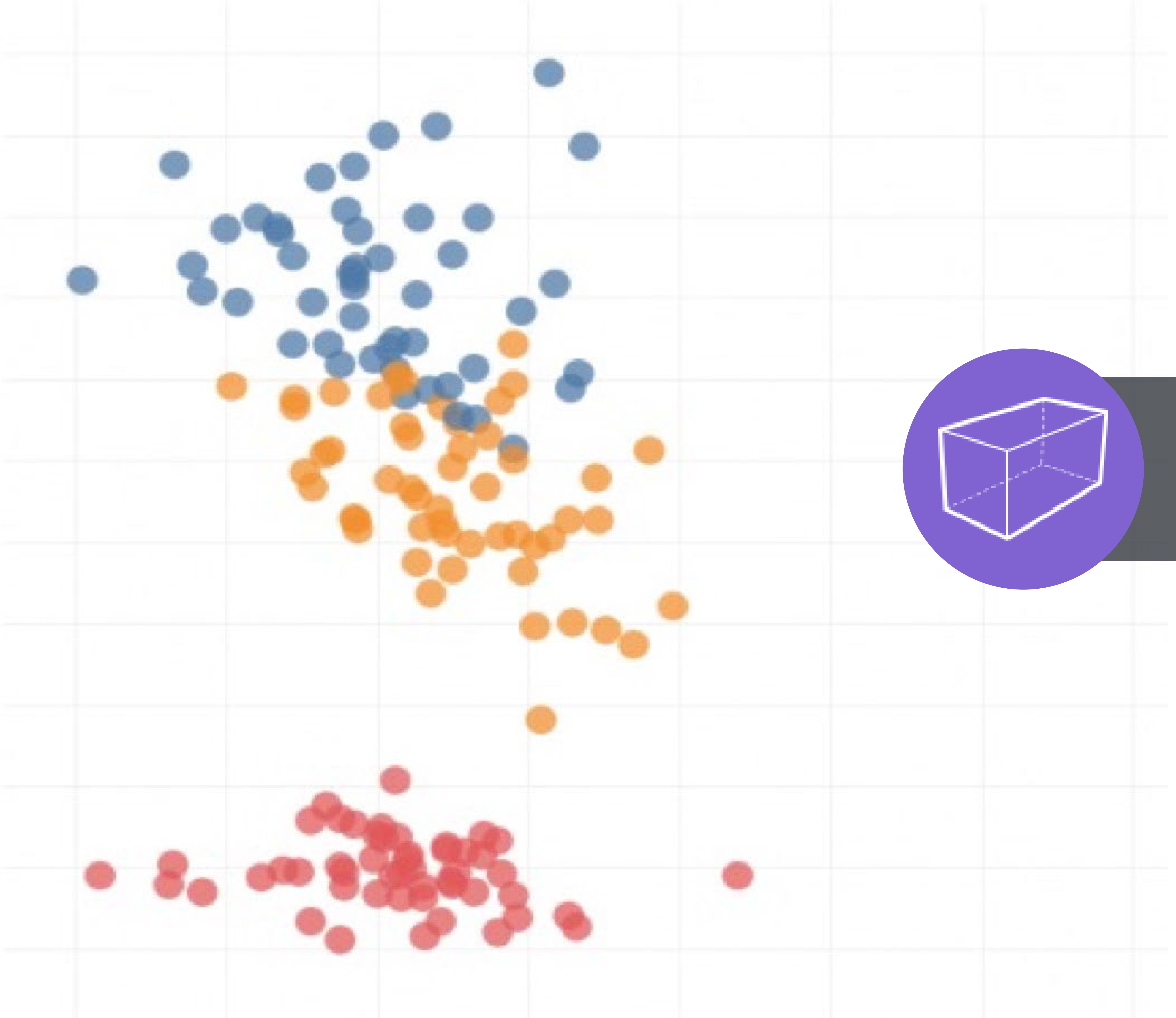
ALGES
ADVANCED LABORATORY FOR GEOSTATISTICAL SUPERCOMPUTING

Test de Hipótesis

Análisis Estadístico y Geoestadístico de Datos

Auxiliar: Fabián Soto F.
Profesor: Xavier Emery

Ingeniería de Minas
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE



Test de hipótesis

Resumen

Conceptos importantes

Test de Hipótesis → Valida o invalida la hipótesis

$\boldsymbol{H_0}$: Hipótesis nula

$\boldsymbol{H_1}$: Hipótesis alternativa

Conceptos importantes

Si $\boldsymbol{H_1}$ es del estilo $x_1 \neq x_2$ → Test bilateral

(α = 0.025)

Si $\boldsymbol{H_1}$ es del estilo:

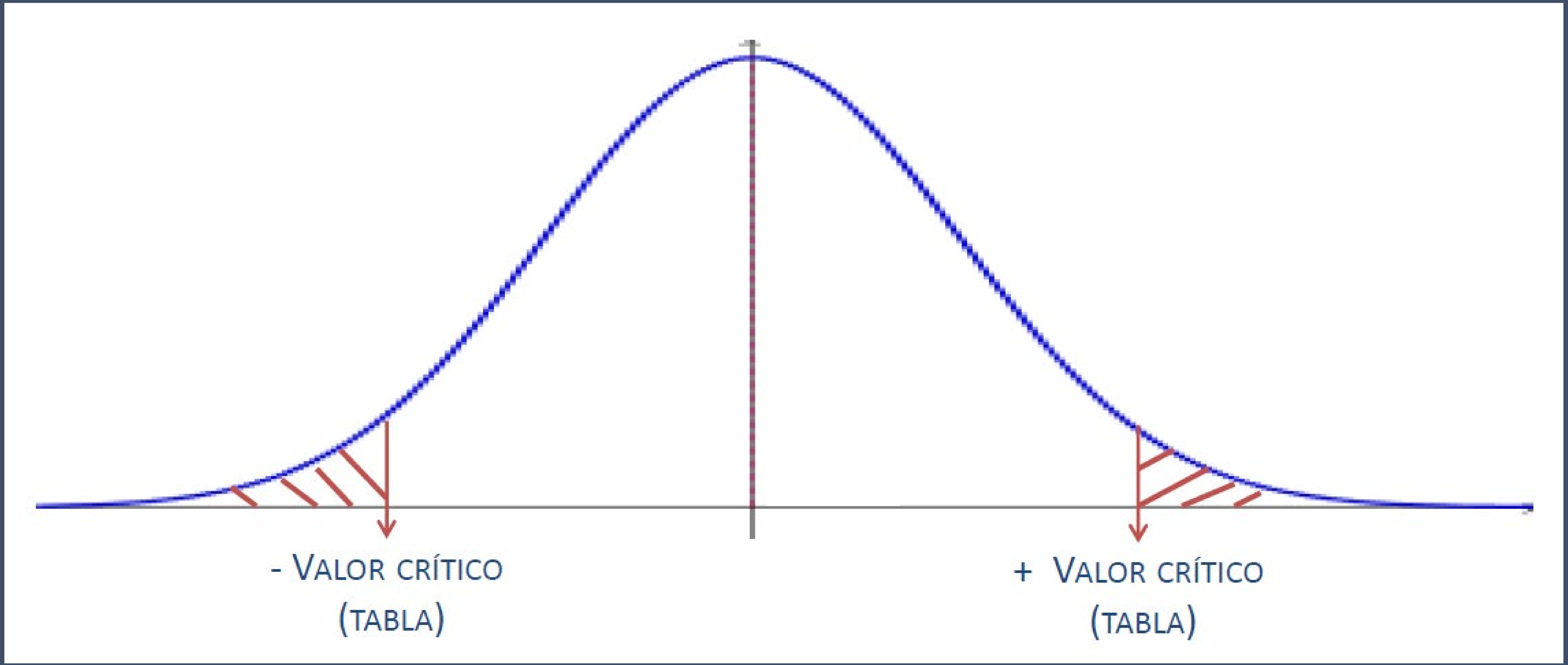
- $x_1 < x_2$
- $x_1 \leq x_2$
- $x_1 > x_2$
- $x_1 \geq x_2$

→ Test unilateral

(α = 0.05)

Propagación de errores

- ¿Cómo concluir al hacer un test de hipótesis?¿qué es el nivel de significancia?

