Inferencia estadística

Trata de estimar o inferir mediante una muestra (aleatoria) el valor (desconocido) de un parámetro poblacional

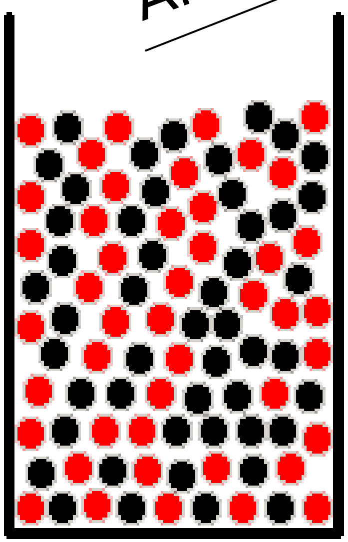
Probabilidad e Inferencia, el problema de la “Inversión”

Muestra Aleatoria

$$X_1, \ldots, X_n$$

Estimador

$$\widehat{\theta}_n = h_n(X_1, \ldots, X_n)$$

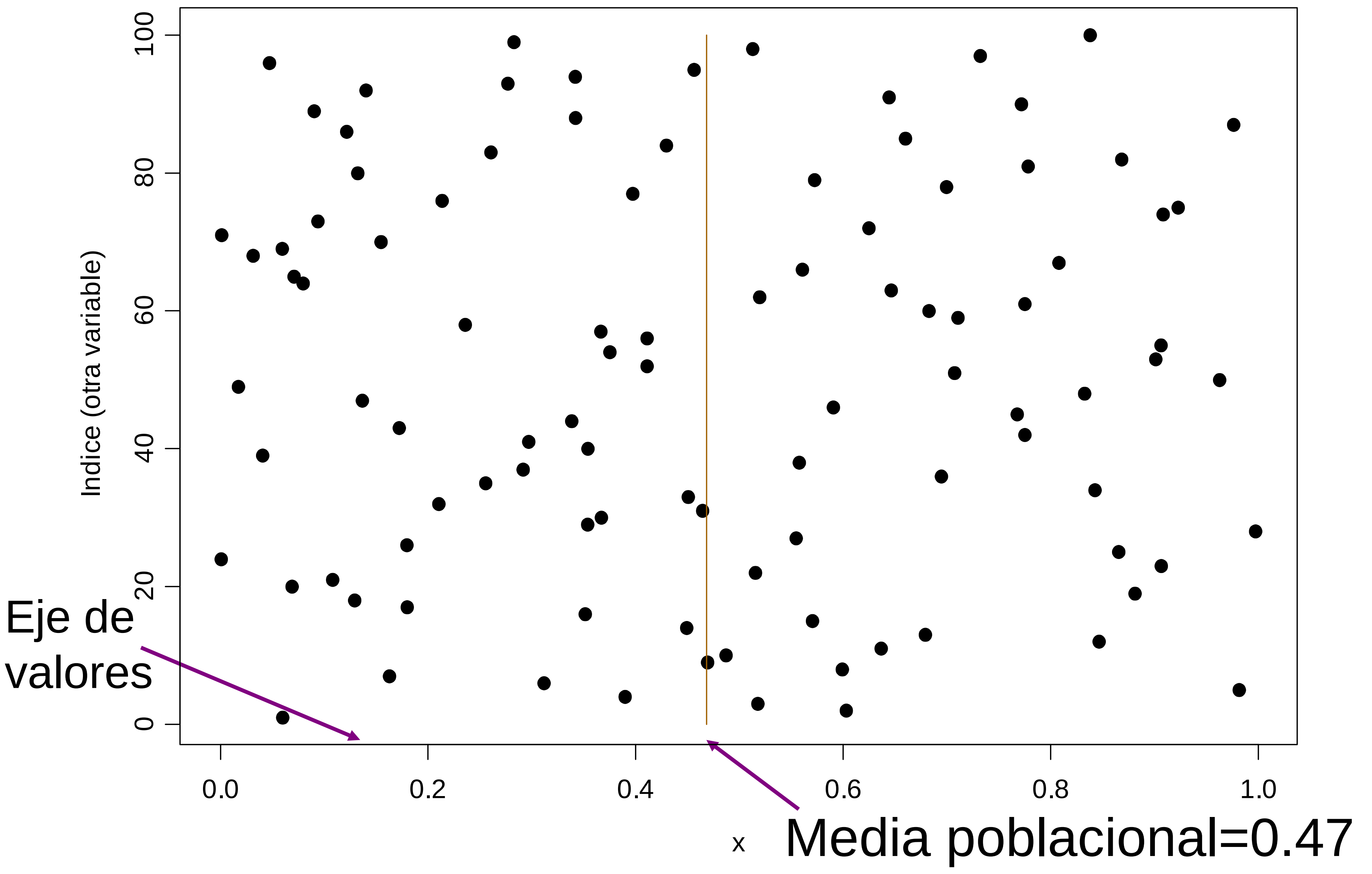


Estimación

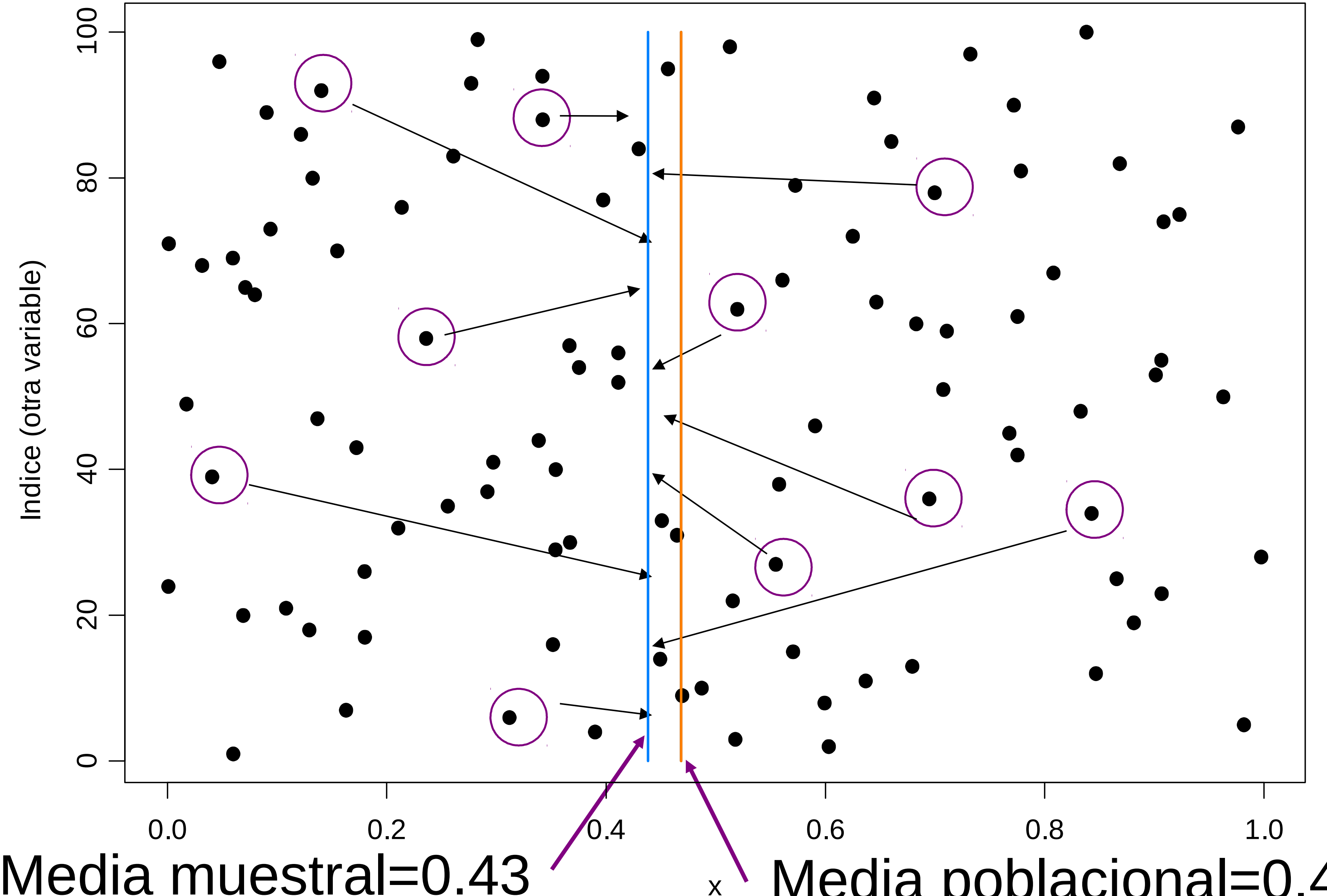
θ = Proporción poblacional de bolitas rojas

$h_n\left(x_1, x_2, \ldots, x_n\right) =$ Proporción muestral de bolitas rojas

