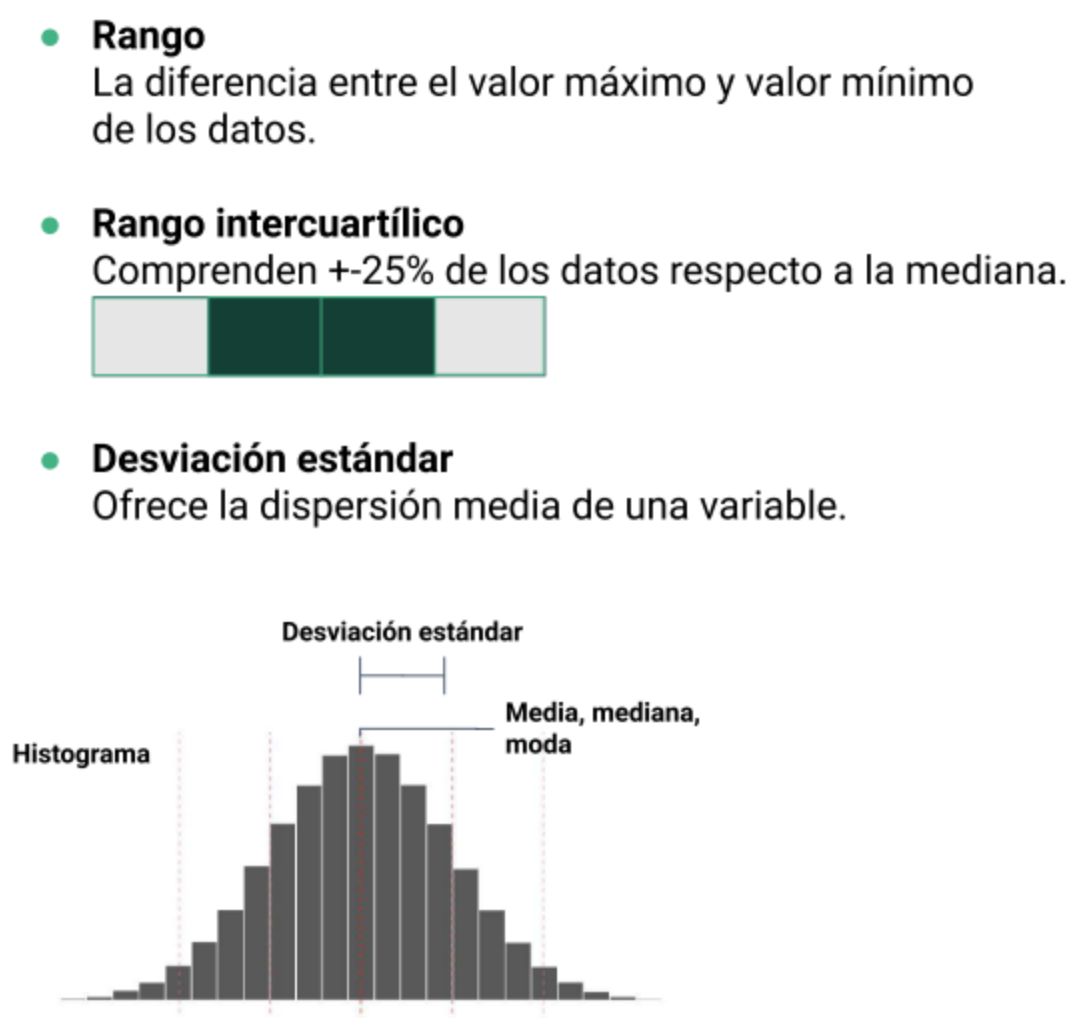
Estadística descriptiva aplicada: medidas de dispersión

Ahora surge la pregunta de

¿Qué tan cerca o lejos se encuentran las medidas de tendencia central, de los datos?

Las medidas que responden esta pregunta son las medidas de dispersión. Principalmente usamos 3:

- **Rango:**: Es la diferencia entre el valor máximo y el valor mínimo de los datos.
- **Rango intercuartílico:** comprenden el ± 25 de los datos respecto a la mediana. Este parámetro lo corta en 4 partes iguales y nos quedamos con las partes centrales, en donde la linea central es la mediana.
- **Desviación estándar:** Nos va a ofrecer una medida de dispersión respecto a la media de una variable.



