
Universidad del País Vasco
Euskal Herriko Unibertsitatea
The University of the Basque Country

---

# E.U.I.T.I. Bilbao

## Asignatura:
## MÉTODOS ESTADÍSTICOS DE LA INGENIERÍA

E.U.I.T.I. Bilbao

# Asignatura:
# MÉTODOS ESTADÍSTICOS DE LA INGENIERÍA

## PARTE 2:
## ESTADÍSTICA INFERENCIAL

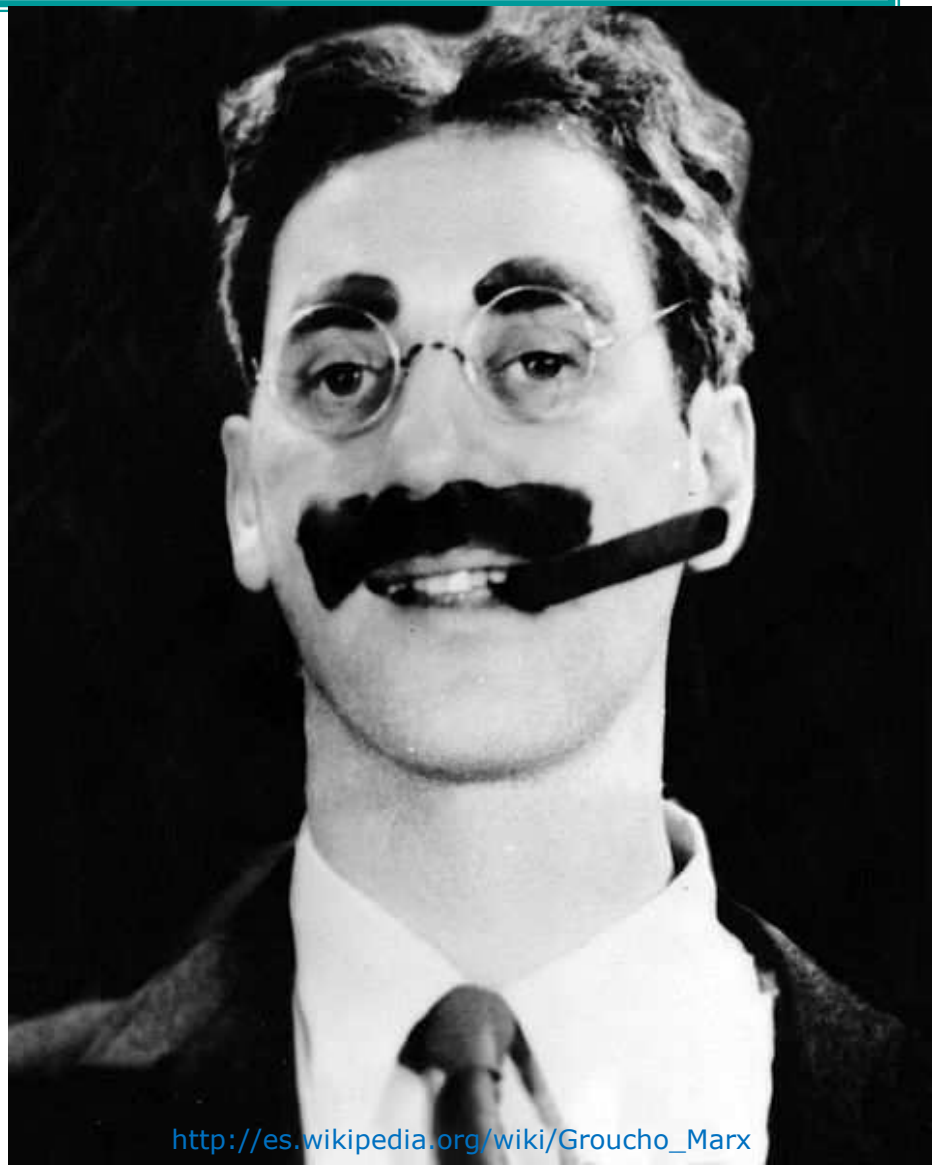
# 0. RECORDATORIO

**Estadística inferencial**. Rama de la Estadística que estudia procedimientos para hacer inferencias, ésto es, para deducir propiedades de una población a partir de la información obtenida de una muestra

**Objetivos básicos.** A partir de la información proporcionada por una muestra:

- **estimación**: estimar (evaluar) un parámetro desconocido de la población
- **contraste de hipótesis**: realización de una prueba estadística cuyo objeto es estudiar si una determinada afirmación sobre una población se confirma o se rechaza

*El matrimonio es la principal causa de divorcio*




*Groucho* Marx
Julius Henry Marx

E.U.I.T.I. Bilbao

Asignatura:
MÉTODOS ESTADÍSTICOS
DE LA INGENIERÍA

# TEMA 5:
# DISTRIBUCIONES MUESTRALES


*La estadística en comic*
L. Gocking, W. Smith
(2002)

*Pocas observaciones y mucho razonamiento conducen al error; muchas observaciones y poco razonamiento, a la verdad.*

http://es.wikipedia.org/wiki/Alexis_Carrel

Alexis Carrel
*Premio Nobel en Medicina (1912)*

# 1. RESUMEN

Se llama la atención acerca de que los estadísticos muestrales son, en realidad, variables aleatorias. Así, se analiza la distribución de probabilidad de la media muestral y de la varianza muestral en diversas situaciones.

**Palabras clave**:

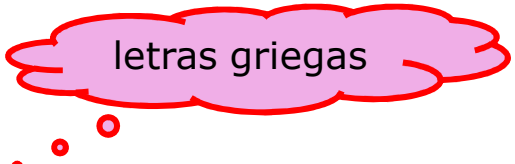
- distribuciones en el muestreo
- *t* de Student
- *F* de Snedecor
- *chi–cuadrado*

# 2. ÍNDICE DEL TEMA

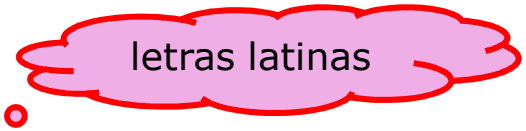
# 2. ÍNDICE DEL TEMA

# 3. INTRODUCCIÓN

- **población**: conjunto de todos los elementos objeto de un estudio
- **muestra**: subconjunto de una población

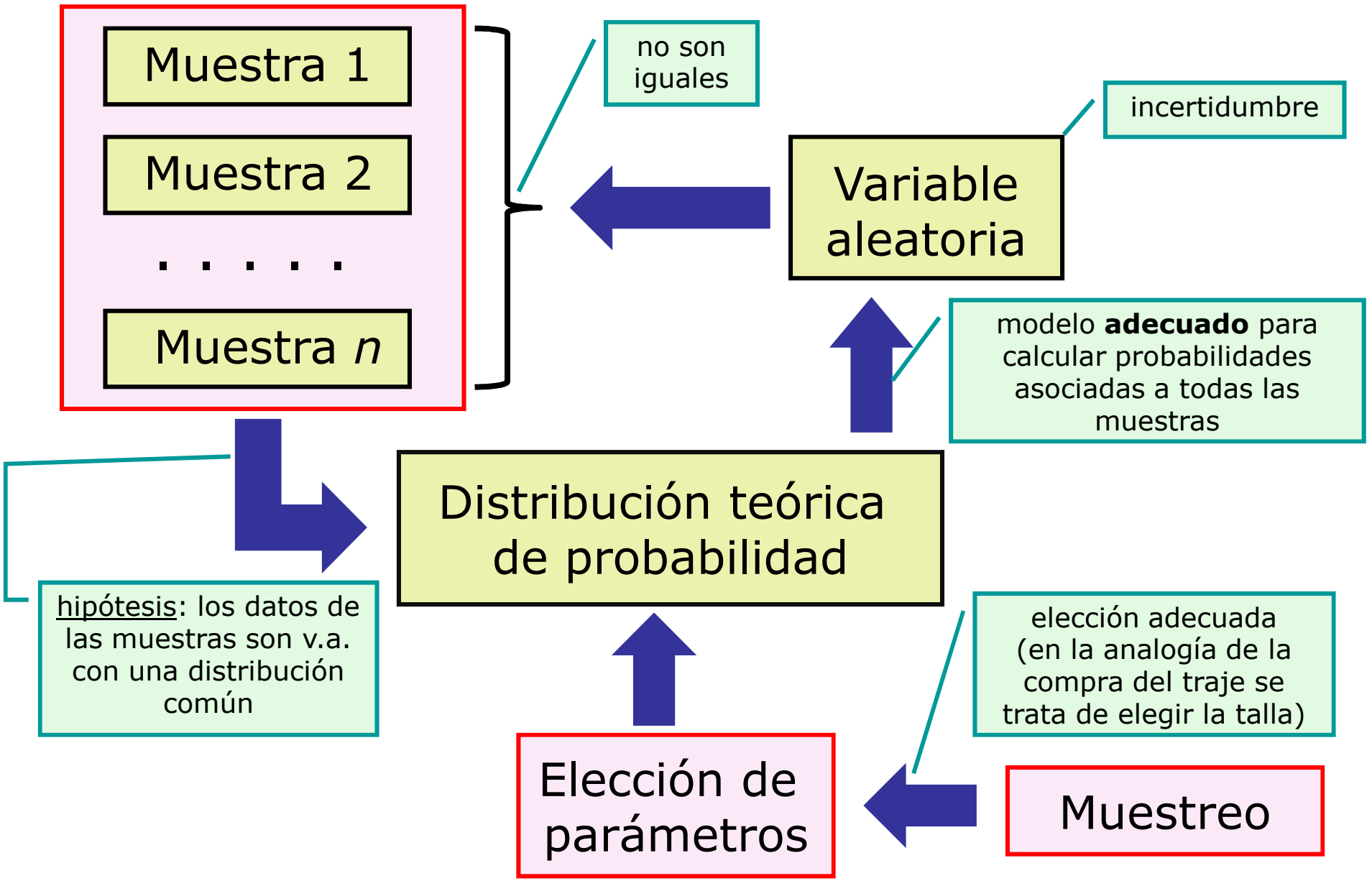



- **parámetro**: es una magnitud numérica obtenida de una población (es una cantidad fija, generalmente desconocida)

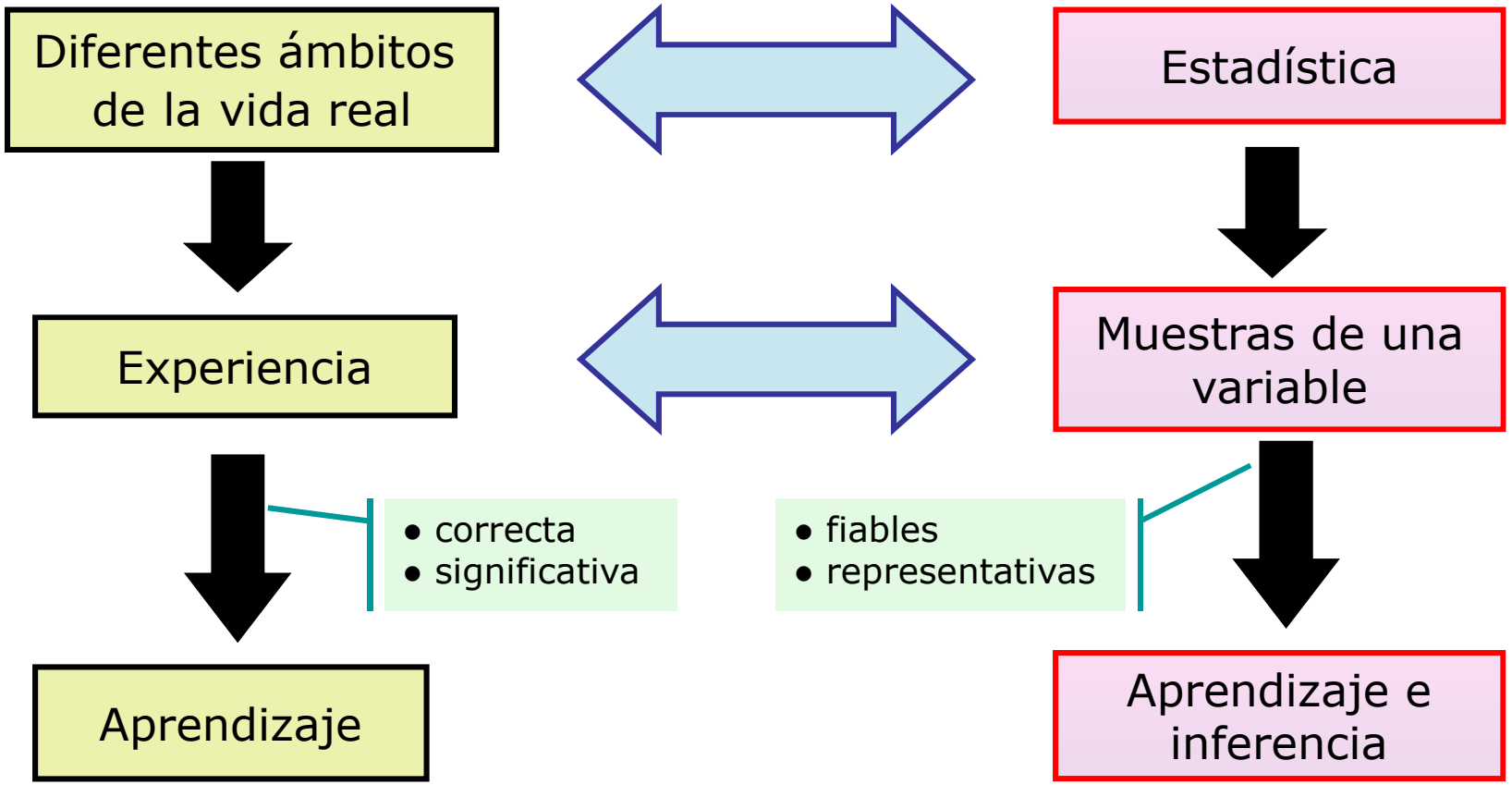



- **estadístico**: es una magnitud numérica cuyo valor viene determinado por la muestra (media muestral y varianza muestral, por ejemplo); se utilizan para estimar los parámetros de una población