La población (N=100)



La muestra (n=10)



Media muestral=0.43

Media poblacional=0.47

El estimador y la estimación

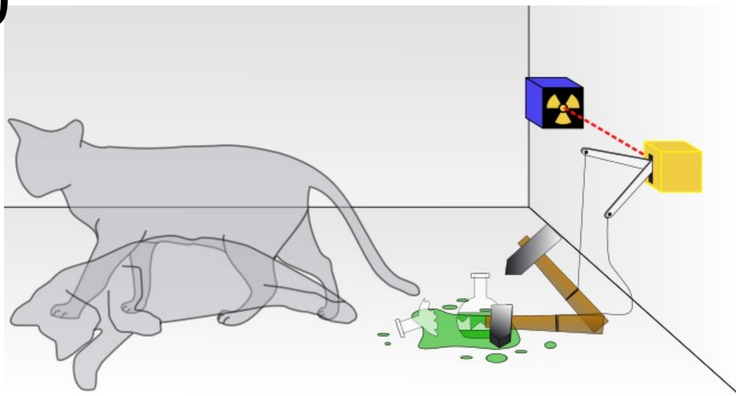
Dados los datos: $x_1, x_2 \ldots x_{10}$

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{x_1}{10} + \frac{x_2}{10} + \ldots + \frac{x_{10}}{10} = 0.43$$

Dados las variables aleatorias: $x_1, x_2 \ldots x_{10}$

$$\bar{X} = \frac{\sum_{i=1}^{10} X_i}{10} = \frac{X_1}{10} + \frac{X_2}{10} + \ldots + \frac{X_{10}}{10}$$