Aquí tenemos un histograma.

Este nos va a permitir representar la:

- media
- mediana
- moda

Dentro de un mismo lugar, para el caso de la distribución normal, todos estos valores se encuentran en el valor central o en el centro de la distribución y además son iguales.

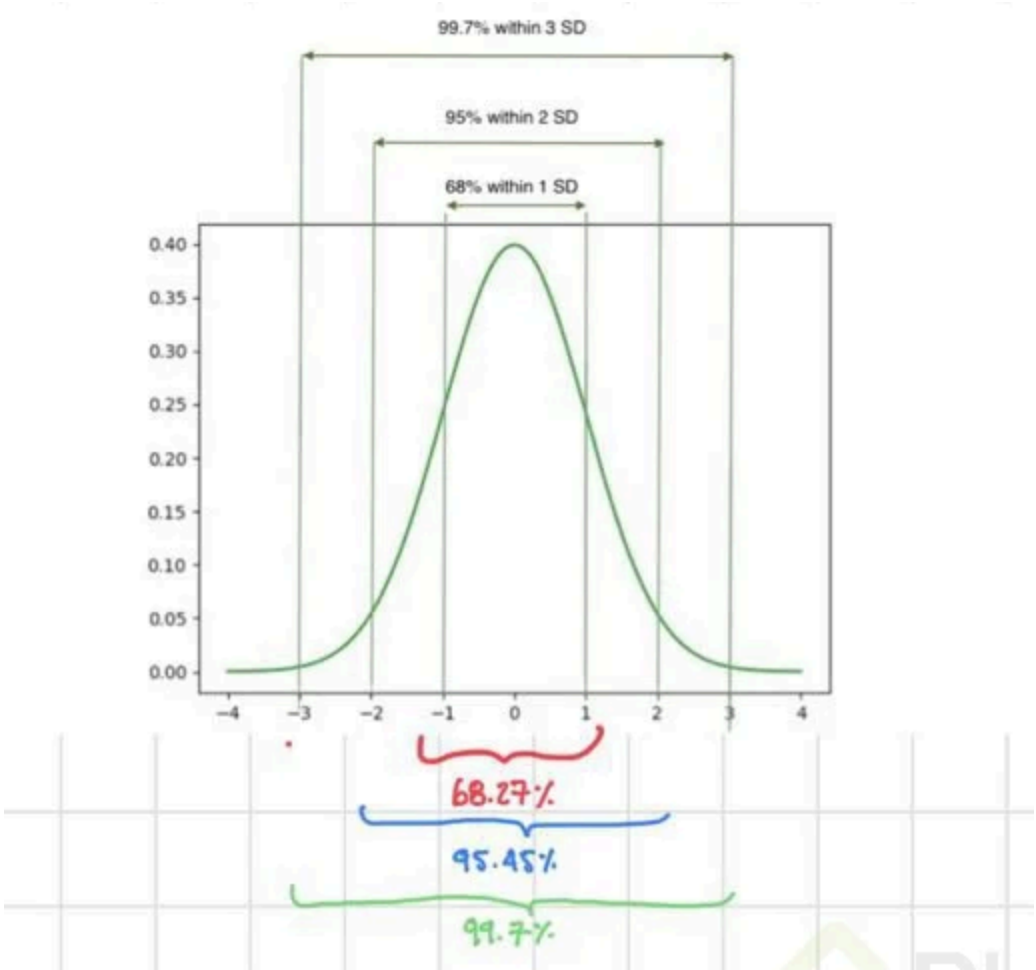
Después tenemos a la **desviación estándar**, que nos va a permitir saber como están distribuidos los datos.

Existe una regla **68, 95, 99.7**, en el que se utiliza para recordar el **%** de valores que se encuentran dentro de una banda alrededor de la **media** en una distribución normal con un ancho de: **2, 4 y 6** veces la **desviación típica** respectivamente.