# 3. INTRODUCCIÓN



# 3. INTRODUCCIÓN



# 3. INTRODUCCIÓN

- la mayor parte de las veces, la distribución de la población no es conocida completamente
- se usará una muestra para realizar inferencias sobre ella
- Ejemplo. Se fabrica un nuevo tipo de baterías para coches eléctricos
  - durarán un número aleatorio de kilómetros siguiendo una distribución de probabilidad desconocida
  - para averiguar cuál es esa distribución de probabilidad se puede fabricar un cierto número de baterías y probarlas en carretera
  - los datos resultantes (número de kms. recorridos con cada batería) constituyen una muestra extraída de la distribución

# 4. MUESTREO

**Muestreo** : conjunto de técnicas empleadas para la selección de una muestra a partir de una población

- al seleccionar una muestra, se espera que sus propiedades se puedan extrapolar a la población
- este proceso permite ahorrar recursos, obteniéndose resultados parecidos a los que se tendrían en el caso de realizar un estudio de toda la población
- **tipos**
  - muestreo no aleatorio o de juicio: se emplea el conocimiento y la opinión personal para identificar aquellos elementos de la población que deben incluirse en la muestra
  - muestreo aleatorio o de probabilidad: todos los elementos de la población tienen la oportunidad de ser escogidos

# 5. MUESTREO ALEATORIO: TIPOS

**Muestra aleatoria simple**: recopilación de datos de una v.a. *X* de una determinada población mediante la repetición del experimento al que está asociada.
Dos condiciones básicas:

- todos los elementos de la población deben tener la misma probabilidad de estar en la muestra
- las distintas observaciones de la muestra deben ser independientes entre sí

- ventajas
  - sencillo y de fácil comprensión, cálculo rápido de media y varianza, existen paquetes informáticos para su análisis
- desventajas
  - debe conocerse de antemano el listado de toda la población, muestras pequeñas pueden no ser representativas

# 5. MUESTREO ALEATORIO: TIPOS

**Muestreo estratificado**: la población se divide en grupos homogéneos (estratos) y después se toma una muestra aleatoria simple de cada estrato cuyo tamaño depende de un número o cuota asignado. Criterios para la elección de la cuota:

- proporcional al tamaño relativo del estrato en la población
- proporcional a la variabilidad del estrato , de forma que los más variables están más representados

- ventajas
  - tiende a asegurar que la muestra represente a la población de forma adecuada, estimaciones más precisas
- desventajas
  - debe conocerse la distribución en la población de las variables usadas para la estratificación

# 5. MUESTREO ALEATORIO: TIPOS

**Muestreo por conglomerado**: la población se divide en conglomerados (clusters) de elementos y, luego, se selecciona una muestra aleatoria de estos clusters.

- la variabilidad dentro de cada grupo es grande y entre los grupos es pequeña
- es como si cada grupo fuese un pequeña representación de la población en si misma

- ventajas
  - eficiente cuando la población es grande y dispersa, reduce costes, no es necesario el listado completo de la población
- desventajas
  - el error estándar es mayor que en el muestreo estratificado ó en el aleatorio y, además, su cálculo es complejo

# 5. MUESTREO ALEATORIO: TIPOS

## Conglomerado vs. estratificación

- muestreo por conglomerado
  - en la muestra sólo está representado un subconjunto de todos los conglomerados
  - variabilidad grande en cada grupo y pequeña entre los grupos
  - funciona si hay pocas diferencias entre los *clusters* y son muy heterogéneos
- muestreo estratificado
  - en la muestra están representados todos los estratos
  - variabilidad pequeña en cada estrato y grande entre ellos
  - funciona tanto mejor cuanto mayor sean las diferencias entre los estratos y más homogéneos sean internamente

# 5. MUESTREO ALEATORIO: TIPOS

**Muestreo sistemático**: los elementos de la muestra se seleccionan de la población de una forma uniforme medida respecto al tiempo, al orden, al espacio, …

- procedimiento:
  - listado de los $N$ elementos de la población
  - determinar el tamaño, $n$, de la muestra
  - definir un intervalo de salto, $k$ tal que: $k \approx \frac{N}{n}$ $donde$ $k \in Z$
  - arranque aleatorio, $r$: $1 \leq r \leq k$
  - hasta completar la muestra, se toman los elementos:

$$r+k,\ r+2k,\ r+3k,\ \ldots$$

# 5. MUESTREO ALEATORIO: TIPOS

**Muestreo sistemático**: los elementos de la muestra se seleccionan de la población de una forma uniforme medida respecto al tiempo, al orden, al espacio, …

- ventajas
  - fácil de aplicar, no siempre es necesario tener un listado de toda la población, cuando la población está ordenada según una tendencia conocida se asegura una cobertura de elementos de todos los tipos
- desventajas
  - sensible a las posibles periodicidades en el listado de los elementos de la población con lo que puede obtenerse una muestra que no sea representativa de la población

# 5. DISTRIBUCIONES MUESTRALES

- para obtener información sobre un parámetro de la población:
  - se selecciona una muestra representativa de la población
  - se obtiene el estadístico adecuado de la muestra
  - con ese estadístico se estima el parámetro de la población
- los valores que toma el estadístico en cada una de las muestras de tamaño *n* ($x_1$, $x_2$, … , $x_n$) son, a su vez, v.a. independientes que siguen la misma distribución de probabilidad (**distribución poblacional**)
- la muestra se usará para inferir sobre esa distribución y, al menos, aproximar su forma

# 5. DISTRIBUCIONES MUESTRALES

- para valorar el grado de validez de las inferencias es necesario conocer la distribución de probabilidad del estadístico considerado
- esa distribución de probabilidad se denomina distribución muestral
- la distribución muestral depende de:
  - la distribución de la población
  - el estadístico utilizado
  - el tamaño muestral

# 5. DISTRIBUCIONES MUESTRALES

- se observa una v.a. $X$
  - muestra aleatoria simple 1:

$$\left\{x_1^1, x_2^1, \cdots, x_n^1\right\} \Rightarrow \begin{cases} \text{media} : \bar{x}_1 \\ \text{desviación típica} : s_1 \end{cases}$$

  - muestra aleatoria simple 2:

$$\left\{x_1^2, x_2^2, \cdots, x_n^2\right\} \Rightarrow \begin{cases} \text{media} : \bar{x}_2 \\ \text{desviación típica} : s_2 \end{cases}$$

  - muestra aleatoria simple $n$:

$$\left\{x_1^n, x_2^n, \cdots, x_n^n\right\} \Rightarrow \begin{cases} \text{media} : \bar{x}_n \\ \text{desviación típica} : s_n \end{cases}$$

# 5. DISTRIBUCIONES MUESTRALES

- los valores resultantes en cada muestra son resultado del azar; entonces, generalmente, no coinciden
- para cada muestra se obtienen diferentes medias y desviaciones típicas
- por tanto, la media muestral y la varianza muestral (en general, cualquier estadístico de una muestra aleatoria simple) son, en realidad variables aleatorias y, como tales, tienen su distribución, su media, su varianza, …

$$\{\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n\} \qquad \{s_1, s_2, \cdots, s_n\}$$

muestras de dichas variables

# 5. DISTRIBUCIONES MUESTRALES

- Recordatorio

**Estadístico muestral:** valor referido a una muestra de una variable aleatoria (media, varianza, …)

**Parámetro poblacional:** valor referido a la distribución poblacional de una variable aleatoria (media, varianza, …)

- Asociados a estos dos conceptos se tiene:

**Distribución muestral** de un estadístico muestral es su distribución de probabilidad

**Error estándar** de un estadístico muestral es la desviación típica de su distribución muestral

# 5. DISTRIBUCIONES MUESTRALES

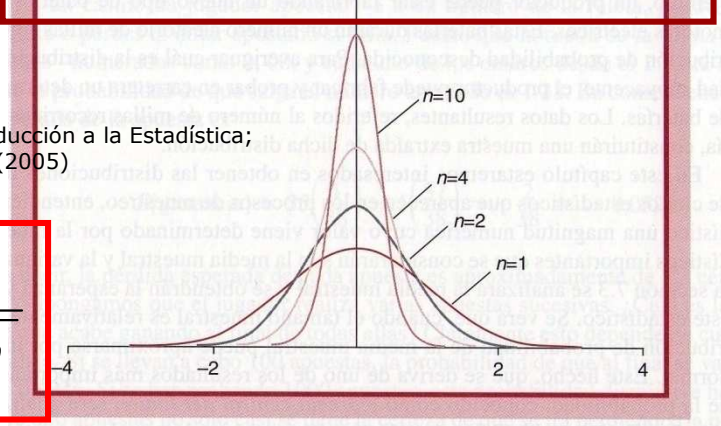
- Problema
  - bastante difícil conocer la distribución en el muestreo de los estadísticos muestrales
- Solución
  - si la variable que se observa sigue una distribución normal se conocen de forma exacta las distribuciones en el muestreo de los dos parámetro más importantes: la media y la varianza (es el caso más sencillo y, a su vez, el más importante)
  - si la variable observada no sigue una distribución normal: el teorema central del límite indica que si una variable es suma de otras variables su distribución es aproximadamente normal con lo que, todavía, se puede confiar en que lo que hagamos para variables normales pueda ser válido

# 6. MEDIA MUESTRAL

- sea una muestra de tamaño *n* ($x_1$, $x_2$, … , $x_n$) de una población de tamaño *N*
  - media poblacional: $\mu$
  - varianza poblacional: $\sigma^2$
  - media muestral: $\overline{X} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

- $\overline{X}$ : también es una v.a.; se demuestra que:
  - esperanza: $E\left[\overline{X}\right] = \mu$
  - varianza: $Var\left[\overline{X}\right] = \dfrac{\sigma^2}{n}$
  - desviación típica: $SD\left[\overline{X}\right] = \sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$

La dispersión disminuye a medida que aumenta el tamaño muestral, *n*




Densidades de las medias muestrales procedentes de una población normal estándar.