Schrödinger's cat

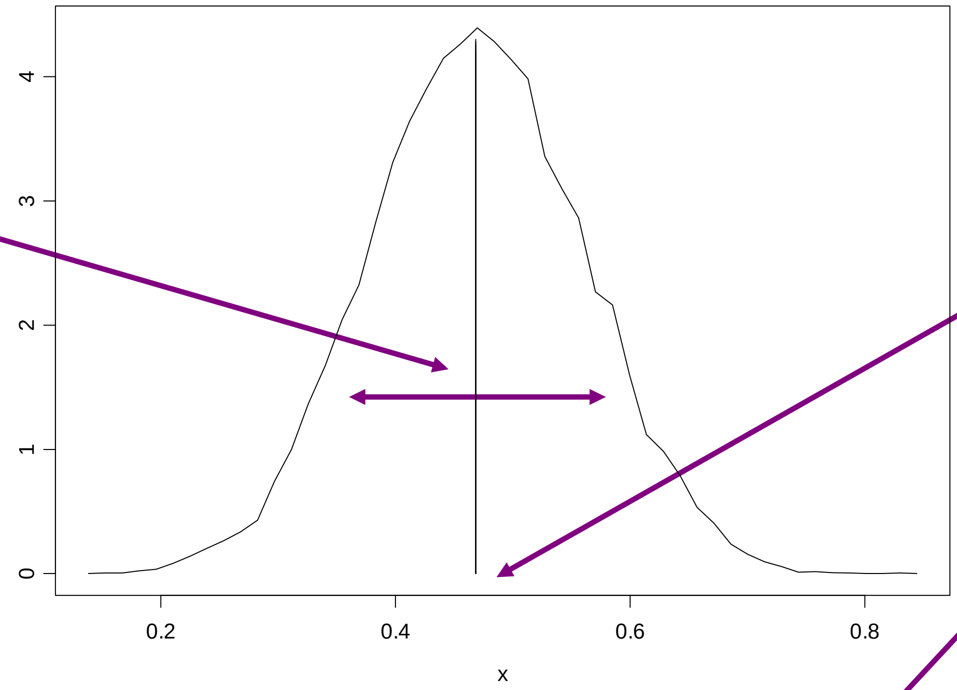


¿ Que tan bueno es ?

¿ Cuales son sus propiedades?

Repito el experimento 10000 veces

Desvío = 0.09



Media poblacional = 0.468

Insesgado

Min.	st Qu.	Median	Mean	3rd Qu.	Max.
0.15705	0.40697	0.46900	0.46879	0.53011	0.82568

Propiedades de los Estimadores

Parámetro

- Consistencia $$\boxed{\widehat{\theta}_n \to \theta}$$
- Insesgadez $$\boxed{\mathbb{E}(\widehat{\theta}_n) = \theta}$$
-

Estimador

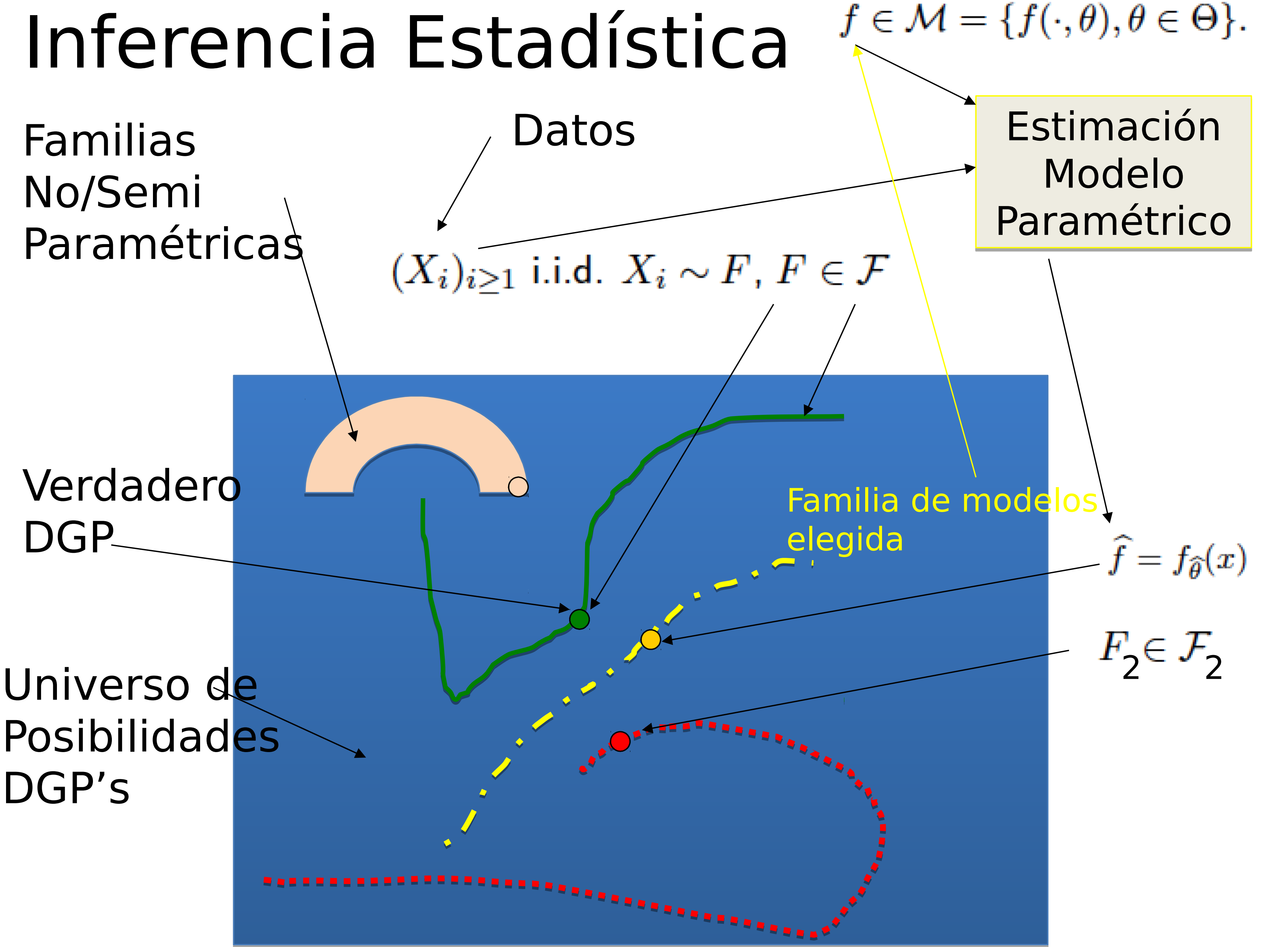
- Error Cuadrático Medio $$\boxed{\mathsf{ECM}(\widehat{\theta}_n) = \mathbb{V}(\widehat{\theta}_n) + \left\{\mathbb{E}(\widehat{\theta}_n) - \theta\right\}^2}$$