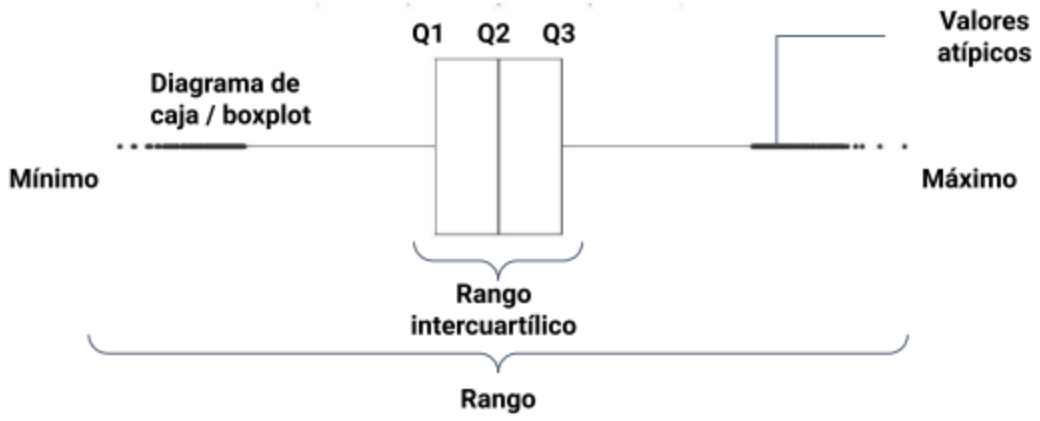
Exactamente el **68.27%, 95.45% y 99.73%** de los valores se encuentran dentro de las bandas con semiancho de **1, 2 y 3** veces la desviación estándar respecto de la media.

Donde:

- 1SD: 1 Vez la Desviación estándar
- 2SD: 2 Veces la Desviación estándar
- 3SD: 3 Veces la Desviación estándar



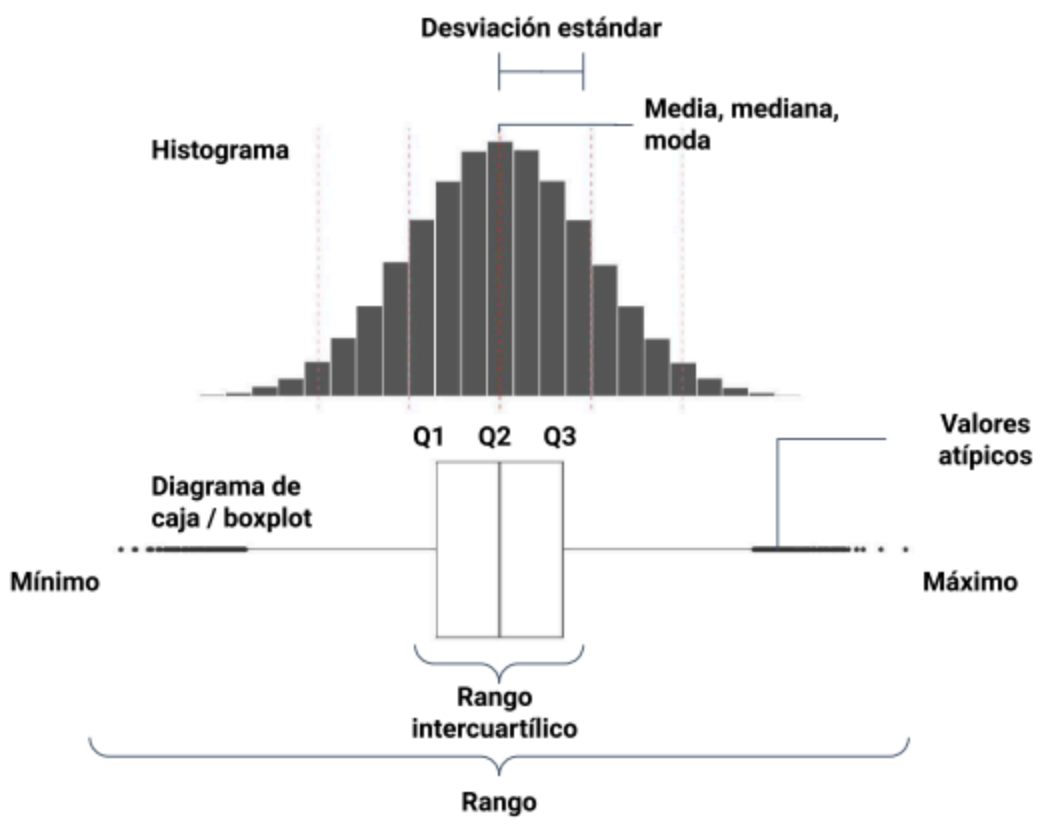
Después tenemos nuestro diagrama de caja o **box plot**



De está manera podemos visualizar el valor mínimo y máximo de una distribución, también podemos visualizar el rango, en este caso que tan dispersos están nuestros datos. En el centro observaremos que tenemos una caja, esta caja representa entonces el **rango intercuartílico**, en el centro encontramos a la mediana que la divide a la mitad. Generalmente cada punto que la intersecta son conocidos como **cuartiles** (Q1,Q2,Q3).