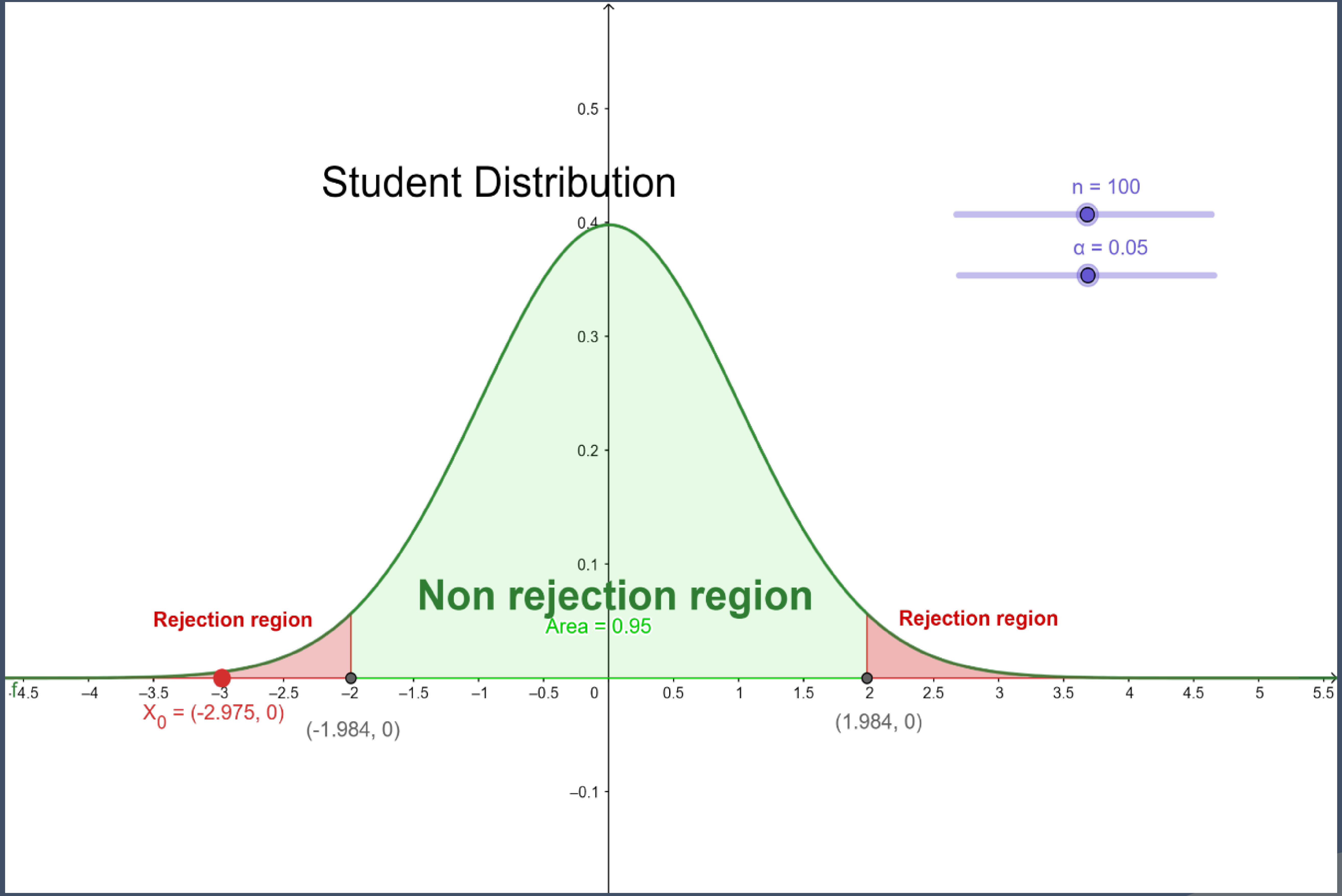


Rechaza H_0 ← → Rechaza H_0

No se puede Rechazar H_0

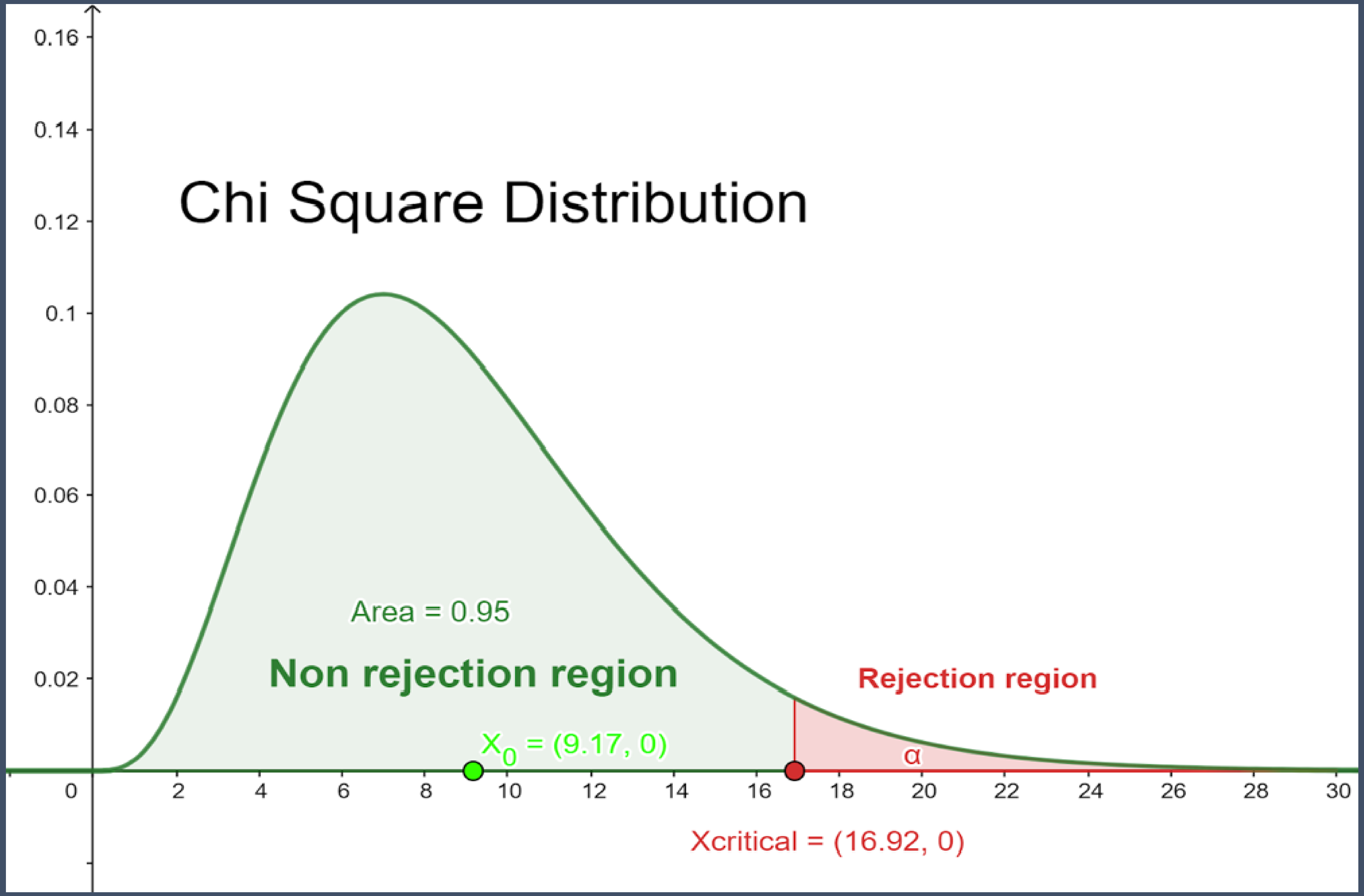
Test bilateral

- Aplica en distribuciones simétricas
- Si el nivel de confianza es del 95%, cada una de las partes que componen la región de rechazo deben ser 2.5%



Test unilateral

- Aplica en distribuciones simétricas o asimétricas
- Si el nivel de confianza es de 95%, la región de rechazo debe ser de un 5%



Pasos de un test de hipótesis

- El procedimiento general para efectuar un test de hipóstesis consiste en 4 etapas principales:
- Etapas:

1. Definir las hipótesis nula y alternativa
2. Definir/calcular el estadístico a analizar
3. Identificar la región de rechazo para un nivel de significancia
4. Aceptar o rechazar la hipótesis nula

Tests paramétricos:

- Sobre las medias y varianzas en datos gaussianos
- Sobre proporciones

Tests de ajustes de una distribución:

- Test de Kolmogorov y Cramer Von Mieses
- Test de Chi cuadrado

Tests de Outliers:

- Test de Grubbs
- Test de Chauvenet

Test Chi Cuadrado

Se tienen $\boldsymbol{k}$ datos agrupados en clases. Se busca comparar si el número de ocurrencias observadas de cierto parámetro en cada rango $\boldsymbol{O_i}$ difieren de los valores esperados $\boldsymbol{E_i}$. Bajo un determinado modelo de distribución. Con $i \in \{1, \ldots, k\}$

$\boldsymbol{H_0}$: El modelo esperado ajusta a los datos observados.

$\boldsymbol{H_1}$: El modelo esperado no ajusta a los datos observados