$$\boxed{\mathbb{E}(\widehat{\theta}_n - \theta)^2}$$

Varianza

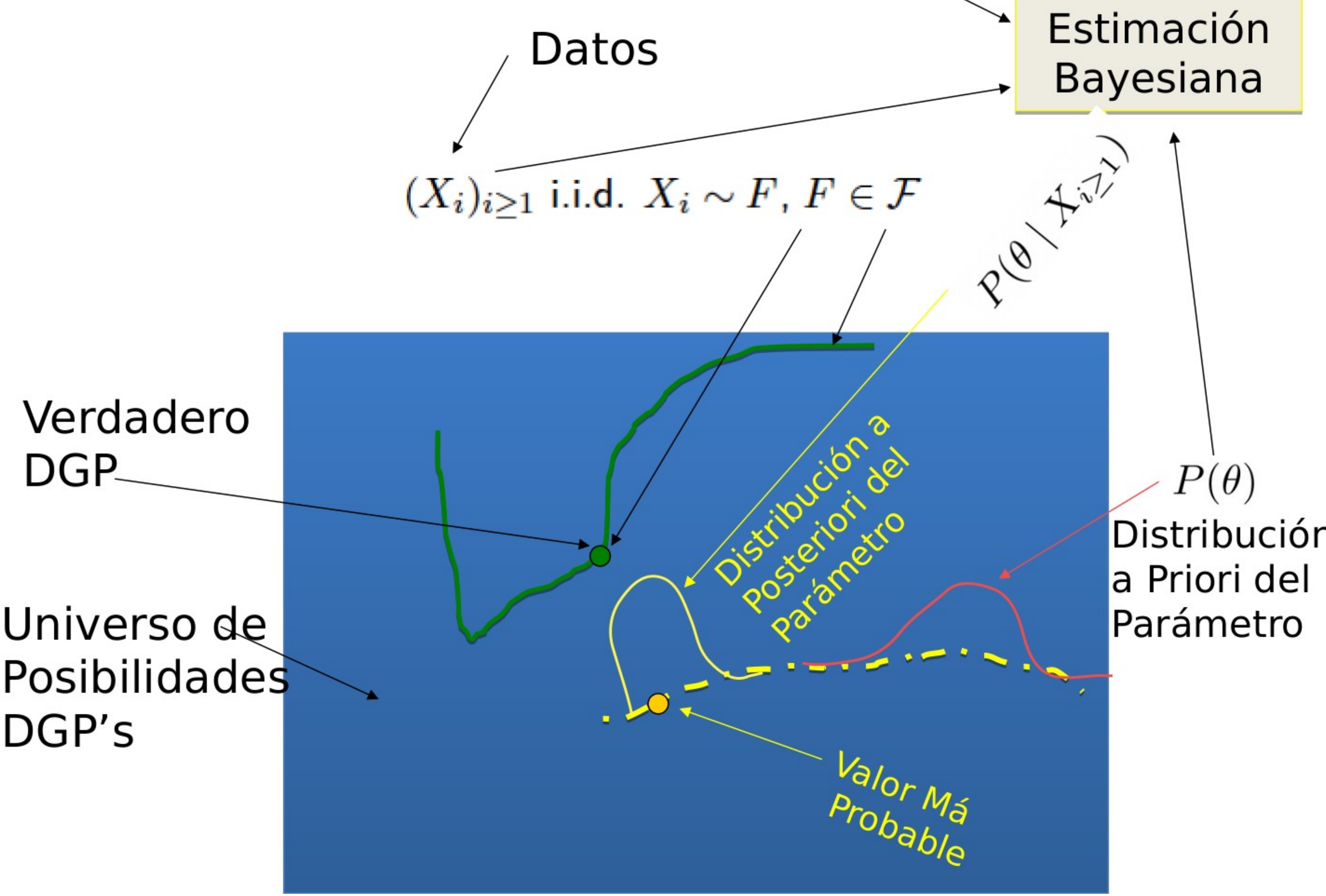
Sesgo^2

Inferencia Estadística

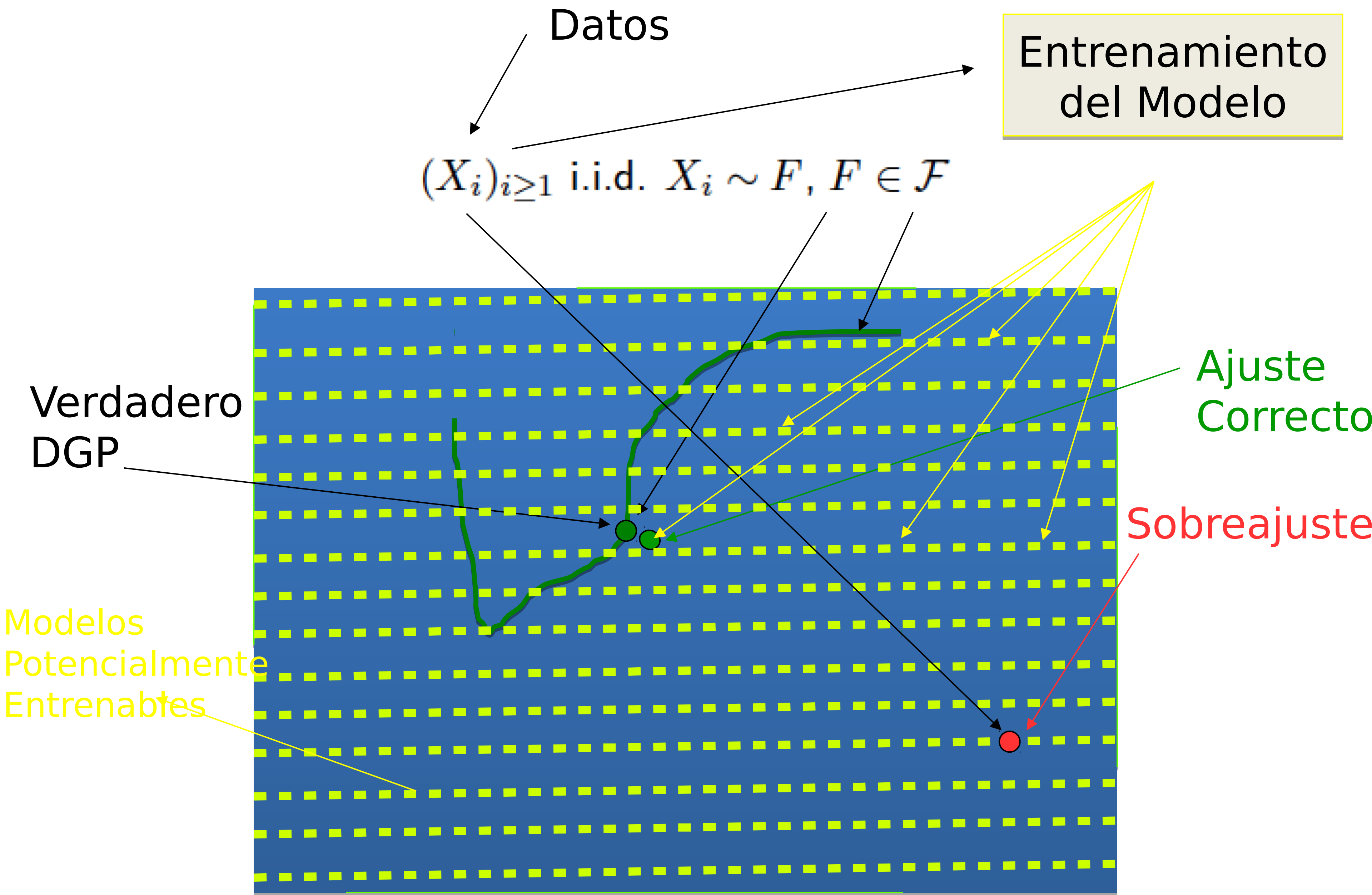


Enfoque Bayesiano

$f \in \mathcal{M} = \{f(\cdot, \theta), \theta \in \Theta\}.$



Aproximadores Universales (ANN - MLP)



Verosimilitud

Parámetros

Variables
Aleatorias

- Modelo: $\mathcal{M} = \{p(\cdot, \theta) \, , \theta \in \Theta\}$.
- $\mathbf{x} = x_1, \cdots, x_n$ realización de $X_1, \cdots, X_n$ i.i.d.
- Función de verosimilitud asociada a $\mathbf{x} = x_1, \cdots, x_n$:

$$L(\cdot \, ; \, \mathbf{x}) \; : \Theta \to \mathbb{R}$$

Los
Parámetros
varían !!!

$$L(\theta \, ; \, \mathbf{x}) \; = \; \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) \, , X_i \sim p(\cdot, \theta) \, .$$

