# Notas del curso: Estadística inferencial

# Distribuciones

## Definición

En estadística, cuando hablamos de distribuciones, normalmente nos referimos a distribuciones de probabilidad.

Definición (informal): Una distribución es una función que muestra los valores posibles para una variable y la frecuencia con la que se producen.

Definición (Wikipedia): En teoría de la probabilidad y estadística, la distribución de probabilidad de una variable aleatoria es una función que asigna a cada suceso definido sobre la variable la probabilidad de que dicho suceso ocurra.

**Ejemplos:** Distribución normal, Distribución T de Student, Distribución de Poisson, Distribución uniforme, Distribución binomial

## Representación gráfica

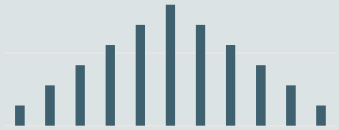
Es un error común creer que la distribución es el gráfico. De hecho, la distribución es la "regla" que determina cómo se posicionan los valores entre sí.

Muy a menudo, utilizamos un gráfico para visualizar los datos. Dado que las diferentes distribuciones tienen una representación gráfica particular, a los estadísticos les gusta dibujarlas.

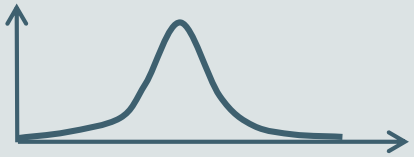
**Ejemplos:**

Distribución uniforme

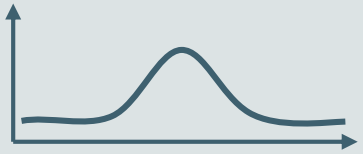
Distribución binomial



Distribución normal



Distribución de T de Student



# La distribución normal

La distribución Normal también se conoce como distribución Gaussiana o curva de Campana. Es una de las distribuciones más comunes debido a las siguientes razones:

- Se aproxima a una amplia variedad de variables aleatorias.
- Las distribuciones de las medias muestrales con tamaños de muestra suficientemente grandes podrían aproximarse a los normales.
- Todas las estadísticas computables son elegantes.
- Muy utilizada en el análisis de regresión
- Buen historial.

Ejemplos:

- Biología. La mayoría de las medidas biológicas están distribuidas en forma normal, tales como: altura; longitud de los brazos, piernas, uñas; presión arterial; espesor de la corteza de los árboles, etc.
- Pruebas de Coeficiente Intelectual (CI).
- Información bursátil.

$$N \sim (\mu, \sigma^2)$$

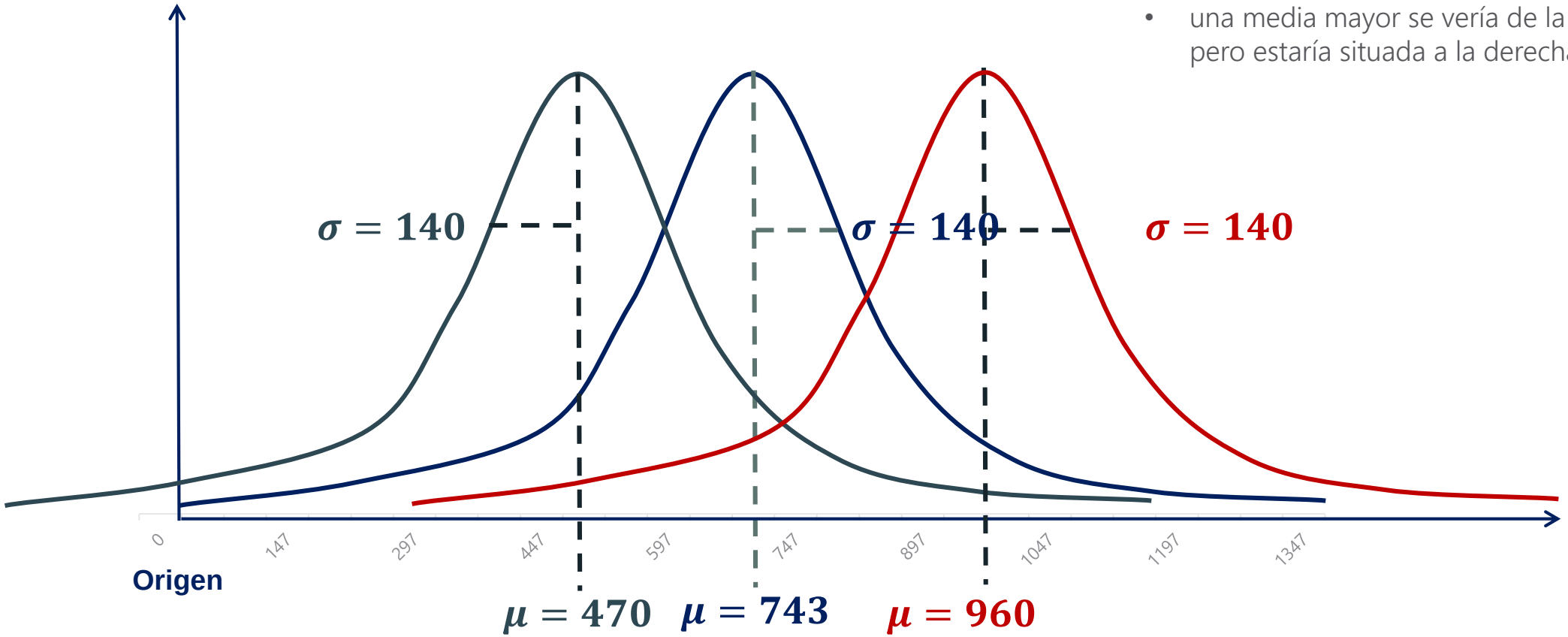
*N* significa normal;
~ representa una distribución;
μ es la media;
$\sigma$^2 es la variación.