$$\chi^2 = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \frac{(\boldsymbol{O_i} - \boldsymbol{E_i})^2}{\boldsymbol{E_i}}$$

Se compara el estadístico con una distribución Chi Cuadrado de k-1 grados de libertad

Test Chi Cuadrado

Propiedades y límites de validez del test chi cuadrado:

1. Se trata de un test libre, independiente del modelo.

2. Idealmente, se requiere al menos 4 a 5 individuos por clase. En caso contrario se agrupan varias clases para cumplir esta condición.

3. Si además de la distribución se estiman l parámetros en forma independiente (media, varianza, etc.), entonces el chi cuadrado tiene k-1-l grados de libertad en lugar de k-1.

4. El rechazo de $\boldsymbol{H_0}$ es contundente. En cambio, la aceptación de $\boldsymbol{H_0}$ no es muy potente (se pueden aceptar varios modelos distintos, por falta de prueba para rechazarlos.

Test Chi Cuadrado de dos variables

- Se consideran observaciones clasificadas según dos variables categóricas (A y B) resumidas en una tabla de contingencia:

	$\boldsymbol{A_1}$	$\boldsymbol{A_2}$	…	$\boldsymbol{A_r}$	Total
$\boldsymbol{B_1}$	O_{11}	O_{12}	…	O_{1r}	$O_{1\cdot}$
$\boldsymbol{B_2}$	O_{21}	O_{22}	…	O_{2r}	$O_{2\cdot}$
…					
$\boldsymbol{B_s}$	O_{s1}	O_{s2}	…	O_{sr}	$O_{s\cdot}$
Total	$O_{\cdot 1}$	$O_{\cdot 2}$		$O_{\cdot r}$	O

Test Chi Cuadrado de dos variables

- Se extiende la variable chi cuadrado para dos variables:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

- Esta variable tiene una distribución del chi cuadrado con $(\mathrm{r}\text{-}1)(s\text{-}1)$ grados de libertad.

