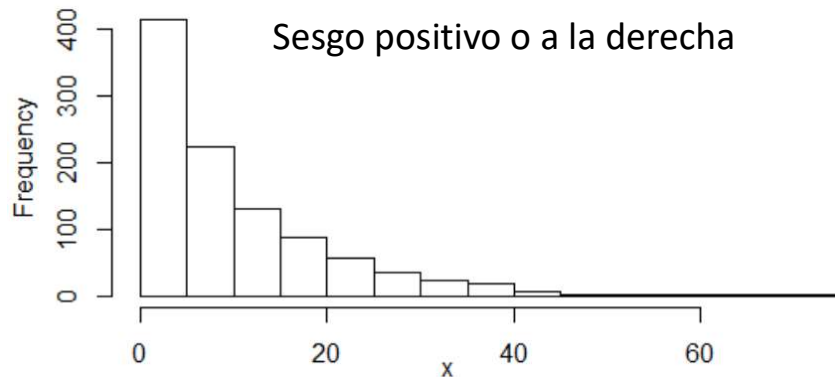
Histograma con sesgo  
positivo o a la derecha.



Histograma con sesgo  
negativo o a la izquierda.

Otro tipo de transformaciones, como las que veremos a continuación, ayudan a enfrentar el sesgo de una distribución de datos.

### Histograma de datos originales $X$



- $X' \leftarrow \log(X)$

- $X' \leftarrow \log(X + 1)$

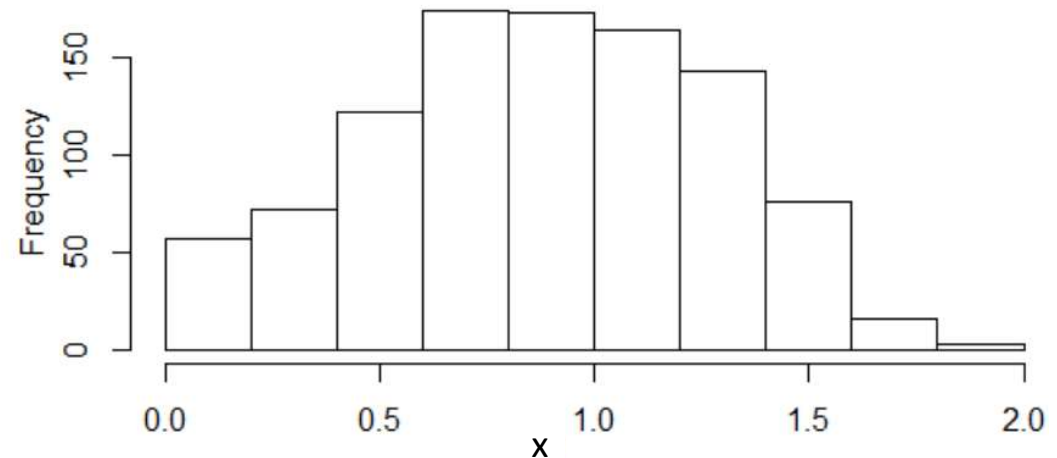
- $X' \leftarrow \log(X + c)$

si se tienen mínimos de valor negativo  
tomar  $c = |\min(X)| + 1$ ,

### Transformación Logarítmica $\log(x)$

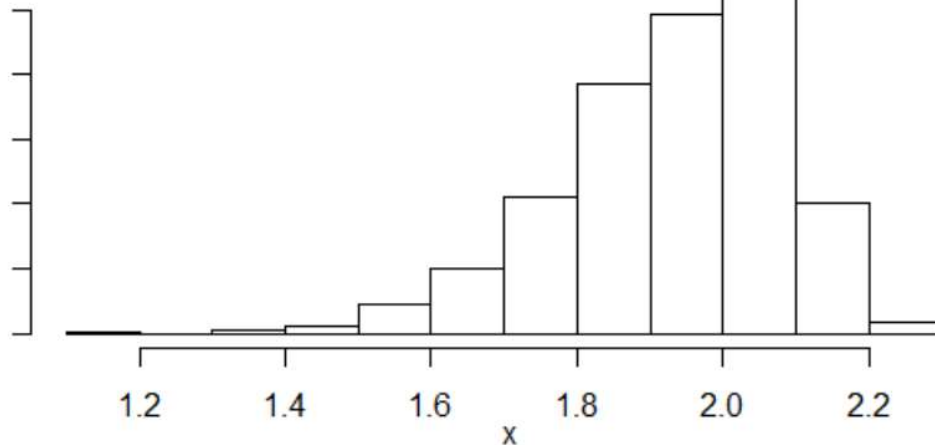
En ocasiones se suma la constante 1 o alguna otra constante positiva a los datos originales antes de aplicar el logaritmo, para generar un nuevo conjunto de datos positivos o no-negativos. Sobre todo si el mínimo es negativo o entre cero y uno.