# La distribución normal

Controlando por la desviación estándar

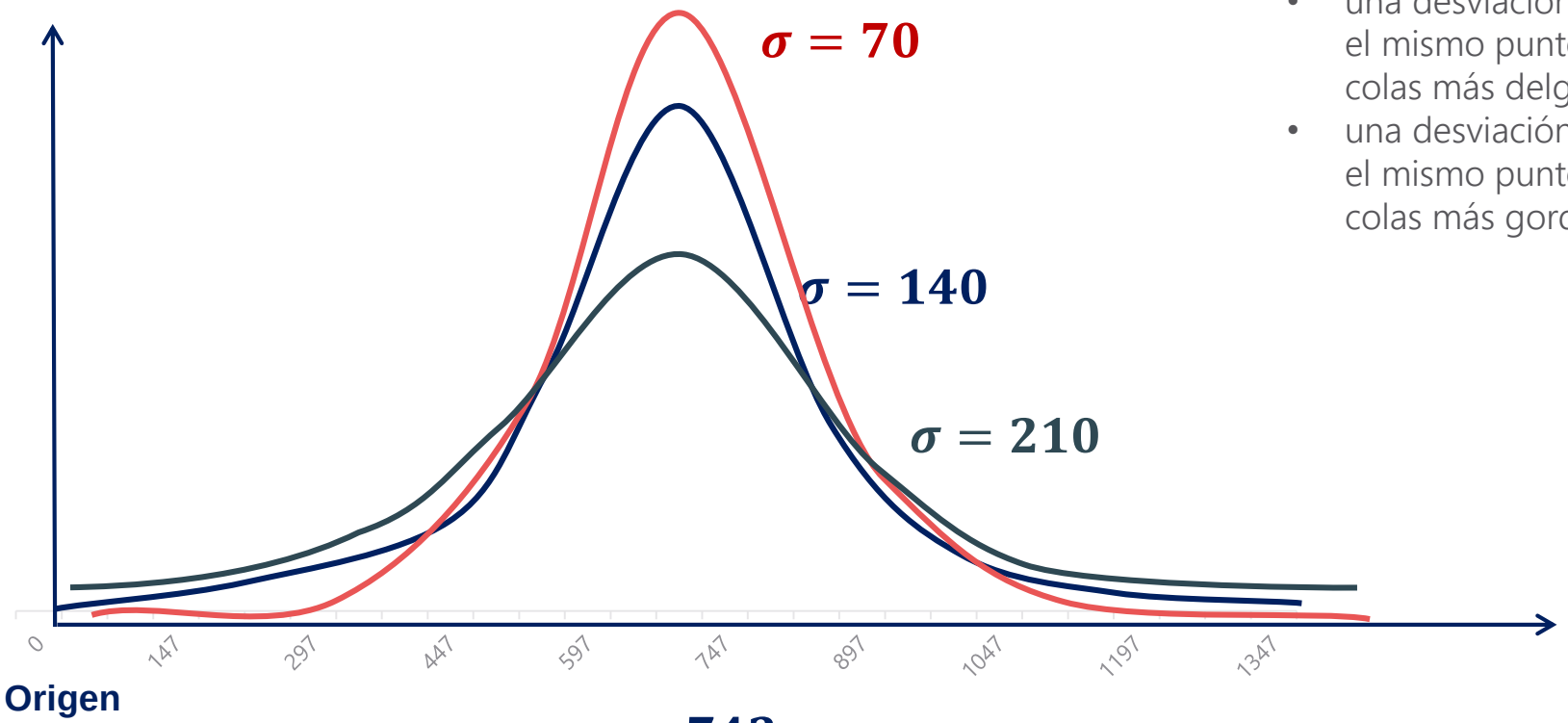
Manteniendo constante la desviación estándar, el gráfico de una distribución normal con:

- una media menor se vería de la misma manera, pero estaría situada a la izquierda (en gris)
- una media mayor se vería de la misma manera, pero estaría situada a la derecha (en rojo)



# La distribución normal

Controlando por la media




$\mu = 743$

$\mu = 743$

$\mu = 743$

Manteniendo constante la media, una distribución normal con:

- una desviación estándar menor estaría situada en el mismo punto, pero tendría un pico más alto y colas más delgadas (en rojo).
- una desviación estándar mayor estaría situada en el mismo punto, pero con un pico más bajo y colas más gordas (en gris).

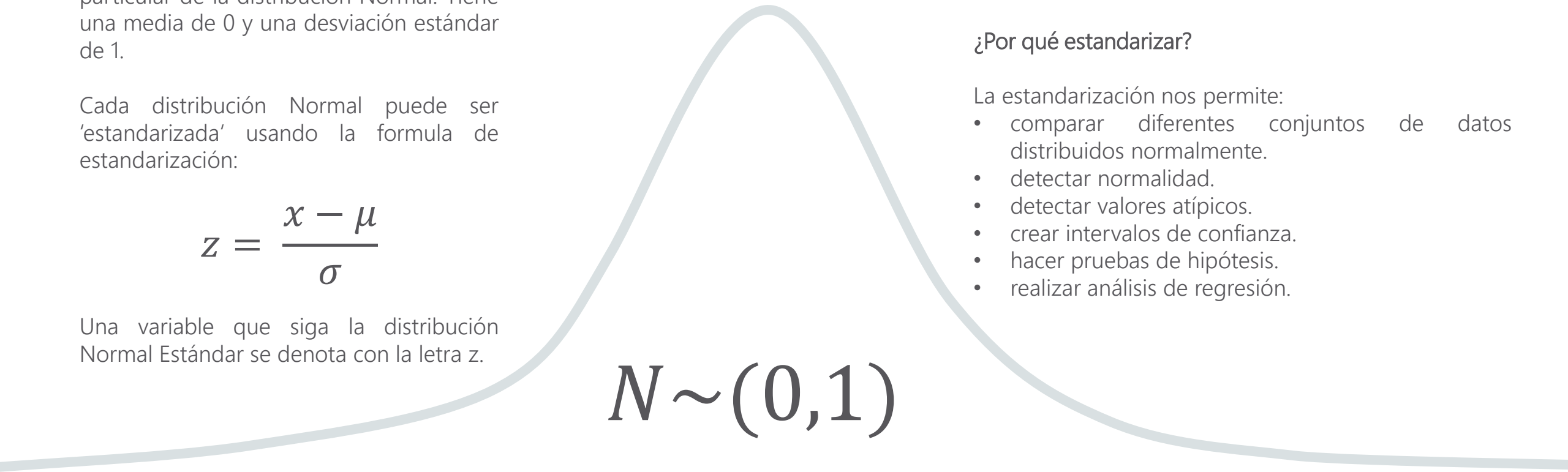
# La distribución normal estándar

La desviación normal estándar es un caso particular de la distribución Normal. Tiene una media de 0 y una desviación estándar de 1.

Cada distribución Normal puede ser 'estandarizada' usando la formula de estandarización:

$$z = \frac{x - \mu}{\sigma}$$

Una variable que siga la distribución Normal Estándar se denota con la letra z.




¿Por qué estandarizar?

La estandarización nos permite:

- comparar diferentes conjuntos de datos distribuidos normalmente.
- detectar normalidad.
- detectar valores atípicos.
- crear intervalos de confianza.
- hacer pruebas de hipótesis.
- realizar análisis de regresión.