Ross, M.S.; Introducción a la Estadística; Ed. Reverté S.A. (2005)

# 6. MEDIA MUESTRAL

- **<u>Ejemplo de comprobación</u>.** Sea una población formada por los números:

$$\{x_1 = 2, x_2 = 4, x_3 = 6, x_4 = 8\} \rightarrow N=4$$

$$\mu = \frac{1}{N}\sum_{i=1}^{i=4} x_i = \frac{2+4+6+8}{4} = 5$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{i=4} x_i^{\,2} - \bar{x}^{\,2}} = \sqrt{5} = 2.236$$

- se consideran todas las muestras posibles de tamaño n=2

| | ELEMENTO 1 | ELEMENTO 2 | MEDIA | | ELEMENTO 1 | ELEMENTO 2 | MEDIA |
|---|---|---|---|---|---|---|---|
| MUESTRA 1 | 2 | 2 | 2 | MUESTRA 9 | 6 | 2 | 4 |
| MUESTRA 2 | 2 | 4 | 3 | MUESTRA 10 | 6 | 4 | 5 |
| MUESTRA 3 | 2 | 6 | 4 | MUESTRA 11 | 6 | 6 | 6 |
| MUESTRA 4 | 2 | 8 | 5 | MUESTRA 12 | 6 | 8 | 7 |
| MUESTRA 5 | 4 | 2 | 3 | MUESTRA 13 | 8 | 2 | 5 |
| MUESTRA 6 | 4 | 4 | 4 | MUESTRA 14 | 8 | 4 | 6 |
| MUESTRA 7 | 4 | 6 | 5 | MUESTRA 15 | 8 | 6 | 7 |
| MUESTRA 8 | 4 | 8 | 6 | MUESTRA 16 | 8 | 8 | 8 |

# 6. MEDIA MUESTRAL

- **<u>Ejemplo de comprobación</u>.**

  - la media de las medias muestrales es:

$$\overline{X} = \frac{1}{16}\sum_{i=1}^{i=16} \overline{x}_i = \frac{2+3+4+5+\ \cdots\ +6+8}{16} = 5 \quad \Rightarrow \quad \overline{X} = \mu$$

  - la desviación típica de las medias muestrales es:

$$\sigma_{\overline{X}} = \sqrt{\frac{1}{N}\sum_{i=1}^{i=16} \overline{x}_i^{\,2} - \overline{x}^{\,2}} = \sqrt{2.5} = 1.581$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.236}{\sqrt{2}} = 1.581$$

# 6. MEDIA MUESTRAL

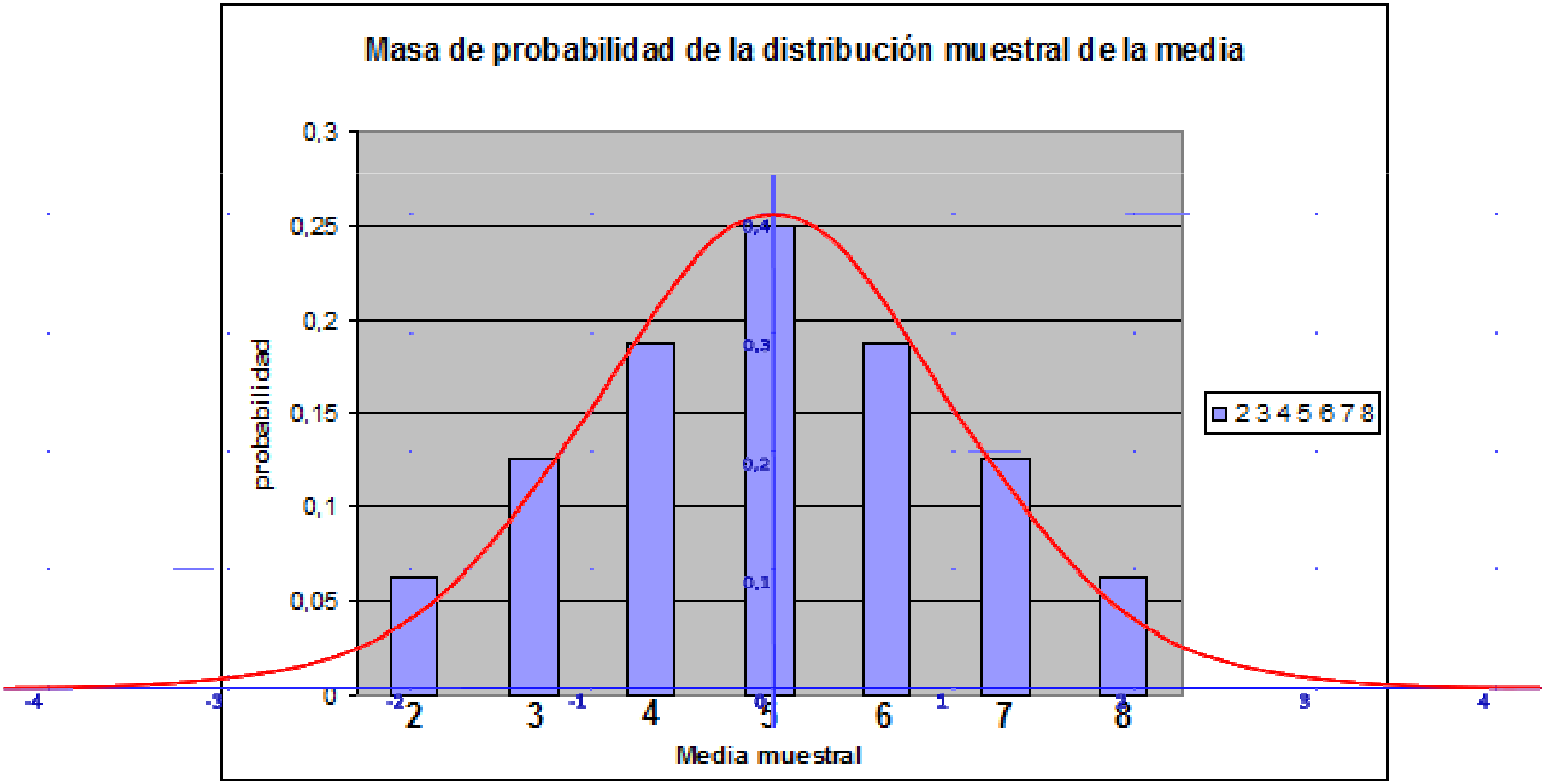
- **<u>Ejemplo de comprobación</u>.**
  - función de masa de probabilidad de la distribución muestral de la media muestral:

| $\bar{x}_i$ | 2 | 3 | 4 |
|---|---|---|---|
| $p(\bar{x}_i) = P(\bar{X} = \bar{x}_i)$ | $\frac{1}{16} = 0.0625$ | $\frac{2}{16} = 0.125$ | $\frac{3}{16} = 0.1875$ |

| $\bar{x}_i$ | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| $p(\bar{x}_i) = P(\bar{X} = \bar{x}_i)$ | $\frac{4}{16} = 0.25$ | $\frac{3}{16} = 0.1875$ | $\frac{2}{16} = 0.125$ | $\frac{1}{16} = 0.0625$ |

# 6. MEDIA MUESTRAL

- **<u>Ejemplo de comprobación</u>.**
  - representación gráfica de la función de masa de probabilidad de la distribución muestral de la media muestral:



# 6. MEDIA MUESTRAL

- en la afirmación: $SD\left[\overline{X}\right]=\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}$ se asume que:
  - la población es infinita (suficientemente grande), ó
  - el muestreo se realiza con reemplazamiento
- en caso contrario, se debe utilizar un *factor de corrección para poblaciones finitas*:

$$SD\left[\overline{X}\right]=\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$$

  - $N$: tamaño de la población
  - $n$: tamaño de la muestra

# 7. TEOREMA CENTRAL DEL LÍMITE

- sea $X_1$, $X_2$, … , $X_n$ una muestra aleatoria procedente de una población de media $\mu$ y desviación típica $\sigma$
- si $n$ tiende a infinito (para valores suficientemente grandes de $n$) la suma:

$$Y = \sum_{i=1}^{i=n} X_i$$

sigue aproximadamente una distribución normal

$$Y \approx N\left(n \cdot \mu; \sqrt{n} \cdot \sigma\right)$$

# 7. TEOREMA CENTRAL DEL LÍMITE

▶ **Ejemplo**. Una compañía aseguradora de automóviles tiene 10000 asegurados. Si el gasto anual que un asegurado ocasiona a la compañía tiene por media 260 euros con una desviación típica de 800 euros, aproximar la probabilidad de que el gasto total que la compañía debe afrontar en un año sobrepase los 2,8 millones de euros

**Solución**

- $X_i$ : gasto ocasionado por el asegurado $i$ ($i$=1,2, … , 10000)
- por el teorema central del límite

$$\left.\begin{array}{l} n\cdot\mu = 2.6\times10^6 \\ \sigma\cdot\sqrt{n} = 8\times10^4 \end{array}\right\} \Rightarrow \sum_{i=1}^{i=n} X_i = Y \approx N\left(2.6\times10^6 ; 8\times10^4\right)$$

# 7. TEOREMA CENTRAL DEL LÍMITE

- **Ejemplo**. Una compañía aseguradora de automóviles tiene 10000 asegurados. Si el gasto anual que un asegurado ocasiona a la compañía tiene por media 260 euros con una desviación típica de 800 euros, aproximar la probabilidad de que el gasto total que la compañía debe afrontar en un año sobrepase los 2,8 millones de euros

**Solución**

- entonces:

tipificación

$$P\left(Y > 2.8\times10^6\right) = P\left(Z > \frac{2.8\times10^6 - 2.6\times10^6}{8\times10^4}\right) \approx P\left(Z > \frac{0.2\times10^6}{8\times10^4}\right) =$$

$$= P\left(Z > 2.5\right) = 1 - P\left(Z \le 2.5\right) = 1 - 0.9938 = 0.0062$$

tablas

***Excel***: =1-DISTR.NORM(2800000;2600000;80000;1)

# 7. TEOREMA CENTRAL DEL LÍMITE

- <u>generalización</u>
  - se puede demostrar que, aunque las v.a. $X_1$, $X_2$, … , $X_n$ sigan distribuciones distintas, la variable $Y$ sigue aproximadamente una distribución normal siendo, como antes:

$$Y = \sum_{i=1}^{i=n} X_i$$

  - es más, si todas las variables aleatorias tienden a ser, más ó menos, de la misma magnitud, de forma que ninguna de ellas domine el valor de la suma, se puede demostrar que la suma de un gran número de variables aleatorias independientes sigue, aproximadamente, una distribución normal

# 7. TEOREMA CENTRAL DEL LÍMITE

**Distribución de la media muestral**

- sea $X_1$, $X_2$, … , $X_n$ una muestra aleatoria procedente de una población de media $\mu$ y desviación típica $\sigma$
- sea la media muestral: $\overline{X} = \frac{1}{n} \sum_{i=1}^{i=n} X_i = \frac{1}{n} \cdot Y$
- el producto de una normal, $Y$, por una constante, $\frac{1}{n}$ , sigue siendo normal: $X \approx N(\mu; \sigma) \Rightarrow k \cdot X \approx N(k \cdot \mu; k \cdot \sigma)$
- entonces, si el tamaño muestral es grande, por el teorema central del límite se tiene: $\overline{X} \approx N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$

# 7. TEOREMA CENTRAL DEL LÍMITE

**Distribución de la media muestral**

- tipificación

$$\overline{X} \approx N\left(\mu; \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = Z \approx N(0;1)$$

- además:

$$P\left(\overline{X} \le a\right) = P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{a-\mu}{\sigma/\sqrt{n}}\right) \approx P\left(Z \le \frac{a-\mu}{\sigma/\sqrt{n}}\right)$$

# 7. TEOREMA CENTRAL DEL LÍMITE

## Distribución de la media muestral

▶ **Ejemplo**. El nivel de colesterol en la sangre de una población de trabajadores tiene media 202 mg/dl y desviación típica 14.

**a.** si se selecciona una muestra de 36 trabajadores, aproximar la probabilidad de que la media muestral de sus niveles de colesterol esté comprendida entre 198 y 206 mg/dl

**b.** repetir el apartado anterior para un tamaño muestral de 64

**Solución**

**a.** del teorema central del límite se deduce, como $n$=36:

$$\overline{X} \approx N\left(202; \frac{7}{3}\right) \Rightarrow \frac{\overline{X} - 202}{7/3} = Z \approx N(0;1)$$

# 7. TEOREMA CENTRAL DEL LÍMITE

## Distribución de la media muestral

▶ **Ejemplo**.

**Solución**

**a.** $$P\left(198 \le \overline{X} \le 206\right) = P\left(\frac{198-202}{7/3} \le \frac{\overline{X}-202}{7/3} \le \frac{206-202}{7/3}\right) \approx$$

tipificación

$$\approx P\left(-1.714 \le Z \le 1.714\right) = \Phi(1.714) - \Phi(-1.714) =$$

$$= \Phi(1.714) - \left[1 - \Phi(1.714)\right] = 2 \cdot \Phi(1.714) - 1 = 0.9128$$

tablas

**b.** procediendo de forma análoga para n=64:

$$P\left(198 \le \overline{X} \le 206\right) \approx 0.9780$$

**Nota**. Al aumentar el tamaño muestral también aumenta la probabilidad de que la media muestral difiera de la poblacional en menos de 4 unidades

# 7. TEOREMA CENTRAL DEL LÍMITE

## Distribución de la media muestral

▶ **Ejemplo**.