Test de Outlier

Un *outlier* es un dato inusual, generalmente un valor extremo. La principal decisión a tomar es si este dato debería ser eliminado o modificado (*capping*).El test de *outlier* es sobre el dato en el que se alcanza el máximo de **G**

$\boldsymbol{H_0}$: El dato no es un outlier

$\boldsymbol{H_1}$: El dato sí es un outlier

$$\boldsymbol{G} = \frac{\max(|\boldsymbol{X} - \bar{\boldsymbol{X}}|)}{\boldsymbol{S}}$$

El conocimiento experto y la experiencia personal son los factores más importantes para tomar la decisión. Son más relevantes que el criterio estadístico.

Test de outlier:

Test de outlier: decide si el dato es outlier, no qué hacer con él → Considera una muestra de variables gaussianas de tamaño n, media $\bar{X}$ y varianza S^2

$$G = \frac{\max(|X - \bar{X}|)}{S}$$

Chauvenet

- Es un outlier si

$$\mathbf{nP} < 1/2$$

- **P** es la probabilidad de que una gaussiana estándar sea mayor a G

Grubbs

- Es un outlier si

$$G > \frac{n-1}{\sqrt{n}}\sqrt{\frac{t^2_{n-2,\alpha/2n}}{n-2+t^2_{n-2,\alpha/2n}}}$$

Distribución de probabilidad

- Si x_0 es un posible valor de la variable aleatoria X entonces la función de distribución (o acumulada) F representa la probabilidad de que la variable aleatoria esté bajo ese valor:

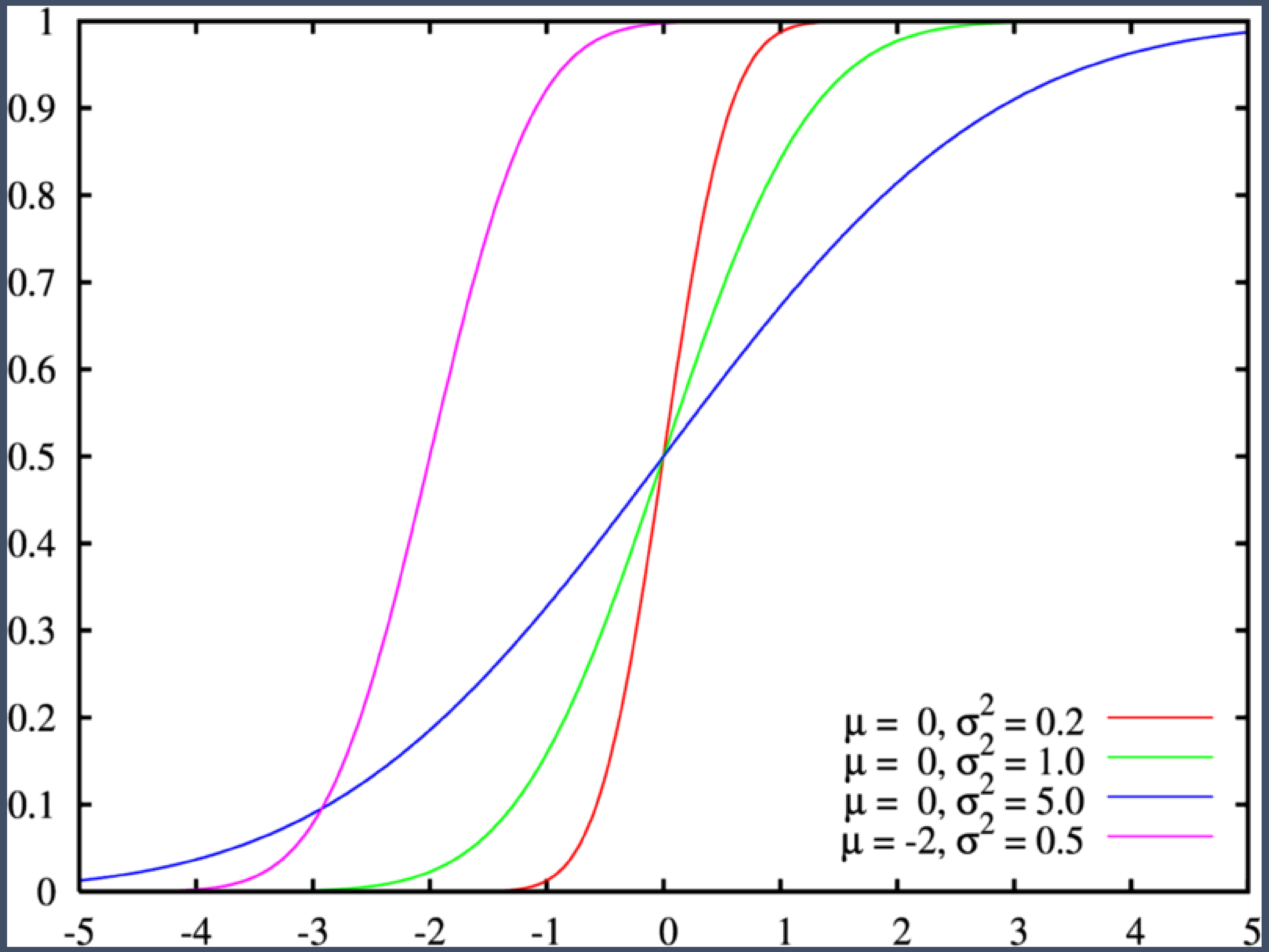
$$F(x_0) = P(X < x_0)$$

- F también se puede calcular a través de la función de densidad de probabilidad f de la variable aleatoria X:

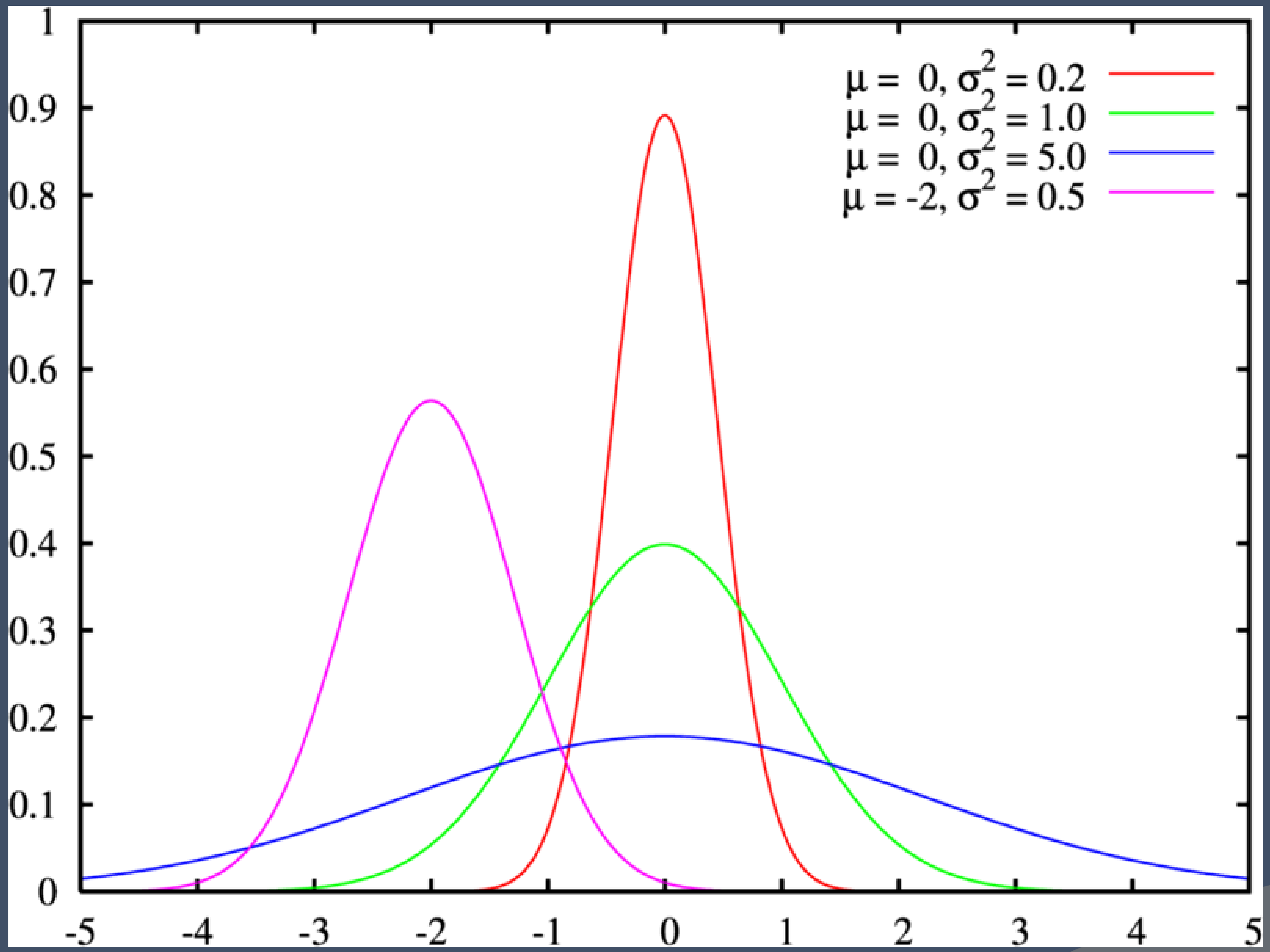
$$F(x_0) = \int_{-\infty}^{x_0} f(u)du$$