Explicación para la formula de estandarización:

Queremos transformar una variable aleatoria de $N \sim (\mu, \sigma^2)$ a $N \sim (0,1)$. Restar la media de todas las observaciones causaría una transformación de $N \sim (\mu, \sigma^2)$ a $N \sim (0, \sigma^2)$, moviendo el gráfico al origen. Subsecuentemente, dividir todas las observaciones por la desviación estándar causaría una transformación de $N \sim (0, \sigma^2)$ a $N \sim (0,1)$, estandarizando el pico y las colas del gráfico.

# El teorema de límite central

El Teorema del Límite Central (TLC) es una de las mejores ideas de la estadísticas. Este afirma que, independientemente de la distribución subyacente del conjunto de datos, la distribución muestral de las medias se aproximaría a una distribución normal. Además, la media de la distribución muestral sería igual a la media de la distribución original y la varianza sería n veces menor, siendo n el tamaño de las muestras. El TLC se aplica siempre que tengamos una suma o un promedio de muchas variables (por ejemplo, la suma de los números que resultan de lanzar los dados).

## El teorema

- No importa la distribución
- La distribución de $\overline{x_1}$, $\overline{x_2}$, $\overline{x_3}$, $\overline{x_4}$, …, $\overline{x_k}$ tendería a $N \sim \left(\mu, \frac{\sigma^2}{n}\right)$
- A mayor cantidad de muestras, más cercana será a la normal ( k -> ∞ )
- A mayor tamaño de muestras, más cercana será a la Normal ( n -> ∞ )

## ¿Por qué es útil?

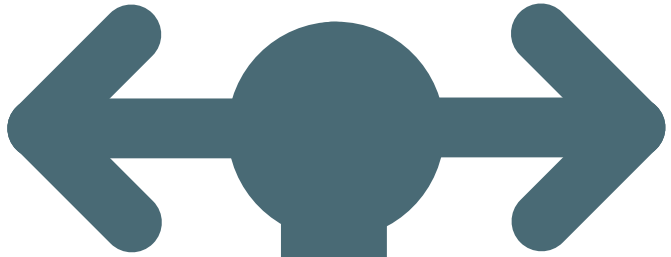
El TLC nos permite asumir la normalidad de muchas variables diferentes. Esto es muy útil para intervalos de confianza, pruebas de hipótesis y análisis de regresión. De hecho, la distribución Normal es tan predominante a nuestro alrededor debido al hecho de que debido a la CLT, muchas variables convergen a Normal.

Haz clic aquí para ir a un simulador de TLC.

## ¿Dónde podemos verlo?

Puesto que muchos conceptos y eventos son una suma o una media de efectos diferentes, El TLC aplica y observamos la normalidad todo el tiempo. Por ejemplo, en el análisis de regresión, la variable dependiente se explica a través de la suma de términos de error.

# Estimadores y estimaciones



## Estimadores

En términos generales, un estimador es una función matemática que se aproxima a un parámetro de población dependiendo únicamente de la información de la muestra.

Ejemplos de estimadores y los parámetros correspondientes:

| Término | Estimador | Parámetro |
|---|---|---|
| Media | $\bar{x}$ | $\mu$ |
| Varianza | $s^2$ | $\sigma^2$ |
| Correlación | $r$ | $\rho$ |

Los estimadores tienen dos propiedades importantes:

- Sesgo

El valor esperado de un estimador imparcial es el parámetro de población. El sesgo en este caso es 0. Si el valor esperado de un estimador es (parámetro + b), entonces el sesgo es b.

- Eficiencia

El estimador más eficiente es el que tiene la menor varianza.

## Estimaciones

Una estimación es el resultado que se obtiene del estimador (cuando se aplica la fórmula). Existen dos tipos de estimaciones: estimaciones puntuales y estimaciones de intervalos de confianza.

### Estimados puntuales

Un valor puntual.

Ejemplos:

- 1
- 5
- 122,67
- 0,32

### Intervalos de confianza

Un intervalo.

Ejemplos:

- ( 1 ; 5 )
- ( 12 ; 33)
- ( 221,78 ; 745,66)
- ( - 0,71 ; 0,11 )

Los intervalos de confianza son mucho más precisos que las estimaciones puntuales. Es por eso que se prefieren cuando se hacen inferencias.

# Intervalos de confianza y margen de error


Definición: Un intervalo de confianza es un intervalo dentro del cual tenemos seguridad (con un cierto porcentaje de confianza) de que el parámetro de población caerá.

Construimos el intervalo de confianza **alrededor** de la estimación puntual.