**Solución.** Sin tipificar, usando ***Excel***

**a.** $P\left(198 \leq \overline{X} \leq 206\right) \approx 0{,}91352373 =$

=DISTR.NORM(206;202;7/3;1)-DISTR.NORM(198;202;7/3;1)

**b.** $P\left(198 \leq \overline{X} \leq 206\right) \approx 0{,}97772902 =$

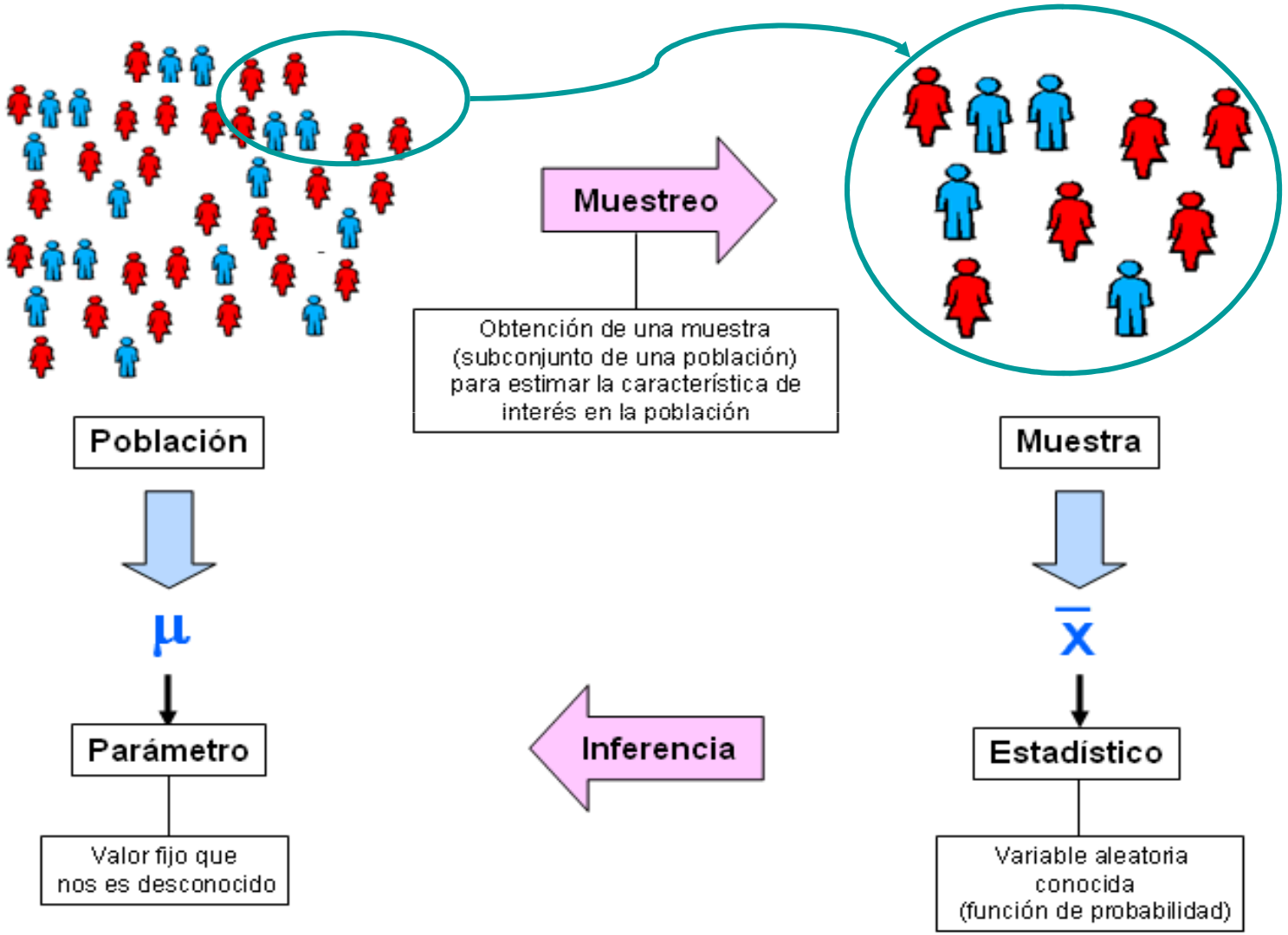
=DISTR.NORM(206;202;7/4;1)-DISTR.NORM(198;202;7/4;1)

# 7. TEOREMA CENTRAL DEL LÍMITE

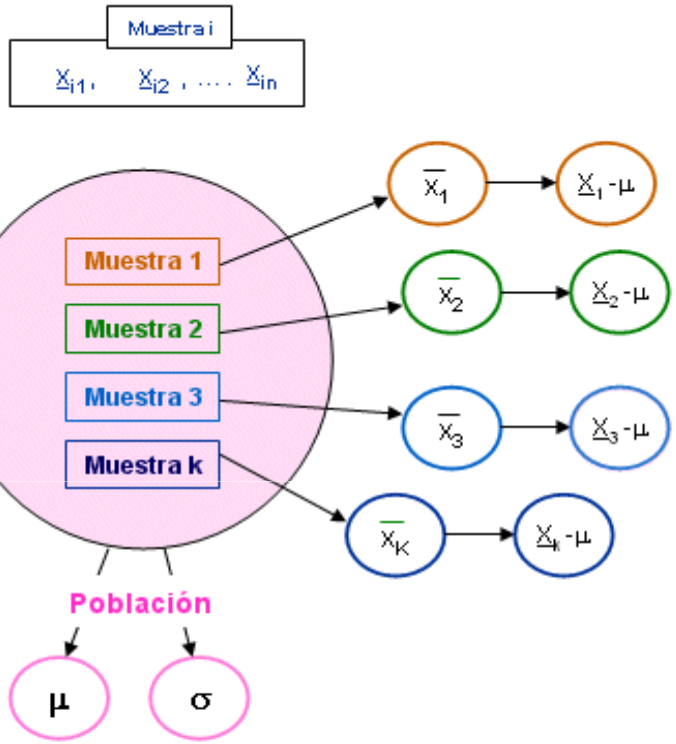
## Tamaño de la muestra

- el teorema central del límite deja abierta la cuestión de cuán grande debe ser el tamaño, *n*, de la muestra para que la aproximación normal sea válida
- respuesta: depende de la distribución de la población que subyace a los datos muestrales
  - si la distribución subyacente es normal, la media muestral es siempre normal independientemente del tamaño de la muestra
  - regla empírica: puede usarse la aproximación normal siempre que el tamaño muestral sea, como mínimo, 30 ($n \geq 30$)
  - en la mayoría de los casos la aproximación normal es válida para tamaños muestrales mucho más reducidos

# 8. RESUMEN


Muestreo
Obtención de una muestra (subconjunto de una población) para estimar la característica de interés en la población
Población
Muestra
μ
x̄
Parámetro
Inferencia
Estadístico
Valor fijo que nos es desconocido
Variable aleatoria conocida (función de probabilidad)

# 8. RESUMEN



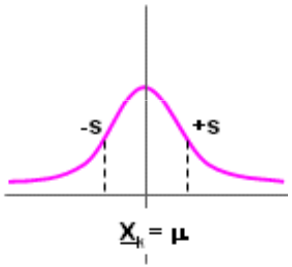

Muestra i = $(\underline{X}_{i1}, \underline{X}_{i2}, \ldots, \underline{X}_{in})$

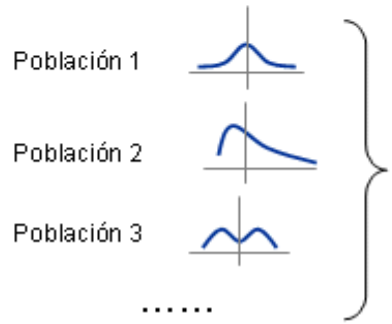
$$\underline{X}_i = \frac{\underline{X}_{i1} + \underline{X}_{i2} + \ldots + \underline{X}_{in}}{n}$$

$$\underline{X}_{\underline{X}} = \frac{\underline{X}_1 + \underline{X}_2 + \ldots + \underline{X}_k}{k} = \mu$$

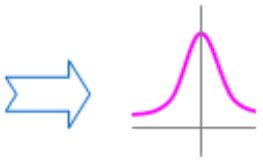
$$s_{\underline{X}} = \sqrt{\frac{\sum_1^k (\underline{X}_i - \mu)^2 n_i}{\Sigma n_i}}$$

$$s_{\underline{X}} = \frac{\sigma}{\sqrt{n}}$$

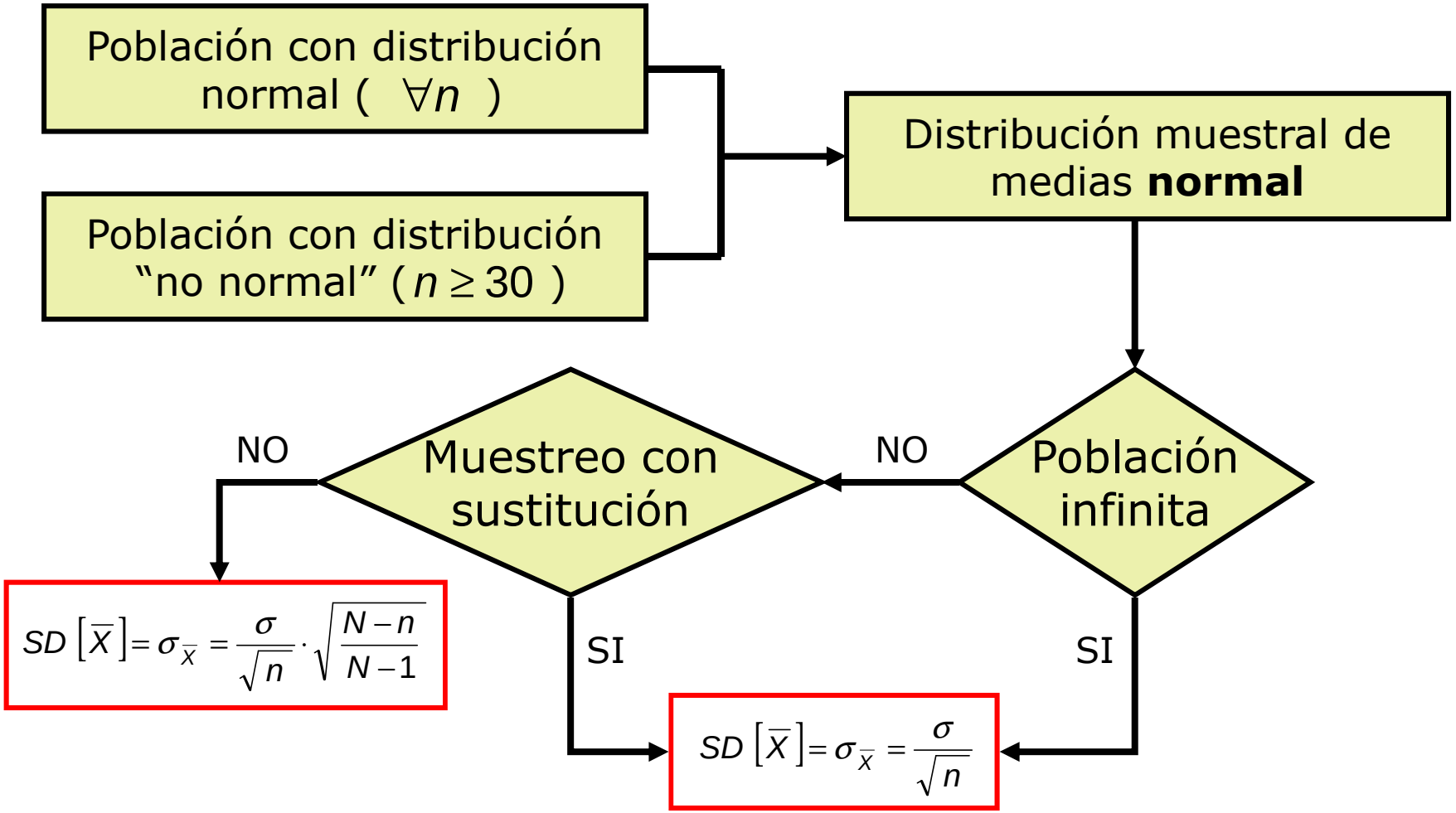



Independientemente de la distribución poblacional la distribución muestral de medias. si **n>30**, es siempre "**normal**"