$$\boxed{L(\theta \, ; \, \mathbf{x}) \; = \; \prod_{i=1}^{n} p(x_i, \theta) \, ,}$$

Observaciones están fijas !!!

Estimador de Máxima-Verosimilitud

- Función de verosimilitud: $L(\cdot\,;\,\mathbf{x}\,) \,:\Theta \to \mathbb{R}$

$$L(\theta\,;\,\mathbf{x}\,) \;=\; f_{X_1,\cdots,X_n}(x_1,\ldots,x_n)\,, X_i \sim f(\cdot,\theta)\,.$$

$$L(\theta\,;\,\mathbf{x}\,) \;=\; \prod_{i=1}^{n} f(x_i,\theta)\,,$$

- Propuesta de Máxima Verosimilitud:

$$\boxed{h_n(\mathbf{x}) = \underset{\theta\in\Theta}{\operatorname{argmax}}\; L(\theta\,;\,\mathbf{x}\,)\,.}$$

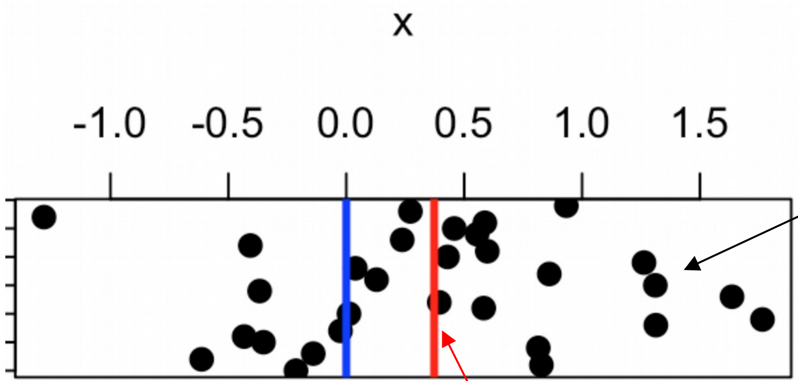
o sea

$$L(h_n(\mathbf{x}),\mathbf{x}) \geq L(\theta,\mathbf{x})$$

- Definimos el EMV siendo $\widehat{\theta}_n = h_n(X_1,\cdots,X_n)$.