### Histograma de datos transformados $X'$



### Datos originales:

Sesgo negativo o a la izquierda



Para transformaciones con sesgo negativo se pueden intentar transformaciones logarítmicas como sigue:

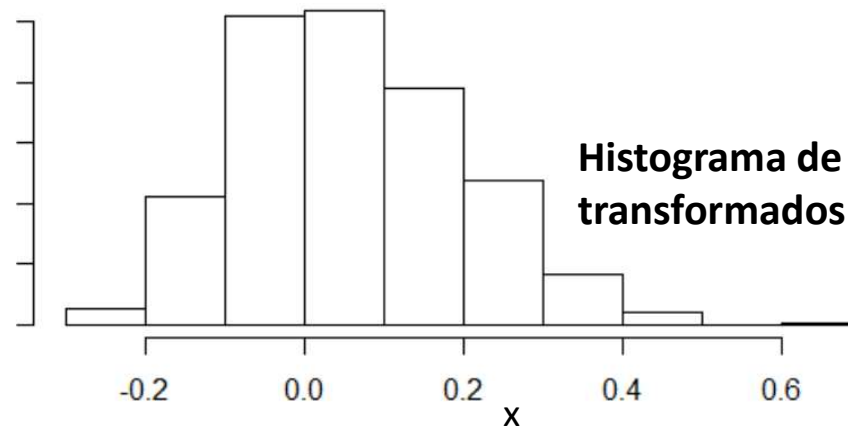
- $X' \leftarrow \log(c - X)$

donde  $c > \max(X)$

y dicho máximo es positivo.

Si el máximo es negativo, se puede intentar multiplicar simplemente por menos uno y después aplicar el logaritmo.