Después tenemos los puntos negros a los costados de nuestro box plot, estos puntos son **valores atípicos**, son aquellos valores que están fuera del rango de valores esperados de la distribución.

Así también podemos observar la relación de las gráficas.

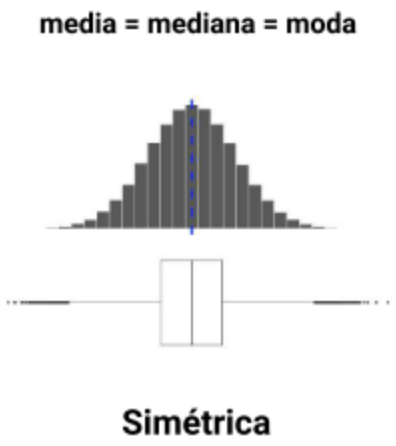


Asimetría estadística

Finalmente podemos hablar de como luce la distribución, para entenderlo un poco nos apoyaremos de una ilustración.

Distribución simétrica

Cuando una distribución es simétrica existe un equilibrio entre el rango de valores, y al existir una gráfica podemos ver su comportamiento equilibrado.

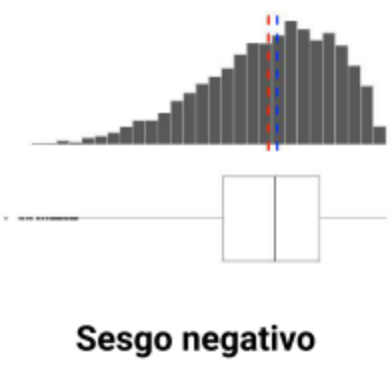


Simétrica

¿Pero que pasa cuando los datos no están distribuidos de manera simétrica? entonces tenemos una asimetría. Para el caso de una **asimetría estadística**, hablaremos de sesgo **negativo o positivo.**

Sesgo negativo

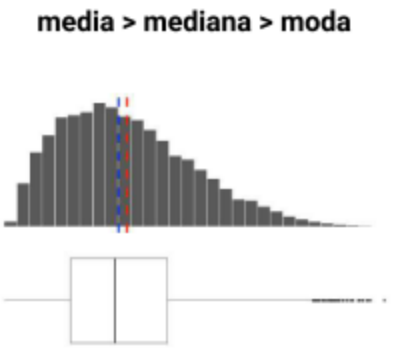
Cuando la distribución tiene un sesgo negativo, la media o promedio, será el valor de **estadísticos de tendencia central** que es más pequeño, después se encontrará la mediana y al último la moda.



Sesgo negativo

Sesgo positivo

Cuando la distribución tiene un sesgo positivo, la media o promedio, será el valor de **estadísticos de tendencia central** que es más grande, después se encontrará la mediana y al último la moda.