# 8. RESUMEN


Población con distribución normal ( $\forall n$ )
Población con distribución "no normal" ( $n \geq 30$ )
Distribución muestral de medias **normal**
Población infinita
NO
Muestreo con sustitución
NO
SI
SI
$SD\left[\overline{X}\right] = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$
$SD\left[\overline{X}\right] = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

# 8. RESUMEN

- **Ejemplo**. En un municipio, el consumo medio por vivienda de energía eléctrica es de 25Kwh con una desviación típica de 6kwh. Se toma una muestra de 50 viviendas:

  **a.** obtener la distribución muestral de las medias de consumo de la muestra en Kwh

  **b.** calcular la probabilidad de que el consumo medio del grupo de viviendas sea superior a 26Kwh

**Solución**

**a.** del teorema central del límite se deduce, como $n=50$:

- $\mu=25$
- $\sigma=6$
- $n=50$

$$\bar{X} \approx N\left(25; \frac{6}{\sqrt{50}}\right) = N\left(25; 0.849\right)$$

# 8. RESUMEN

- **Ejemplo**. En un municipio, el consumo medio, por vivienda, de energía eléctrica es de 25Kwh con una desviación típica de 6kwh. Se toma una muestra de 50 viviendas:

  **a.** obtener la distribución muestral de las medias de consumo de la muestra en Kwh

  **b.** calcular la probabilidad de que el consumo medio del grupo de viviendas sea superior a 26Kwh

**Solución**

**b.** del teorema central del límite se deduce, como $n=50$:

$$P\left(\overline{X}>26\right)=P\left(\frac{\overline{X}-25}{0.849}>\frac{26-25}{0.849}\right)\approx P\left(Z>1.178\right)=$$

$$=1-P\left(Z\leq 1.178\right)=1-0.881=0.119$$

***Excel***: =1-DISTR.NORM(26;25;6/RAIZ(50);1)~ 0,11929641

# 8. RESUMEN

▶ **Ejemplo**. El peso del azúcar envasado en paquetes es una variable aleatoria normal de media 400gr. y desviación típica 30gr. Se toma una muestra aleatoria simple de 25 paquetes. Calcular la probabilidad de que el promedio quede fuera del intervalo (390,410).

**Solución**

- del teorema central del límite se deduce, como $n$=25:

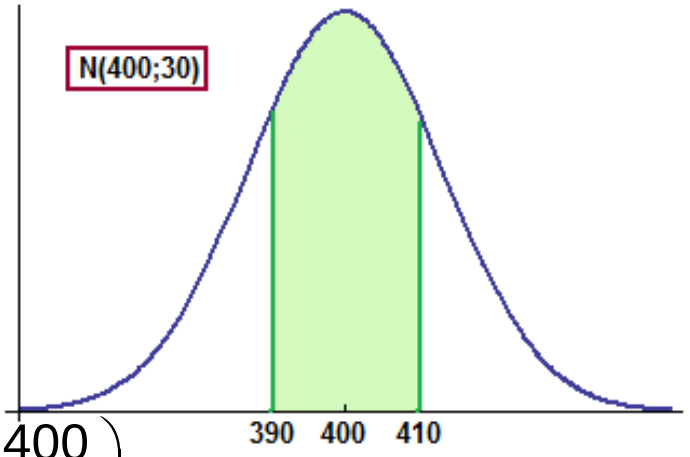
$$\overline{X} \approx N\left(\mu; \frac{\sigma}{\sqrt{n}}\right) = N\left(400; \frac{30}{\sqrt{25}}\right) = N\left(400; 6\right)$$

  - ◇ media poblacional: $\mu$=400
  - ◇ desviación típica de la población: $\sigma$=30
  - ◇ tamaño de la muestra: $n$=25

# 8. RESUMEN

- **Ejemplo**. El peso del azúcar envasado en paquetes es una variable aleatoria normal de media 400gr. y desviación típica 30gr. Se toma una muestra aleatoria simple de 25 paquetes. Calcular la probabilidad de que el promedio quede fuera del intervalo (390,410).

**Solución**




- se pide calcular: $P\left(\overline{X} \notin (390, 410)\right)$

$$P\left(\overline{X} \notin (390, 410)\right) = 1 - P\left(\overline{X} \in (390, 410)\right) =$$

$$= 1 - P\left(390 < \overline{X} < 410\right) = 1 - P\left(\frac{390-400}{6} < Z < \frac{410-400}{6}\right) =$$

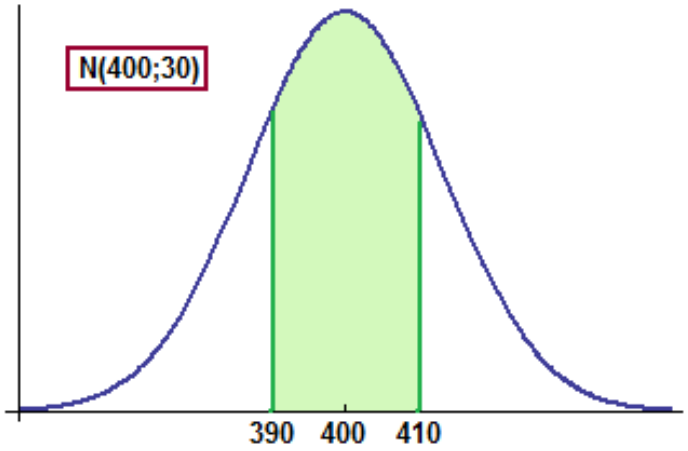
$$= 1 - P\left(-1.66666667 < Z < 1.666666667\right) = 0.0955807$$

# 8. RESUMEN

- **Ejemplo**. El peso del azúcar envasado en paquetes es una variable aleatoria normal de media 400gr. y desviación típica 30gr. Se toma una muestra aleatoria simple de 25 paquetes. Calcular la probabilidad de que el promedio quede fuera del intervalo (390,410).

**Solución**

- se pide calcular: $P\left(\overline{X} \notin (390, 410)\right)$




$$P\left(\overline{X} \notin (390, 410)\right) = 1 - P\left(\overline{X} \in (390, 410)\right) \approx 0{,}0955807$$

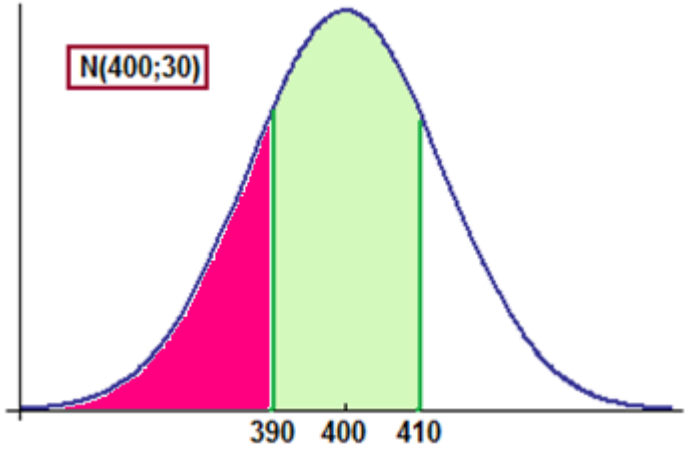
***Excel***: =1-(DISTR.NORM(410;400;6;1)-DISTR.NORM(390;400;6;1))

# 8. RESUMEN

▶ **Ejemplo**. El peso del azúcar envasado en paquetes es una variable aleatoria normal de media 400gr. y desviación típica 30gr. Se toma una muestra aleatoria simple de 25 paquetes. Calcular la probabilidad de que el promedio quede fuera del intervalo (390,410).

**Solución**

- se pide calcular: $P\left(\overline{X} \notin (390, 410)\right)$




$$P\left(\overline{X} \notin (390, 410)\right) = 2 \cdot P\left(\overline{X} < 390\right) \approx 0{,}0955807$$

***Excel***: =2*DISTR.NORM(390;400;6;1)

# 9. DISTRIBUCIONES ASOCIADAS A LA NORMAL

- cuando se va a hacer inferencia estadística se ha visto que la distribución normal aparece de forma casi inevitable.
- dependiendo del problema, se pueden encontrar otras (asociadas):
  - *ji*-cuadrado
  - *t* de Student
  - *F* de Fisher-Snedecor

Sus funciones de densidad son muy poco tratables por lo que los cálculos se realizan con ordenador (también, muchos valores de sus funciones de distribución están tabulados)

- resultan directamente de operar con distribuciones normales
- típicamente aparecen como distribuciones de ciertos estadísticos

# 10. DISTRIBUCIÓN *JI*-CUADRADO

Si $Z_1, Z_2, \ldots, Z_k$ son variables aleatorias normales estándar e independientes entonces se dice que la variable

$$X = \sum_{i=1}^{i=k} Z_i^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

es una variable aleatoria ***ji*-cuadrado** (ó *chi*-cuadrado) con $k$ grados de libertad

- notación: $\chi_k^2 \approx \sum_{i=1}^{i=k} Z_i^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2$ ó $X \approx \chi_k^2$
- parámetro: $k$ (número de grados de libertad)
- es una de las distribuciones más utilizadas en estadística inferencial (describe la distribución muestral de la varianza muestral)

# 10. DISTRIBUCIÓN *JI*-CUADRADO

- ¿qué son los grados de libertad?
  - se pueden definir como el número de valores que se pueden elegir libremente
  - ejemplo: sea una muestra de tamaño $n=2$, los valores de la muestra son $x$ e $y$, y se sabe que tienen una media de 10

$$\frac{x+y}{2}=10 \Rightarrow x+y=20 \Rightarrow \begin{cases} x=0 \Rightarrow y=20 \\ x=1 \Rightarrow y=19 \\ \dots \quad \dots \quad \dots \\ x=20 \Rightarrow y=0 \end{cases}$$

$$si \quad x=5 \Rightarrow y=15$$

*y* no es libre de tomar cualquier valor

  - este ejemplo se puede generalizar para cualquier $n$ en donde dada la media de los valores sólo quedan ($n$-1) elementos que pueden definirse libremente y uno es función de la media y el resto de los elementos