**Transformación Logarítmica**  
 $\log(c - x)$

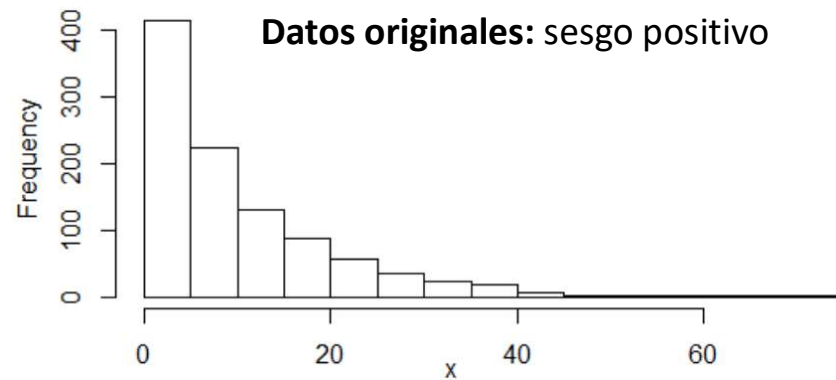


**Histograma de datos transformados  $X'$**



- $X' \leftarrow \text{sqrt}(X)$
- $X' \leftarrow \text{sqrt}(X + c)$

donde  $X + c > 0$



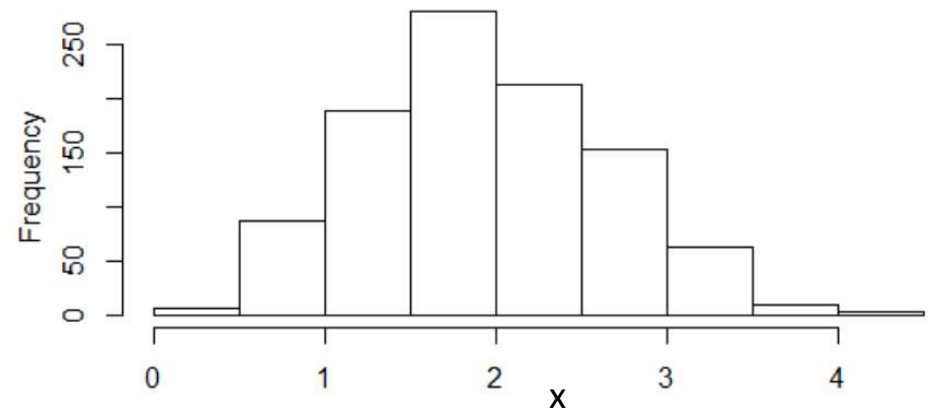
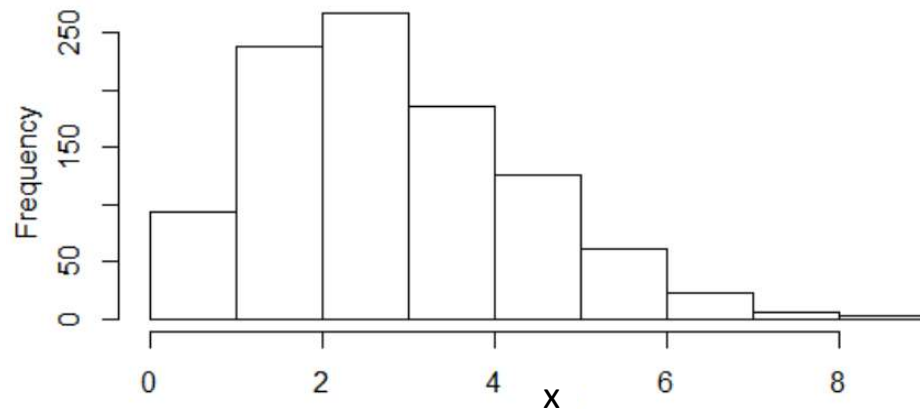
**Raíz cuadrada (square root)**

$$\sqrt{x}$$

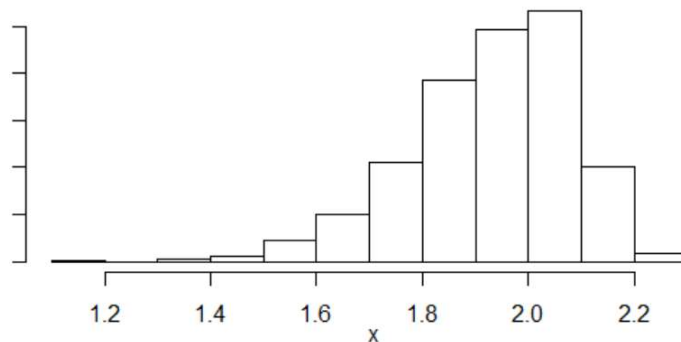
**Raíz cúbica (cube root)**

$$\sqrt[3]{x}$$

**Histogramas de datos transformados  $X'$**



**Datos originales:**  
Sesgo negativo o a la izquierda



**Raíz cuadrada**

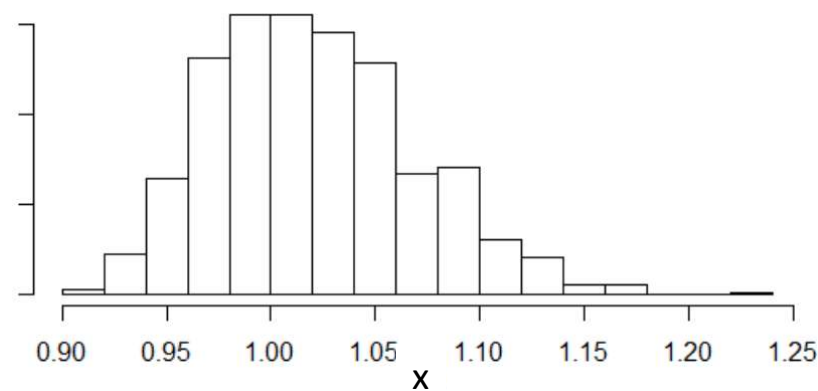
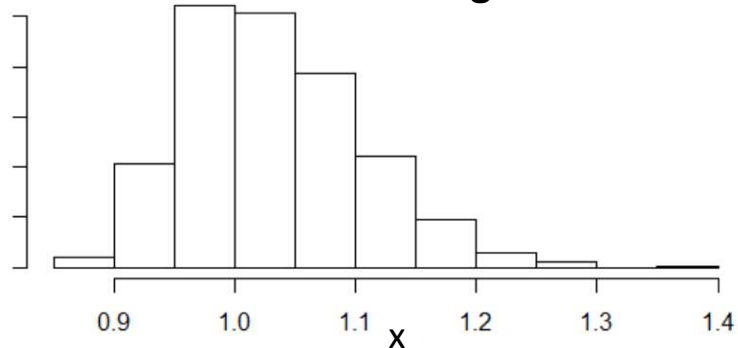
$$\sqrt{c - x}$$