**(1-α)** es el nivel de confianza. Tenemos un (1-α)*100% de confianza de que el parámetro de población caerá en el intervalo especificado. Los alfas comunes lo son: 0.01, 0.05, 0.1.

Fórmula general:

**[ $\bar{x}$ - ME, $\bar{x}$ + ME ]** , donde ME es el margen de error.

**ME** = $factor\ de\ confianza * \frac{desviación\ estándar}{\sqrt{tamaño\ de\ la\ muestra}}$

$$z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

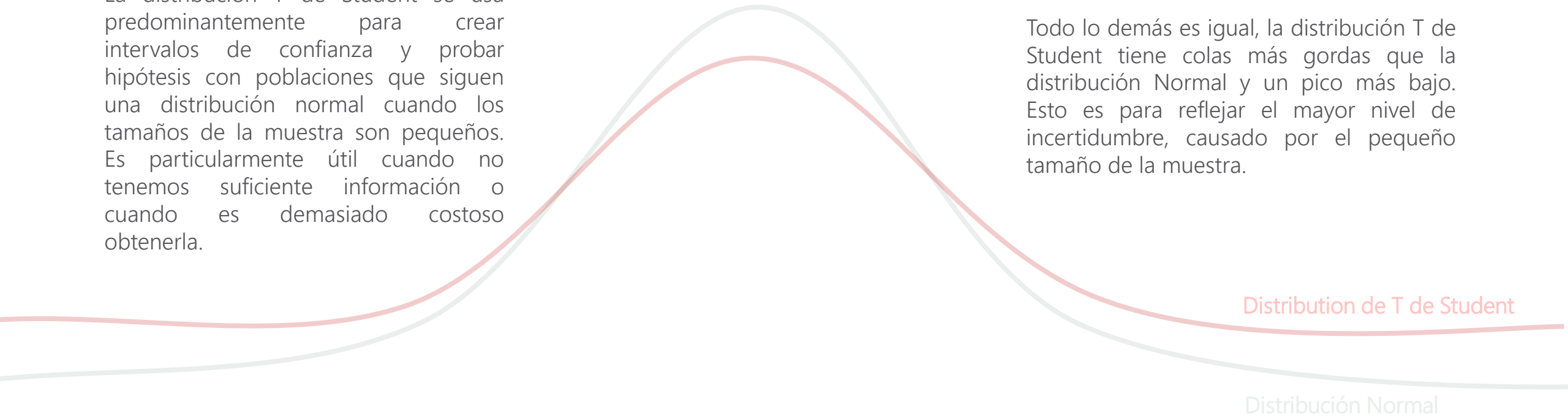
$$t_{\upsilon,\alpha/2} * \frac{s}{\sqrt{n}}$$

| Término | Efecto en la amplitud del intervalo de confianza |
|---|---|
| (1-α) ↑ | ↑ |
| $\sigma$ ↑ | ↑ |
| n ↑ | ↓ |

# Distribución de T de Student's T

La distribución T de Student se usa predominantemente para crear intervalos de confianza y probar hipótesis con poblaciones que siguen una distribución normal cuando los tamaños de la muestra son pequeños. Es particularmente útil cuando no tenemos suficiente información o cuando es demasiado costoso obtenerla.

Todo lo demás es igual, la distribución T de Student tiene colas más gordas que la distribución Normal y un pico más bajo. Esto es para reflejar el mayor nivel de incertidumbre, causado por el pequeño tamaño de la muestra.




Una variable aleatoria que sigue la distribución t se denota con $t_{\upsilon,\alpha}$ , donde $\upsilon$ son grados de libertad.

Podemos obtener la distribución T de Student para una variable con una población que sigue una distribución Normal usando la fórmula:

$$t_{\upsilon,\alpha} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

# Formulas para los intervalos de confianza

| # poblaciones | Varianza poblacional | Muestras | Estadística | Varianza | Fórmula |
|---|---|---|---|---|---|
| Una | Conocida | - | z | $\sigma^2$ | $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ |
| Una | Desconocida | - | t | $s^2$ | $\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$ |
| Dos | - | Dependiente | t | $s^2_{diferencia}$ | $\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$ |
| Dos | Conocida | Independiente | z | $\sigma_x^2$ , $\sigma_y^2$ | $(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$ |
| Dos | Desconocida, se asume igual | Independiente | t | $s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$ | $(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$ |
| Dos | Desconocida, se asume diferente | Independiente | t | $s_x^2$ , $s_y^2$ | $(\bar{x} - \bar{y}) \pm t_{\upsilon,\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$ |