# 10. DISTRIBUCIÓN *JI*-CUADRADO

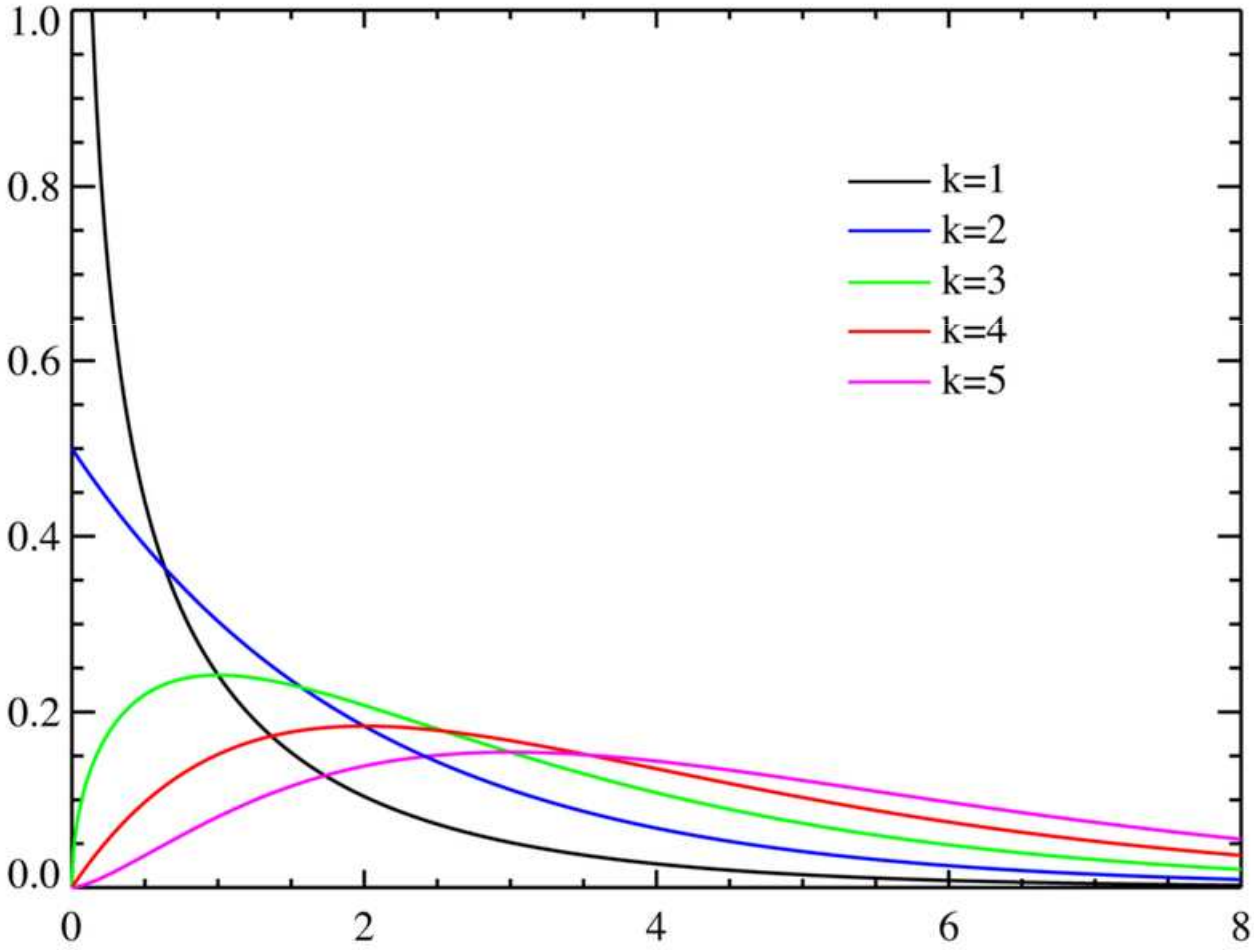
- **función de densidad**

$$f(x)=\frac{x^{\frac{k}{2}-1}\cdot e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\cdot\Gamma\left(\frac{k}{2}\right)} \quad (x\geq 0)$$

  - función gamma: $\Gamma(x)=\int_0^{\infty} t^{x-1}\cdot e^{-t}\,dt$
  - función asimétrica positiva, sólo tienen densidad los valores positivos (suma de cuadrados)
  - se hace más simétrica, incluso casi gaussiana, al aumentar el número de grados de libertad.
  - normalmente se consideran anómalos aquellos valores de la variable de la **cola de la derecha**

# 10. DISTRIBUCIÓN *JI*-CUADRADO

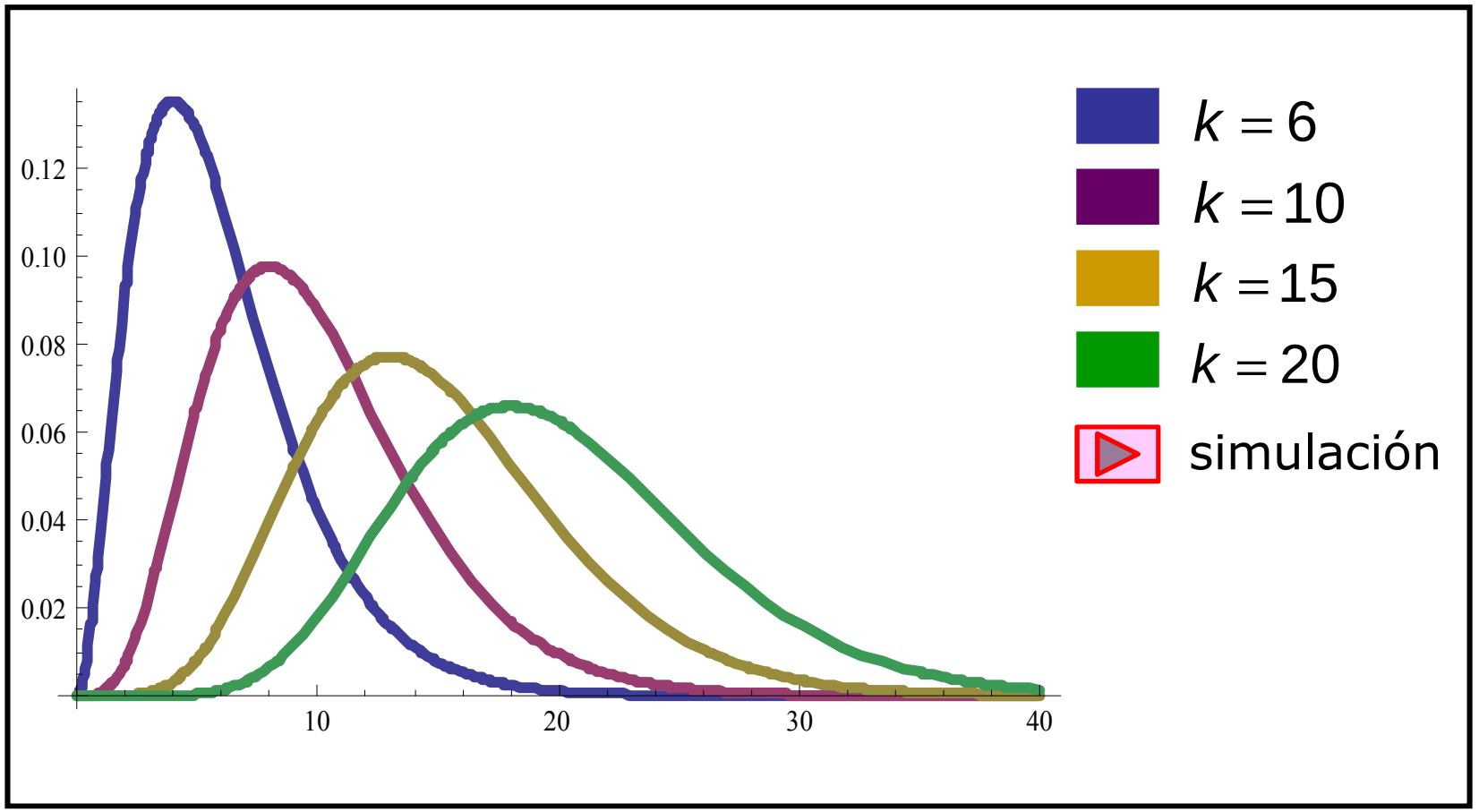
- **<u>función de densidad</u>**




http://es.wikipedia.org/wiki/Distribución_$\chi^2$

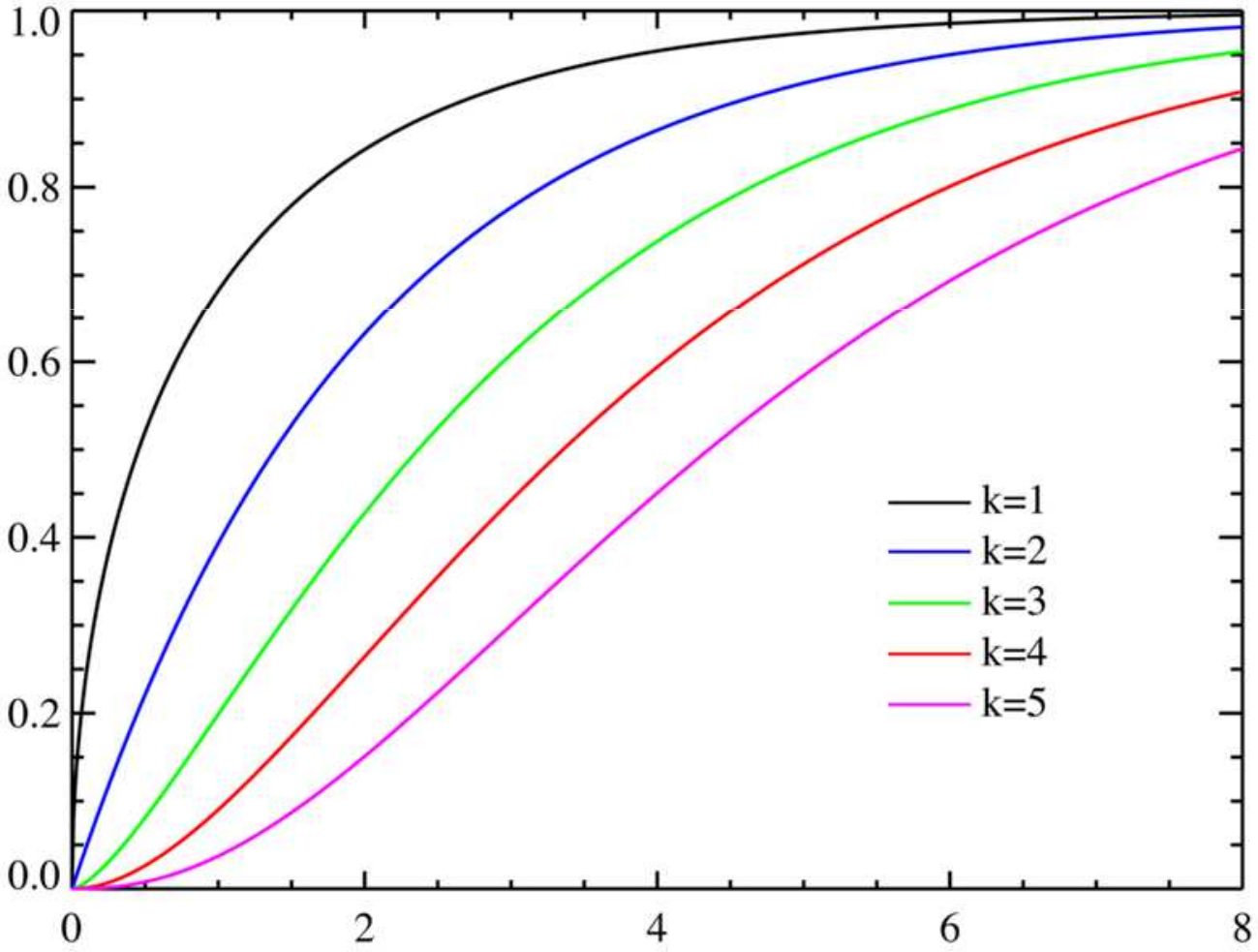
# 10. DISTRIBUCIÓN *JI*-CUADRADO

- **función de densidad**



# 10. DISTRIBUCIÓN *JI*-CUADRADO

- **función de distribución acumulada**




http://es.wikipedia.org/wiki/Distribución_$\chi^2$

# 10. DISTRIBUCIÓN *JI*-CUADRADO

- esperanza matemática (media):

$$E\left[\chi_k^2\right] = k$$

- varianza:

$$Var\left[\chi_k^2\right] = 2k$$

- desviación típica:

$$SD\left[\chi_k^2\right] = \sqrt{2k}$$

# 10. DISTRIBUCIÓN *JI*-CUADRADO

- **<u>propiedades</u>:**

  - la suma de dos v.a. independientes y continuas *ji*-cuadrado es también una v.a. continua *ji*-cuadrado cuyo número de grados de libertad es la suma de los grados de libertad de las dos variables:

$$\chi_k^2 + \chi_n^2 = \chi_{k+n}^2$$

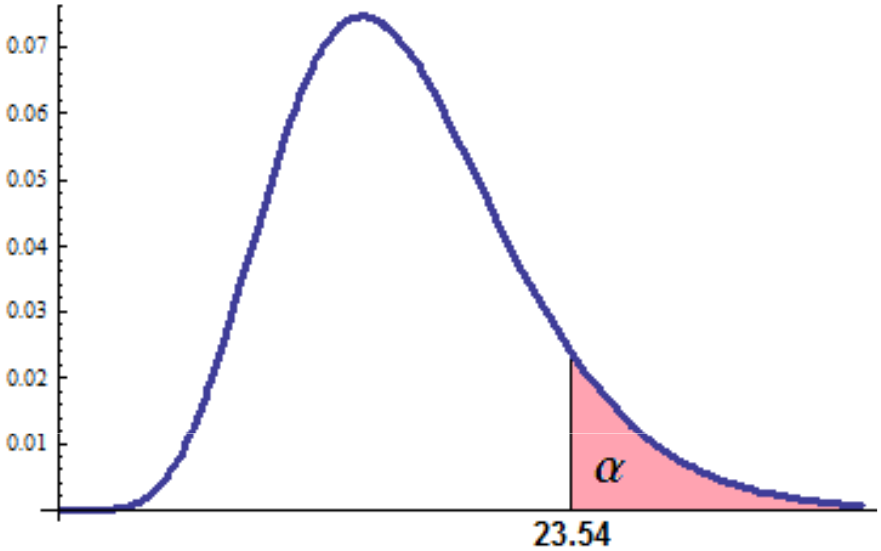
  - por el teorema central del límite, cuando $k>30$:

$$\sqrt{2 \cdot \chi_k^2} \approx N\left(\sqrt{2k-1}\,;1\right)$$

# 10. DISTRIBUCIÓN *JI*-CUADRADO

- **ejemplo**
  - $P\left(\chi^2_{16} \leq 23.54183\right) = 0.90$

| n | 0.30 | 0.20 | 0.10 |
|---|---|---|---|
| 1 | 1.07420 | 1.64238 | 2.70554 |
| 2 | 2.40794 | 3.21888 | 4.60518 |
| 3 | 3.66487 | 4.64163 | 6.25139 |
| 4 | 4.87843 | 5.98862 | 7.77943 |
| 5 | 6.06443 | 7.28927 | 9.23635 |
| 6 | 7.23113 | 8.55806 | 10.64464 |
| 7 | 8.38343 | 9.80325 | 12.01703 |
| 8 | 9.52446 | 11.03009 | 13.36156 |
| 9 | 10.65637 | 12.24214 | 14.68366 |
| 10 | 11.78072 | 13.44196 | 15.98717 |
| 11 | 12.89867 | 14.63142 | 17.27501 |
| 12 | 14.01110 | 15.81199 | 18.54934 |
| 13 | 15.11872 | 16.98479 | 19.81193 |
| 14 | 16.22209 | 18.15077 | 21.06414 |
| 15 | 17.32169 | 19.31065 | 22.30712 |
| 16 | 18.41789 | 20.46507 | 23.54182 |