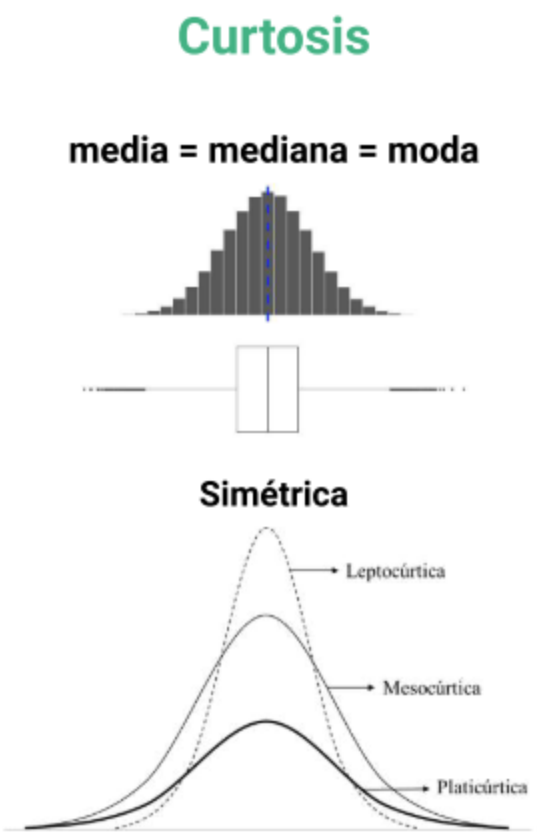
Sesgo positivo

De esta forma podemos describir si nuestros datos están sesgados de manera positiva o negativa. Y esto nos va a servir para tratar de ajustar modelos o tratar de ver si los datos deberían de comportarse de tal forma o se debería de realizar una transformación para que se comporten de una manera esperada y que se pueda controlar.

Curtosis

La curtosis lo que hará es que me dirá que tan juntos o dispersos se encuentran mis datos respecto a la media o promedio. Ls curtosis es un número que cuando tiene un valor de:

- 0, significa que los datos están distribuidos homogéneamente alrededor de la media y se conoce como **Distribución mesocúrtica**.
- >0, significa que los datos están concentrados alrededor de la media, a esto se le conoce **Distribución Leptocúrtica**
- <0, significa que nuestros datos se distribuyen a lo largo de nuestra media y se le conoce como una **Distribución Platicúrtica**



Curtosis en una Distribución

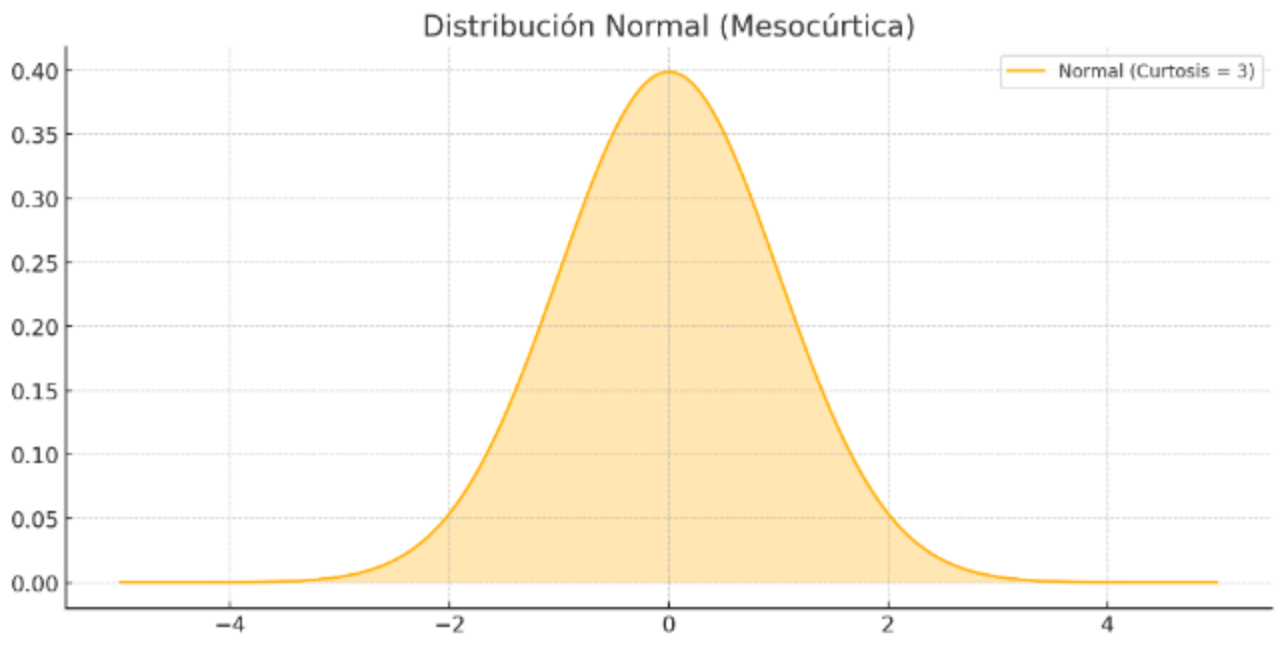
La curtosis es una medida estadística que describe la forma de las colas de una distribución de datos en comparación con una distribución normal. En términos simples, la curtosis nos dice si los datos tienen colas más gruesas (más valores extremos) o más delgadas (menos valores extremos) que una distribución normal.