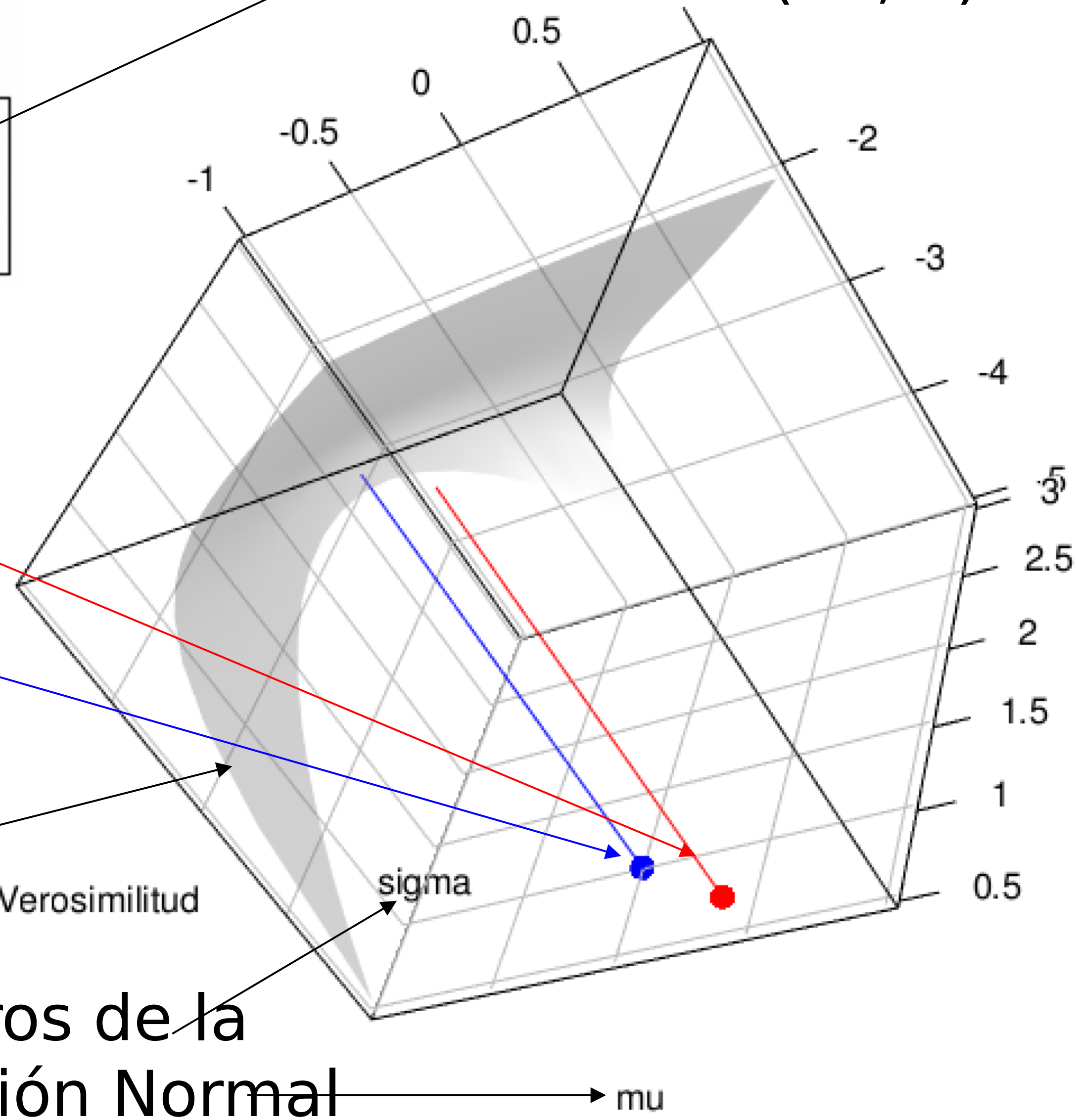
Ejemplo de Verosimilitud

30 realizaciones
$X \sim N(0, 1)$



Media verdadera = 0

Media estimada MV = 0.37



Función de Verosimilitud (Normal) basada en los 30 x’s fijos

Parámetros de la Distribución Normal

Selección de Modelos

- Modelo: $\mathcal{M} = \{f(\cdot,\theta), \theta \in \Theta\}$

- Verosimilitud (likelihood): $L(\theta\ ,\ \mathbf{x}\ )\ =\ \prod_{i=1}^{n} f(x_i,\theta)$

$$\widehat{\theta}(\mathbf{x}) = \underset{\theta\in\Theta}{\operatorname{argmax}}\ L(\theta\ ;\ \mathbf{x}\ )\ , \quad \text{o sea} \quad L(\widehat{\theta}(\mathbf{x}),\mathbf{x}) \geq L(\theta,\mathbf{x}).$$

- log-vero (log- likelihood): $\ell(\theta\ ,\ \mathbf{x}\ )\ =\ \sum_{i=1}^{n} log(f(x_i,\theta))$
- Bondad del Modelo:

$$\text{Bondad}(\mathcal{M},\mathbf{x}) := \ell(\widehat{\theta}(\mathbf{x})\ ,\ \mathbf{x}\ )\ =\ \sum_{i=1}^{n} log(f(x_i,\widehat{\theta}(\mathbf{x})))$$

$$\boxed{\text{AIC} = \text{AIC}(\mathcal{M},\mathbf{x}) := -2\left(\ell(\widehat{\theta}(\mathbf{x}),\ \mathbf{x}\ ) - \#\text{parámetros}\right)}$$

Intervalos de confianza

Dado un parámetro poblacional desconocido, buscamos un intervalo (dependiente de la muestra) que con alta probabilidad contenga al verdadero valor del parámetro.