donde  $c > \max(x)$

**Raíz cúbica**

$$\sqrt[3]{c - x}$$

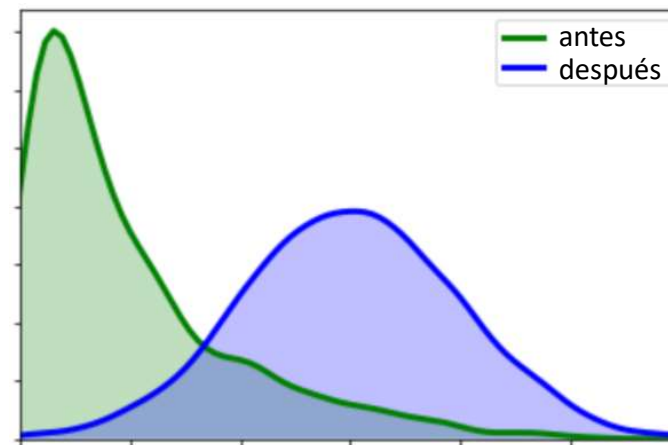
**Histograma de datos transformados  $X'$**



## Otro tipo de transformaciones:

- Transformación Box-Cox / Transformación de potencias (power transform/law):

$$y \leftarrow \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(y) & \text{si } \lambda = 0 \end{cases}$$



Usualmente  $\lambda$  suele variar y buscarse entre  $-5$  y  $+5$ .

Debes asegurarte antes que los valores sean positivos.

Existen otras variantes de dicha transformación, pero esta es de las más usadas.

## Notas varias sobre la transformación de los datos

- Hay que considerar que ninguna de estas transformaciones garantiza que los datos transformados tendrán un comportamiento acampanado.
- Recuerda que también se pueden usar otras técnicas como discretización de variables continuas, o tratamiento inicial de los valores extremos, etc.
- La transformación buscada dependerá siempre de una inspección inicial de los datos en cuestión.
- Hemos visto también que en ocasiones conviene incluir relaciones no-lineales, por ejemplo relaciones polinomiales de variables independientes.
- El método de PCA se utiliza en ocasiones como método de transformación al incluir todas las variables desde el inicio, para posteriormente seleccionar las componentes principales más representativas. Este tema se estudiará más adelante en el curso.

## Centrado & Escalamiento de los datos

$$x_k \leftarrow \frac{x_k - \bar{x}_k}{S_k}$$

} centrado  
} escalamiento

- En ocasiones solamente se lleva a cabo el centrado de los datos. Entre algunas de las razones para considerar solo el centrado podemos mencionar las siguientes: porque los datos de las diferentes variables ya están aproximadamente en la misma escala; o porque se quiere respetar la distribución original y no hacerlos varianza 1.

## Filtrado de información (data leakage)

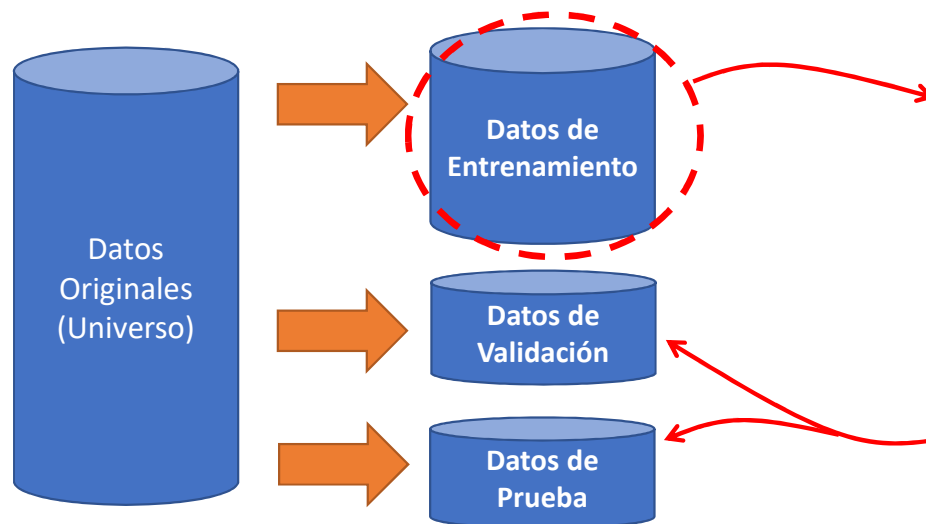
### Transformaciones en los conjuntos de Entrenamiento, Validación y Prueba