Tipos de Curtosis:

Distribución Normal (Mesocúrtica):

Curtosis = 3

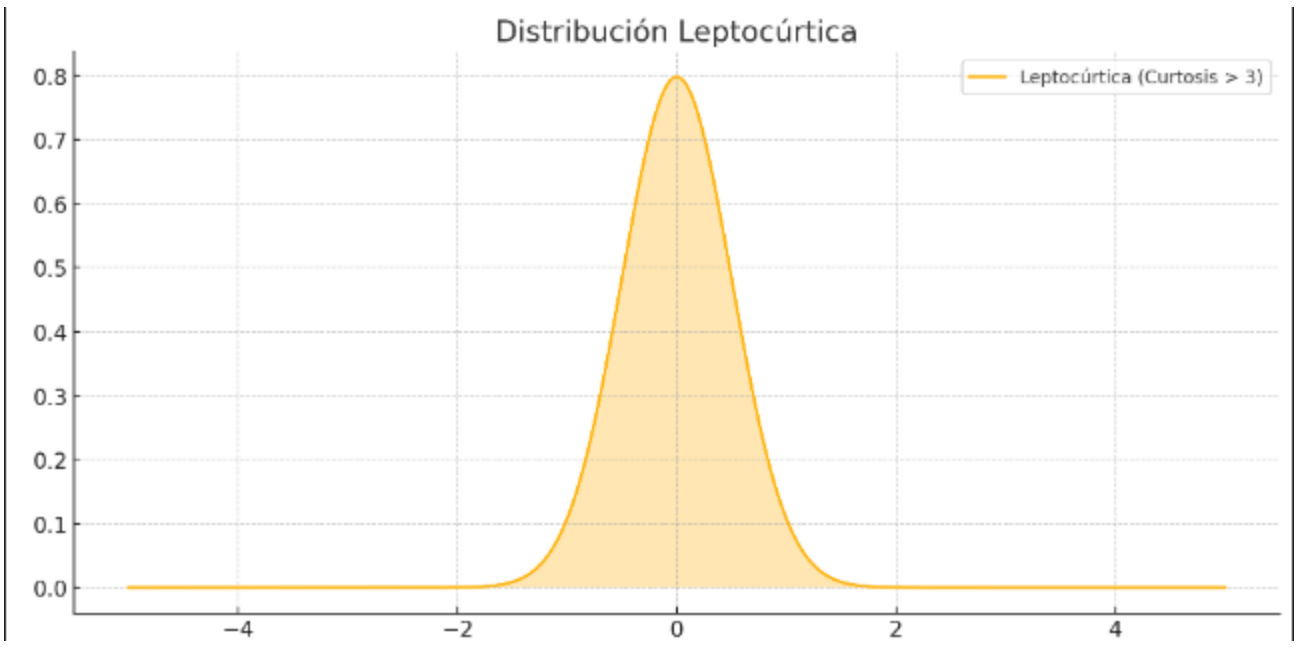
1. **Curtosis Normal (Mesocúrtica):**
   - La distribución normal tiene una curtosis de 3.
   - No tiene colas especialmente gruesas ni delgadas.



Distribución Leptocúrtica:

Curtosis > 3

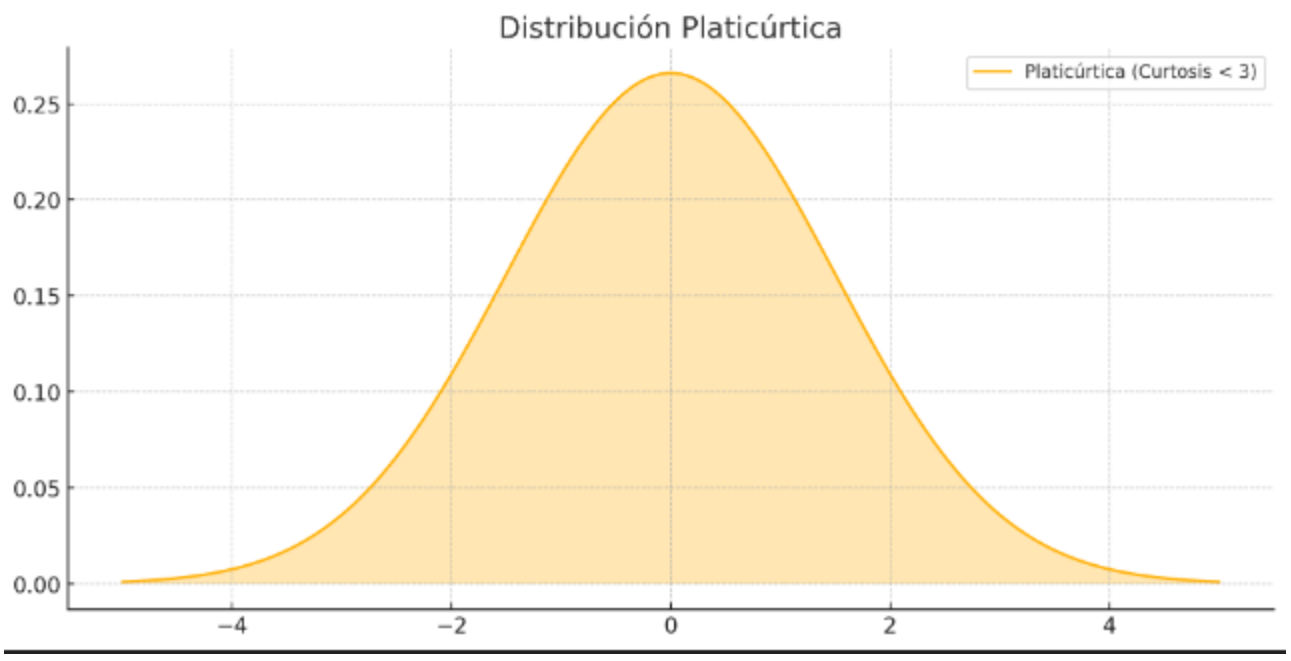
2. **Curtosis Alta (Leptocúrtica):**
   - Tiene una curtosis mayor a 3.
   - Colas más gruesas y picos más altos y estrechos.
   - Indica la presencia de más valores extremos (outliers).



Distribución Platicúrtica:

Curtosis < 3 3. **Curtosis Baja (Platicúrtica):**

- Tiene una curtosis menor a 3.
- Colas más delgadas y picos más bajos y anchos.
- Indica menos valores extremos.



Estas gráficas ayudan a visualizar cómo la curtosis afecta la forma de una distribución y sus colas.

NOTA IMPORTANTE

Curtosis vs. Curtosis Excesiva

En algunas fuentes y contextos, la curtosis puede ser ajustada restando 3 de su valor, lo cual se llama **"curtosis excesiva"**. Esto se hace para que la distribución normal tenga una curtosis de 0, facilitando la comparación con otras distribuciones.

Curtosis (clásica)

- En la definición clásica, una distribución normal tiene una curtosis de 3.
- Las distribuciones leptocúrticas (más picos y colas más gruesas) tienen una curtosis mayor que 3.
- Las distribuciones platicúrticas (más planas y colas más delgadas) tienen una curtosis menor que 3.

Curtosis Excesiva

- La curtosis excesiva se calcula restando 3 de la curtosis clásica.
- Así, la distribución normal tiene una curtosis excesiva de 0.
- Distribuciones leptocúrticas tendrán una curtosis excesiva mayor que 0.
- Distribuciones platicúrticas tendrán una curtosis excesiva menor que 0.

Ejemplo

- Si una distribución tiene una curtosis clásica de 3 (como la normal), su curtosis excesiva será 3 - 3 = 0.
- Si una distribución tiene una curtosis clásica de 5, su curtosis excesiva será 5 - 3 = 2.
- Si una distribución tiene una curtosis clásica de 2, su curtosis excesiva será 2 - 3 = -1.

En resumen

En algunas clases o fuentes, especialmente en estadística aplicada, se usa la curtosis excesiva para simplificar la comparación entre distribuciones. Cuando se dice que una distribución homogénea (como la normal) tiene una curtosis de 0, se está refiriendo a la curtosis excesiva.

Extras:

- Estadística básica en Python para Machine Learning
- Curtosis Wikipedia