Intervalos de confianza

Dados $x_1 \ldots x_n$ muestra aleatoria proveniente de una población con parámetro θ

Límite inferior

Límite superior

$\left[I(x_1 \ldots x_n) \, , D(x_1 \ldots x_n) \right]$ es un intervalo de confianza 0.95 si

Fijo

$$P\left(I(x_1 \ldots x_n) \leq \theta \leq D(x_1 \ldots x_n)\right) = 0.95$$

Aleatorio

Test de Hipótesis

Es un mecanismo para decidir acerca de la validez de una hipótesis, controlando la probabilidad de rechazar la misma siendo que esta es verdadera.

Potencia, Específicidad y Errores

Modelo

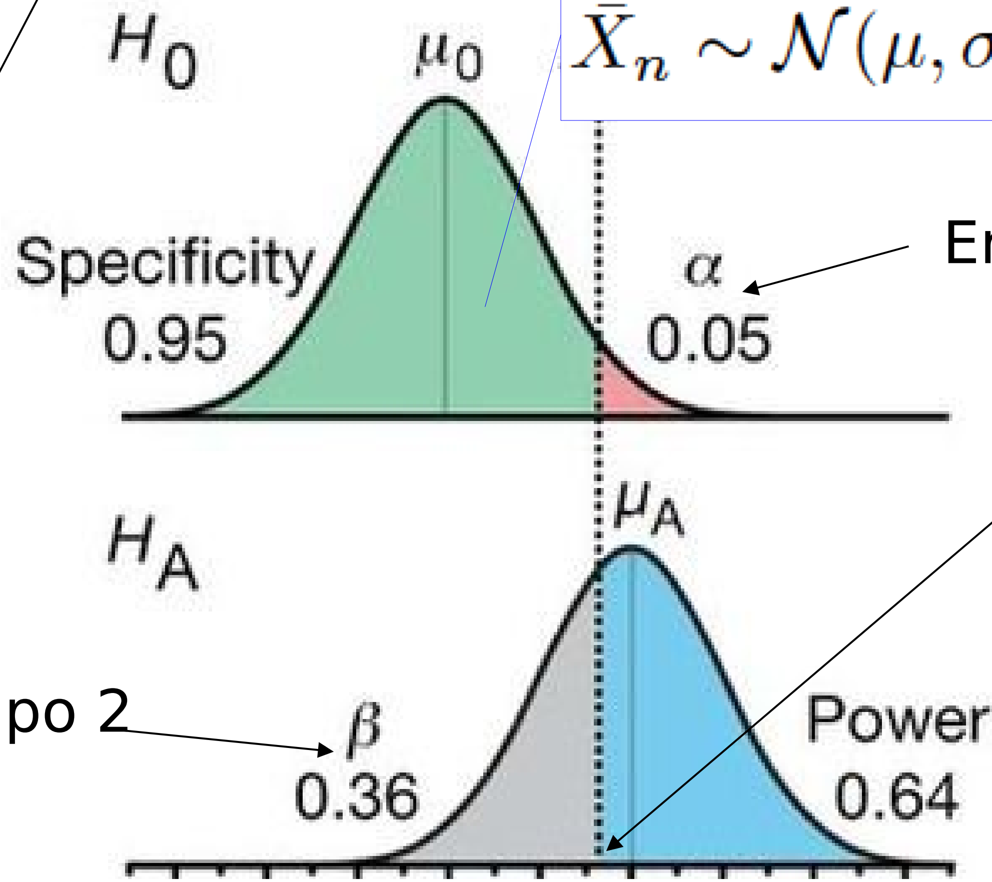
$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