- $P\left(\chi^2_{16} > 23.54183\right) = 0.10$

# 10. DISTRIBUCIÓN *JI*-CUADRADO

- **ejemplo.** Calcular: $P\left(\chi^2_{12} \le 20\right)$
  - tablas

$$P\left(\chi^2_{12} \le 18.54934\right) = 0.900$$

$$P\left(\chi^2_{12} \le 21.02606\right) = 0.950$$

| n | α | | | |
|---|---|---|---|---|
| | 0.30 | 0.20 | 0.10 | 0.05 |
| 1 | 1.07420 | 1.64238 | 2.70554 | 3.84146 |
| 2 | 2.40794 | 3.21888 | 4.60518 | 5.99148 |
| 3 | 3.66487 | 4.64163 | 6.25139 | 7.81472 |
| 4 | 4.87843 | 5.98862 | 7.77943 | 9.48773 |
| 5 | 6.06443 | 7.28927 | 9.23635 | 11.07048 |
| 6 | 7.23113 | 8.55806 | 10.64464 | 12.59158 |
| 7 | 8.38343 | 9.80325 | 12.01703 | 14.06713 |
| 8 | 9.52446 | 11.03009 | 13.36156 | 15.50731 |
| 9 | 10.65637 | 12.24214 | 14.68366 | 16.91896 |
| 10 | 11.78072 | 13.44196 | 15.98717 | 18.30703 |
| 11 | 12.89867 | 14.63142 | 17.27501 | 19.67515 |
| 12 | 14.01110 | 15.81199 | 18.54934 | 21.02606 |

  - el valor pedido no se encuentra en las tablas, por tanto, debe realizarse una interpolación lineal
  - ***Excel***: =1-DISTR.CHI(20;12)=0,93291404

# 10. DISTRIBUCIÓN *JI*-CUADRADO

- **ejemplo.** Calcular: $P\left(\chi^2_{12} \le 20\right)$
  - interpolación lineal: parte de dos puntos conocidos de una función y determina los valores intermedios con la recta que une estos dos puntos

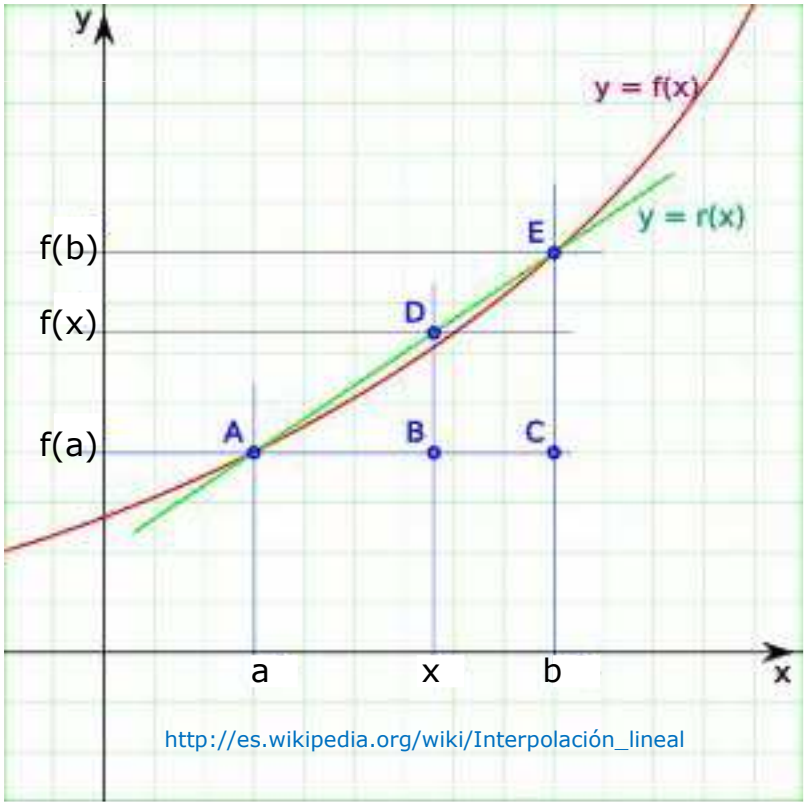
$$f(x) \approx \frac{f(b)-f(a)}{b-a}\cdot(x-a)+f(a) \quad (a<x<b)$$

  - en este caso:

$$\left.\begin{array}{c}(a, f(a)) = (18.54934, 0.900) \\ (b, f(b)) = (21.02606, 0.950) \\ x = 20\end{array}\right\}$$

  - operando:

$$P\left(\chi^2_{12} \le 20\right) \approx f(x) = 0.929286$$




http://es.wikipedia.org/wiki/Interpolación_lineal

# 10. DISTRIBUCIÓN *JI*-CUADRADO

- **ejemplo.** Calcular: $\chi_0^2$ *tal que* $P\left(\chi_{20}^2 \geq \chi_0^2\right) = 0.100$

  $\chi_0^2$ *tal que* $P\left(\chi_{23}^2 \leq \chi_0^2\right) = 0.950$

  - tablas

$$\chi_0^2 = \chi_{20;\,0.100}^2 = 28.41197$$

  - tablas

$$P\left(\chi_{23}^2 \leq \chi_0^2\right) = 0.950 \Rightarrow P\left(\chi_{23}^2 > \chi_0^2\right) = 0.05$$

$$\chi_0^2 = \chi_{23;\,0.050}^2 = 35.17246$$

| n | α | | | |
|---|---|---|---|---|
| | 0.30 | 0.20 | 0.10 | 0.05 |
| 1 | 1.07420 | 1.64238 | 2.70554 | 3.84146 |
| 2 | 2.40794 | 3.21888 | 4.60518 | 5.99148 |
| 3 | 3.66487 | 4.64163 | 6.25139 | 7.81472 |
| 4 | 4.87843 | 5.98862 | 7.77943 | 9.48773 |
| 5 | 6.06443 | 7.28927 | 9.23635 | 11.07048 |
| 6 | 7.23113 | 8.55806 | 10.64464 | 12.59158 |
| 7 | 8.38343 | 9.80325 | 12.01703 | 14.06713 |
| 8 | 9.52446 | 11.03009 | 13.36156 | 15.50731 |
| 9 | 10.65637 | 12.24214 | 14.68366 | 16.91896 |
| 10 | 11.78072 | 13.44196 | 15.98717 | 18.30703 |
| 11 | 12.89867 | 14.63142 | 17.27501 | 19.67515 |
| 12 | 14.01110 | 15.81199 | 18.54934 | 21.02606 |
| 13 | 15.11872 | 16.98479 | 19.81193 | 22.36203 |
| 14 | 16.22209 | 18.15077 | 21.06414 | 23.68478 |
| 15 | 17.32169 | 19.31065 | 22.30712 | 24.99580 |
| 16 | 18.41789 | 20.46507 | 23.54182 | 26.29622 |
| 17 | 19.51102 | 21.61456 | 24.76903 | 27.58710 |
| 18 | 20.60135 | 22.75955 | 25.98942 | 28.86932 |
| 19 | 21.68913 | 23.90042 | 27.20356 | 30.14351 |
| 20 | 22.77454 | 25.03750 | 28.41197 | 31.41042 |
| 21 | 23.85779 | 26.17109 | 29.61509 | 32.67056 |
| 22 | 24.93901 | 27.30145 | 30.81329 | 33.92446 |
| 23 | 26.01837 | 28.42879 | 32.00689 | 35.17246 |
| 24 | 27.09596 | 29.55332 | 33.19624 | 36.41503 |

# 10. DISTRIBUCIÓN *JI*-CUADRADO

- **ejemplo.** Calcular: $P\left(\chi^2_{41} > 32\right)$

  - aproximación con la normal ($k$=41)

$$\sqrt{2\cdot\chi^2_{41}} \approx N\left(\sqrt{2(41)-1}\,;1\right) = N\left(9;1\right)$$

$$P(\chi^2_{41} > 32) \approx P\left(\sqrt{2\chi^2_{41}} > \sqrt{2(32)} = 8\right) =$$

$$= P\left(Z > -1\right) = 0.8413$$

  - ***Excel***:

=DISTR.CHI(32;41)~0,84194476

# 11. DISTRIBUCIÓN *t* DE STUDENT

Si $Y, Z_1, Z_2, \ldots, Z_n$ son $n+1$ variables aleatorias normales estándar e independientes entonces se dice que la v.a.

$$X = \frac{Y}{\sqrt{\frac{1}{n}\sum_{i=1}^{i=n} Z_i^2}} = \frac{Y}{\sqrt{\frac{1}{n}\left(Z_1^2 + Z_2^2 + \cdots + Z_n^2\right)}}$$

sigue una distribución ***t* de Student** con $n$ grados de libertad

- notación: $X \approx t_n$
- parámetro: $n$ (número de grados de libertad)
- es una distribución que surge al estimar la media de una población con un tamaño muestral pequeño

# 11. DISTRIBUCIÓN *t* DE STUDENT

- es una distribución que surge al estimar la media de una población con un tamaño muestral pequeño
- además, normalmente se desconoce la desviación típica $\sigma$
- William Gosset "Student" define una nueva variable aleatoria, la *t*, utilizada en estos supuestos
  - en la variable: $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}} \approx N(0;1)$
  - estima (y sustituye) $\sigma$ utilizando la desviación típica muestral