- Recuerda que en el proceso de generar cualquier modelo de aprendizaje automático, las decisiones que tomes en relación a cualquier tipo de transformación de los datos en la etapa de limpieza y preprocesamiento de los mismos, **deben ser tomadas analizando y extrayendo información únicamente del conjunto de entrenamiento.**
- Es decir, primero debes generar la partición de tu conjunto original de datos en entrenamiento, validación y prueba y después empezar a analizar y a extraer información del conjunto de entrenamiento. Las decisiones que tomes con base al conjunto de entrenamiento, podrás aplicarlas posteriormente a los conjuntos de validación y prueba.
- Esto debe ser así, ya que los conjuntos de validación y prueba se usan para simular datos futuros, a los cuales se supone que aún no tienes acceso.
- Analizar los datos de validación o prueba para extraer información durante el proceso de limpieza o preparación de los datos, se conoce como **filtrado de información** o **filtrado de datos (data leakage** en inglés). Este es un error muy común que se comete por los analistas poco experimentados.

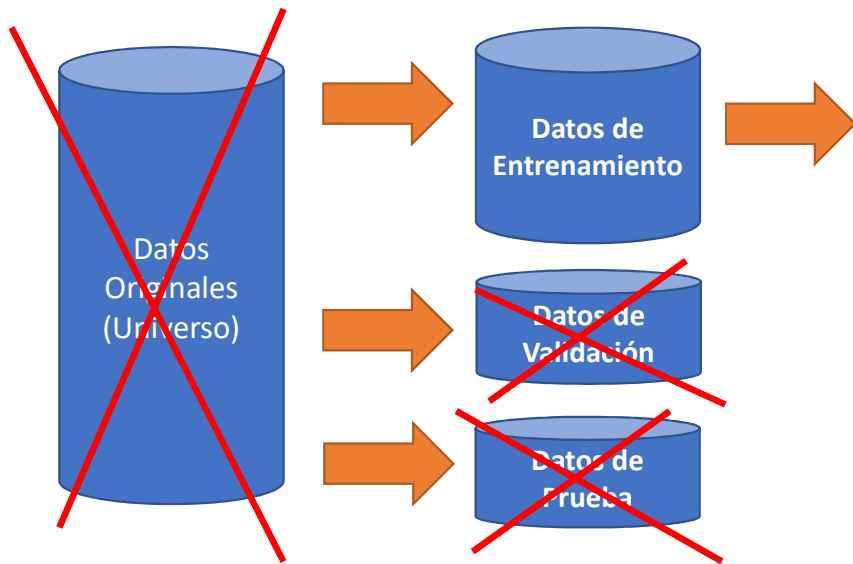
## Filtrado de información (data leakage)

### Transformaciones en los conjuntos de Entrenamiento, Validación y Prueba

- Así, la manera correcta de proceder para evitar el filtrado de información sería como sigue: sin extraer información relevante del conjunto original (universo) de datos, llevar a cabo la partición en el conjunto de entrenamiento, validación y prueba. A continuación **analizar y extraer la información necesaria del conjunto de datos de entrenamiento**. Las decisiones que hayas tomado y extraído con base al conjunto de entrenamiento, las deberás aplicar ahora a los conjuntos de validación y prueba. Por ejemplo, si decides aplicar una transformación gaussiana a los datos numéricos de entrada, entonces **deberás obtener la media y desviación estándar del conjunto de entrenamiento y utilizar estos mismos parámetros** para transformar al conjunto de prueba y al de validación.



Por ejemplo, supongamos que al analizar los datos del conjunto de entrenamiento se decide aplicar la transformación  $\frac{x - MEA}{STD}$  a las variables numéricas de entrada, entonces la media MEAN y la desviación estándar STD deberás obtenerlas usando solamente los datos de entrenamiento y posteriormente estos mismos valores usarlos para transformar los datos de los conjuntos de validación y de prueba.



Recuerda, toda la información que desees extraer de los datos en la etapa de limpieza y preprocesamiento debe proceder del conjunto de entrenamiento: promedio, desviación estándar, varianza, moda, mediana, máximo, mínimo, rango, cuartiles, percentiles, análisis de sesgo de histogramas, análisis de datos perdidos, análisis de valores extremos (outliers), etc.

- La restricción de no extraer información durante la etapa de preprocesamiento aplica también al conjunto (universo) original de datos: tampoco puede extraerse información previa de dicho conjunto, porque nuevamente estarías analizando información futura que pertenece a los conjuntos de validación y prueba.