Estimador

$$\bar{X} = \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

$N\left(0, \sigma\right)$

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$



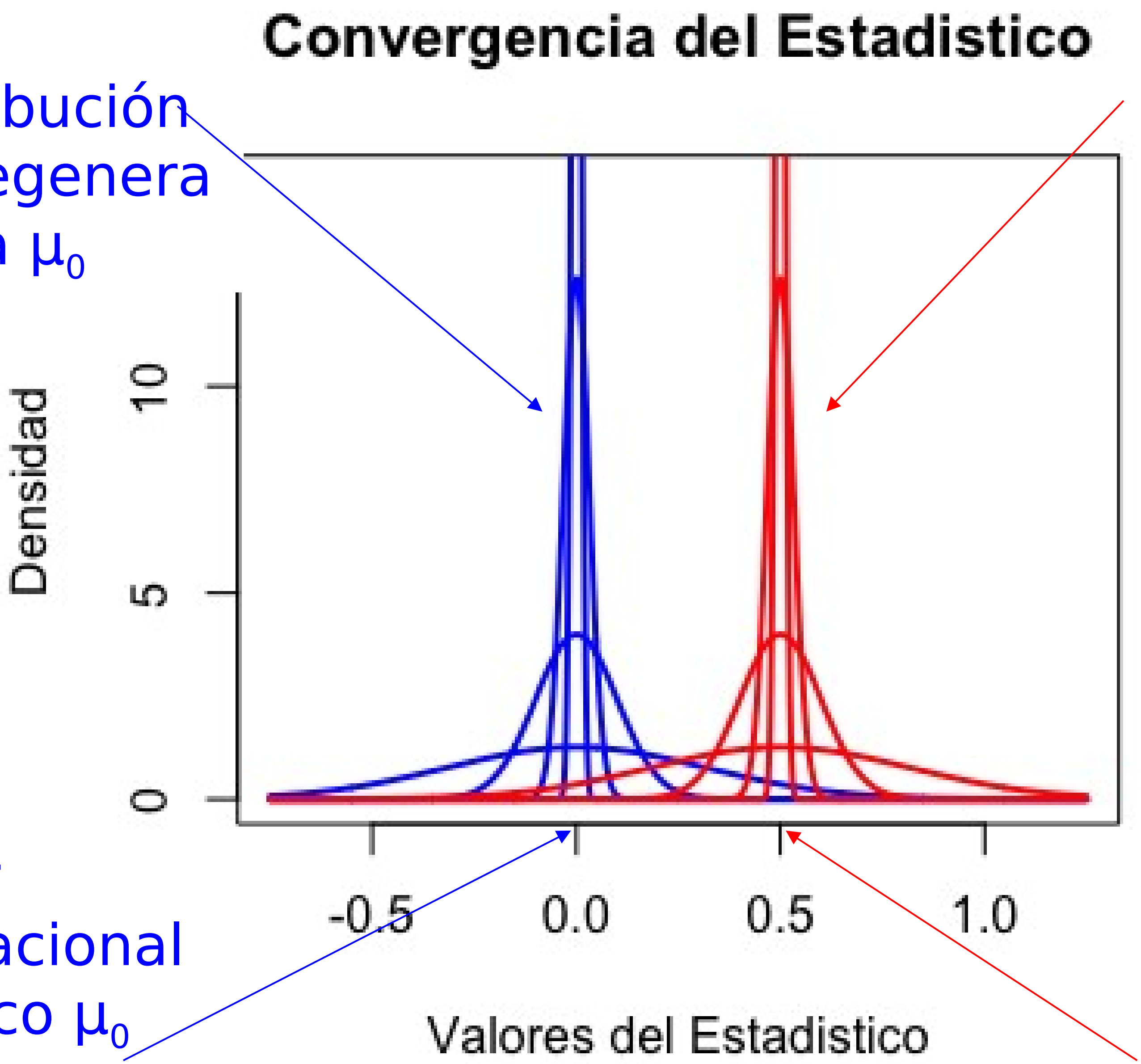
Error Tipo 1

Contra mayor, rechazo si supera el umbral

Error Tipo 2

La Virtud y El Problema de la Consistencia

La distribución se degenera hacia μ_0



La distribución se degenera hacia μ_A

Valor poblacional teórico μ_0

Valor poblacional real μ_A

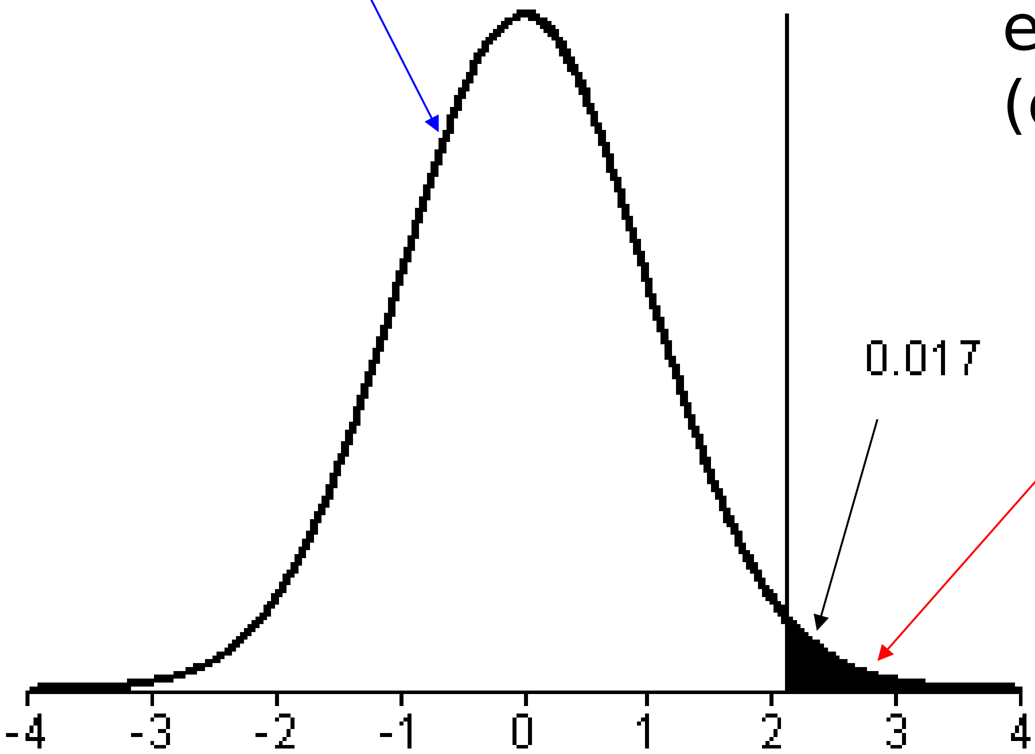
El p-valor

- Concepto fundamental de la Estadística que cuantifica objetivamente la evidencia acerca de la validez de una hipótesis.
- Específicamente, mide en base a las observaciones el graodo de "compatibilidad" de una hipótesis en términos del comportamiento distribucional de un estimador /estadístico.

Gráficamente

Distribución del Estadístico bajo la Hipótesis

Valor observado del estadístico (estimación)



Probabilidad de observar “algo tan o mas extremo” que lo observado basado en la muestra

Evidencia a favor H0   Evidencia contra H0