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

# 11. DISTRIBUCIÓN *t* DE STUDENT

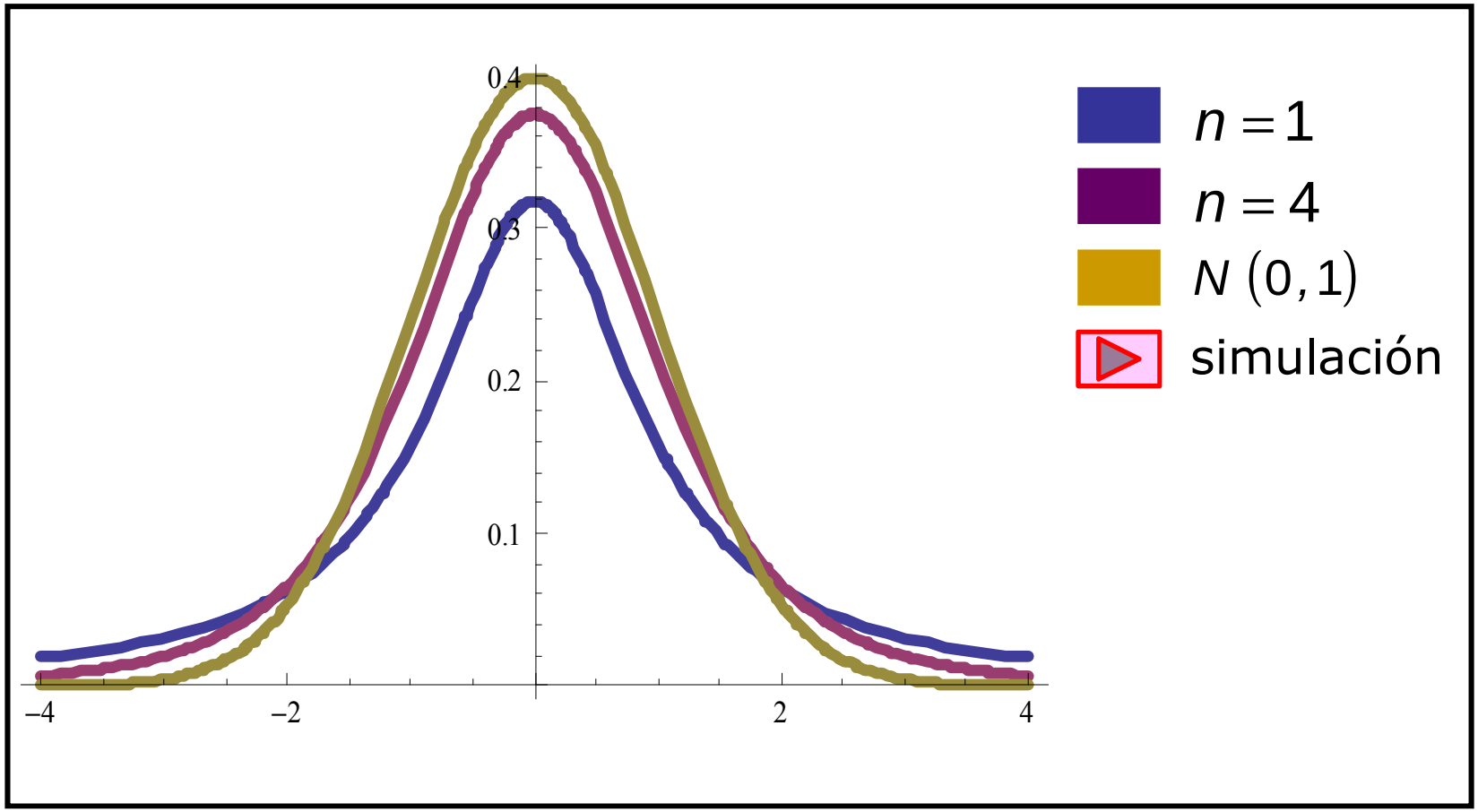
- **función de densidad**

$$f(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\cdot\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

- función gamma: $\Gamma(x)=\int_0^{\infty} t^{x-1}\cdot e^{-t}\,dt$
- simétrica con respecto al cero
- gráfica parecida a la de la normal estándar pero presenta mayor dispersión con lo que es más aplanada
- a medida que aumentan los grados de libertad, su gráfica más se acerca a la de la normal estándar N(0,1); para n≥30 la distribución *t* tiende a N(0,1)
- se consideran valores anómalos los que se alejan de cero (positivos o negativos)

# 11. DISTRIBUCIÓN *t* DE STUDENT

- **función de densidad**



# 11. DISTRIBUCIÓN *t* DE STUDENT

- esperanza matemática (media): $E\left[t_n\right]=0$
- varianza: $Var\left[t_n\right]=\frac{n}{n-2}$

# 12. DISTRIBUCIÓN *F* DE SNEDECOR

Si $Y_1, Y_2, \ldots, Y_n, Z_1, Z_2, \ldots, Z_m$ son $n$+m v.a. normales estándar e independientes entonces se dice que la v.a.

$$X = \frac{\frac{1}{n}\sum_{i=1}^{i=n} Y_i^2}{\frac{1}{m}\sum_{i=1}^{i=m} Z_i^2} = \frac{\frac{1}{n}\left(Y_1^2 + Y_2^2 + \cdots + Y_n^2\right)}{\frac{1}{m}\left(Z_1^2 + Z_2^2 + \cdots + Z_m^2\right)}$$

sigue una distribución ***F* de Snedecor** (ó *F* de Fisher ó *F* de Fisher-Snedecor con $n$ grados de libertad en el numerador y $m$ grados de libertad en el denominador

- notación: $X \approx F_{n;m}$
- parámetros: $n$, $m$ (grados de libertad)
- es una distribución que surge al estimar el cociente de varianzas de dos poblaciones normales

# 12. DISTRIBUCIÓN *F* DE SNEDECOR

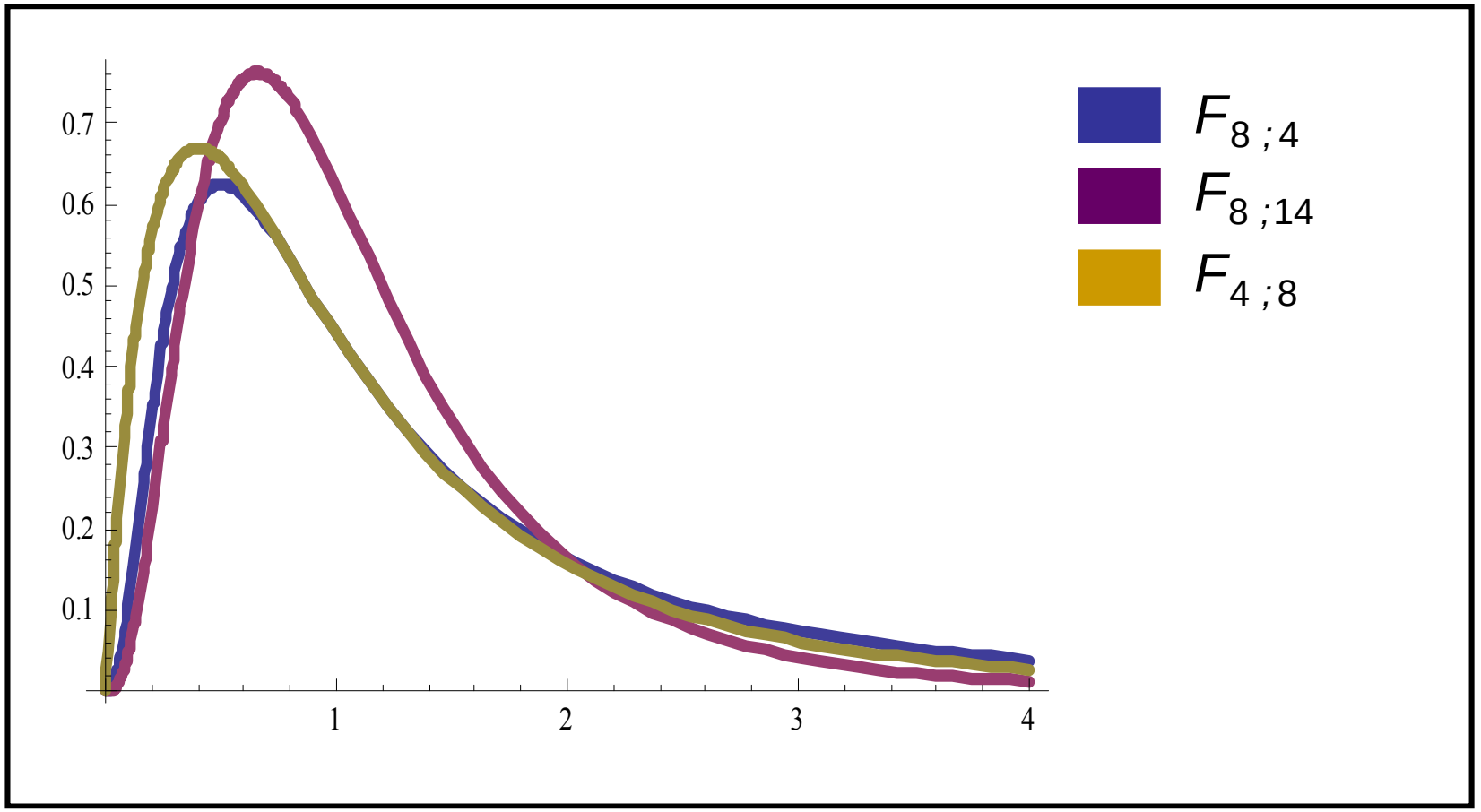
- **función de densidad**

$$f(x)=\frac{1}{B\left(\frac{n}{2},\frac{m}{2}\right)}\cdot\left(\frac{nx}{nx+m}\right)^{\frac{n}{2}}\cdot\left(1-\frac{nx}{nx+m}\right)^{\frac{m}{2}}\cdot x^{-1}$$

- función gamma: $B(x,y)=2\int_0^{\frac{\pi}{2}}\cos^{2x-1}\theta\cdot sen^{2y-1}\theta\,d\theta$
- sólo toma valores positivos
- es asimétrica
- normalmente, se consideran valores anómalos los de la cola de la derecha
- relación con *ji*-cuadrado: $F_{n;\infty}=\chi_n^2$

# 12. DISTRIBUCIÓN *F* DE SNEDECOR

- **función de densidad**



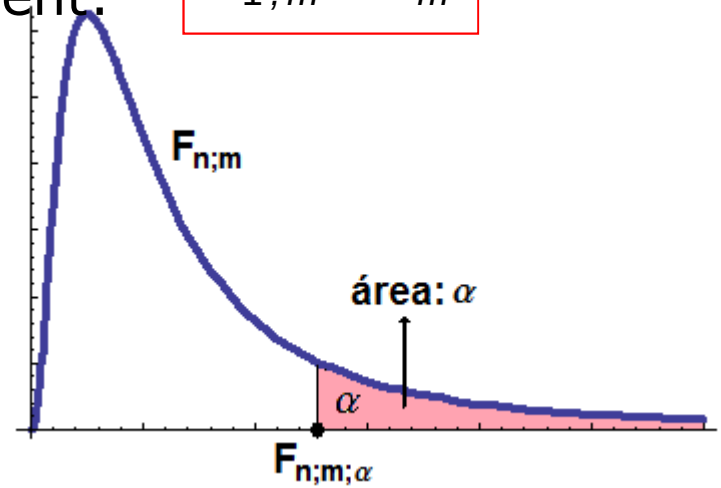
# 12. DISTRIBUCIÓN *F* DE SNEDECOR

- esperanza matemática (media): $E\left[F_{n;m}\right]=\dfrac{m}{m-2} \quad (m>2)$

- varianza: $Var\left[F_{n;m}\right]=\dfrac{2m^2(n+m-2)}{n\cdot(m-2)^2\cdot(m-4)} \quad (m>4)$

- propiedades:
  - relación con la distribución *ji*-cuadrado: $F_{n;\infty}=\chi^2_n$
  - relación con la distribución t de Student: $F_{1;m}=t_m$
  - si $F_{n;m;1-\alpha}$ es el valor que deja a su derecha una probabilidad $\alpha$:

$$F_{n;m;\alpha}=\frac{1}{F_{n;m;1-\alpha}} \qquad P\left(F_{n;m}>F_{n;m;\alpha}\right)=\alpha$$



# 13. DISTRIBUCIONES ASOCIADAS A LA NORMAL

- propiedad reproductiva de la distribución normal: si $X_1$, $X_2$, … , $X_n$ son variables aleatorias normales independientes, con $E[X_i]=\mu_i$ y $Var[X_i]=\sigma_i^2$, entonces:

$$a_1 \cdot X_1 + a_2 \cdot X_2 + \cdots + a_n \cdot X_n = Y \approx N\left( \sum_{i=1}^{i=n} a_i \cdot \mu_i \; ; \sqrt{\sum_{i=1}^{i=n} a_i^2 \cdot \sigma_i^2} \right)$$

# 13. DISTRIBUCIONES ASOCIADAS A LA NORMAL

- sea $X_1$, $X_2$, … , $X_n$ una muestra aleatoria simple de una población normal con media $\mu$ y varianza $\sigma^2$ y sea la cuasivarianza muestral:

$$\hat{s}^2 = s_{n-1}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}\left(X_i - \overline{X}\right)^2$$

- la variable aleatoria $\dfrac{(n-1)\cdot\hat{s}^2}{\sigma^2} = \dfrac{(n-1)\cdot S^2}{\sigma^2}$

sigue una distribución *ji*-cuadrado con $n$-1 grados de libertad

$$\frac{(n-1)\cdot\hat{s}^2}{\sigma^2} = \frac{(n-1)\cdot S^2}{\sigma^2} \approx \chi^2_{n-1}$$