Distribución normal

- Función de distribución acumulada o *cumulative density function (cdf)*

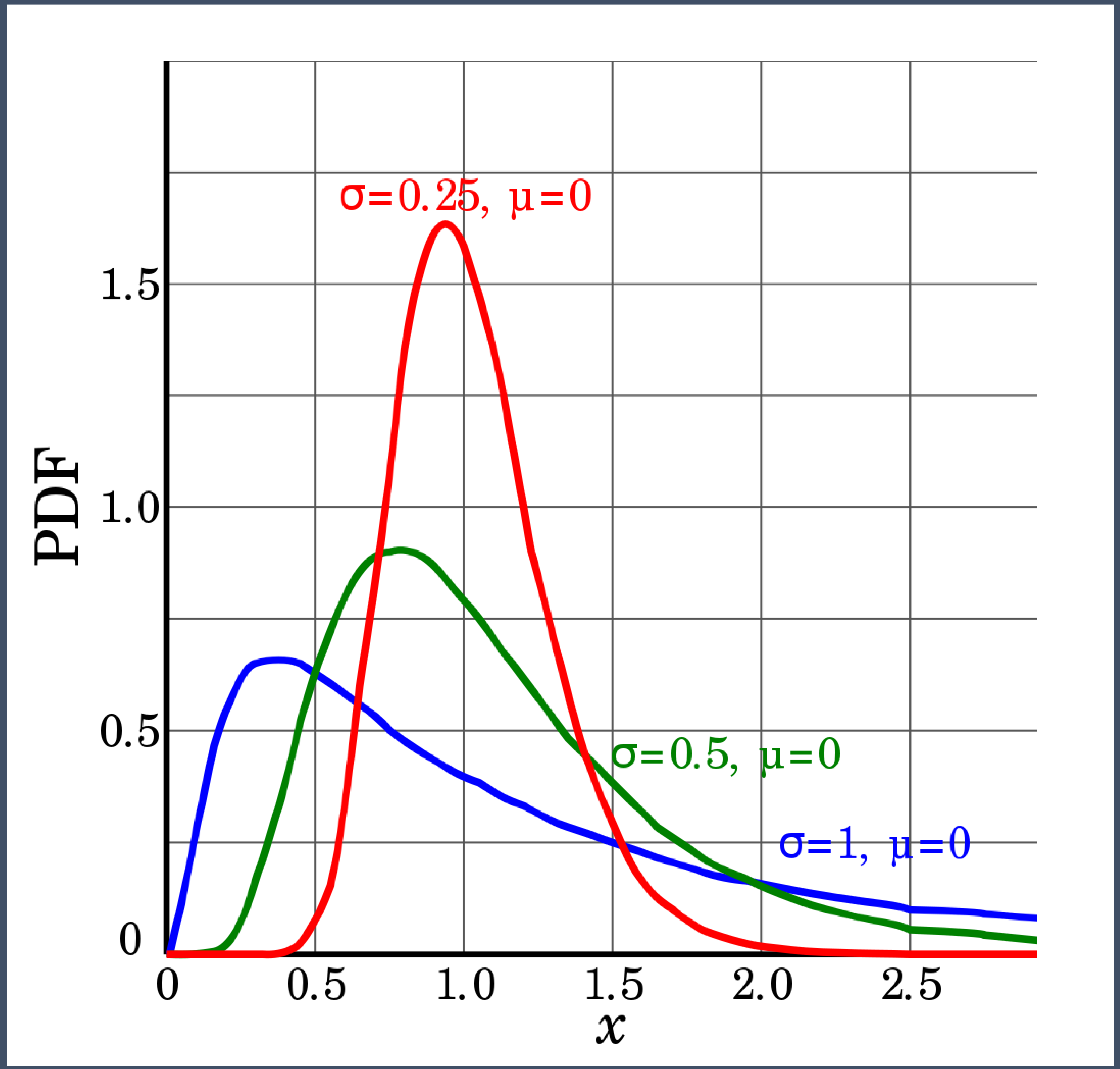


- Función de densidad de probabilidad o *probability density function (pdf)*

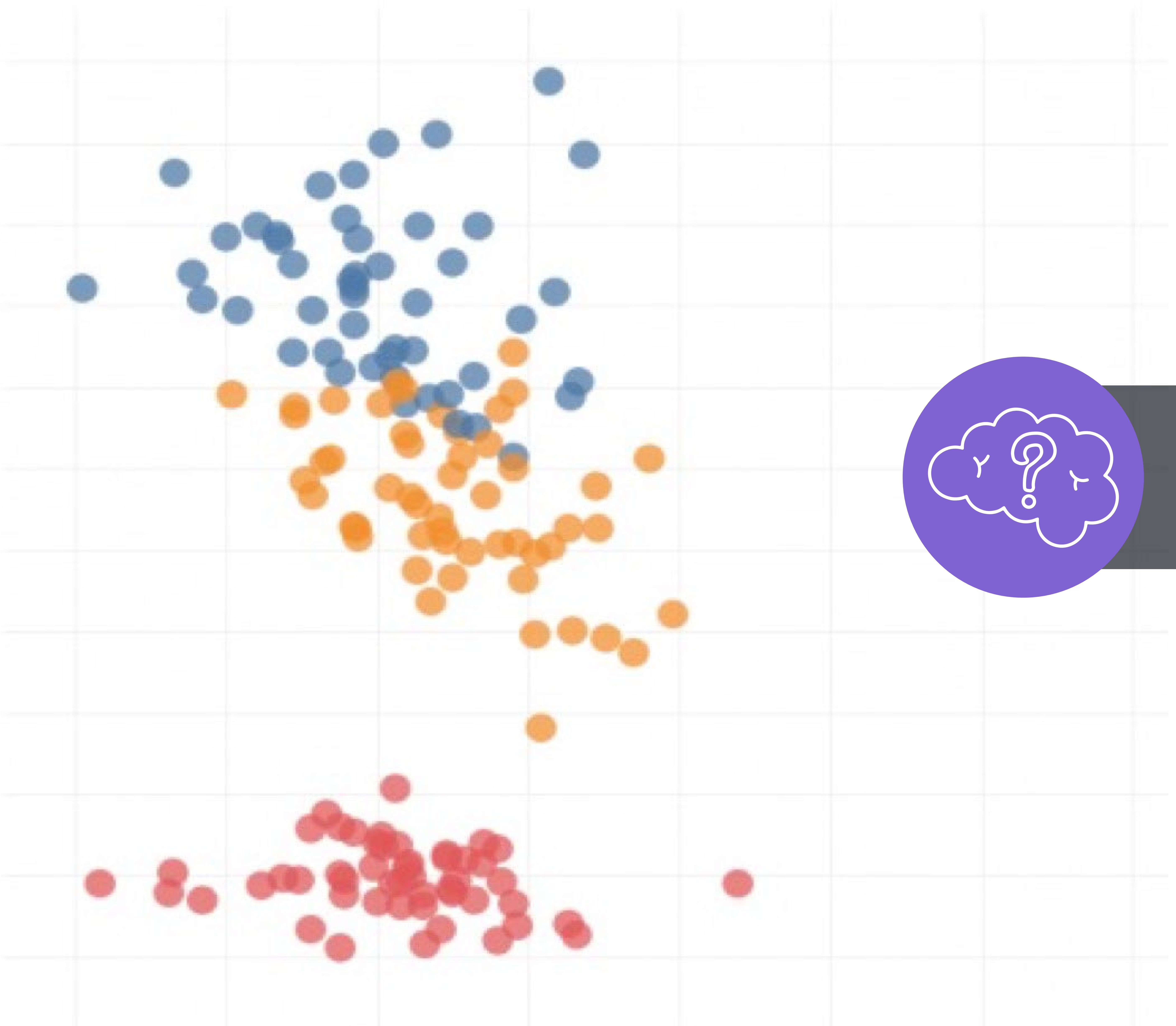


Distribución lognormal

- Función de densidad de probabilidad o *probability density function* *(pdf)*



- El logaritmo de la variable distribuye normal



ALGES
ADVANCED LABORATORY FOR GEOSTATISTICAL SUPERCOMPUTING

Problemas

Problema 1:

- Se tienen los siguientes datos:
- 36.8 32.0 18.7 43.3 25.0 156.2

- ¿Se puede considerar al último dato como aberrante?

Problema 1:

- Se tienen los siguientes datos:
- 36.8 32.0 18.7 43.3 25.0 156.2

- ¿Se puede considerar al último dato como aberrante?

Problema 2:

- Se desea saber si la distribución de las leyes de cobre en los sondajes de exploración es *lognormal*.
- Los datos se encuentran en la tabla Excel *"Leyes de cobre.xls"*

Problema 3:

- (Independencia entre 2 variables). Las fallas de tres máquinas puden tener diferentes naturalezas (fallas mecánicas o eléctricas). Después de un mes, se tienen los siguientes datos de la tabla adjunta.
- ¿La frecuencia de fallas mecánicas depende de la máquina?

Tipo de falla	Máquina 1	Máquina 2	Máquina 3	Total
Mecánica	13	11	10	34
Eléctrica	4	4	9	17
Total	17	15	19	13