TEMA 1: Introducción a la Inferencia Estadística. Estadísticos muestrales

1.1. Planteamiento y modelización de un problema de inferencia.

1.2. Muestra aleatoria simple.

1.3. Función de distribución muestral.

1.4. Estadísticos muestrales. Distribuciones en el muestreo.

1.1. PLANTEAMIENTO Y MODELIZACIÓN

La Estadística es la ciencia que estudia cómo debe emplearse la información disponible sobre una situación práctica que envuelve incertidumbre para dar una guía de acción en dicha situación (Vic Barnett (1973)).

El objetivo de la Estadística es crear procedimientos o métodos que permitan formular juicios acerca del modelo de probabilidad adecuado para describir una situación de incertidumbre concreta.

PROBLEMA DE INFERENCIA ESTADÍSTICA: *Bajo el punto de vista clásico, un problema de inferencia estadística consiste en obtener conclusiones acerca del comportamiento de una o varias características en una determinada población, basándose en la observación de las mismas en un subconjunto de la población.*

▶ **ELEMENTOS BÁSICOS:**

- *Característica* (o características) de interés.
- *Población* de individuos en los que se pretende estudiar dicha característica (Ω).
- *Muestra*, parte de la población en la que se observa la característica para inferir conclusiones sobre su comportamiento en toda la población.

▶ **MODELO ESTADÍSTICO:** $(X, \mathcal{P})$

- $X : (\Omega, \mathcal{A}, P) \longrightarrow (\mathbb{R}, \mathcal{B}, P_X)$, variable aleatoria que describe la característica objeto de estudio.
- $\mathcal{P}$: familia de todas las distribuciones de probabilidad que, según nuestro conocimiento, pueden ser la de X.

Modelos estadísticos paramétricos: se conoce la forma funcional de P_X y el desconocimiento sólo se refiere al valor de uno o varios parámetros:

$$P_X \in \mathcal{P} = \{P_\theta;\ \underbrace{\theta \in \Theta}_{\downarrow\ \textit{espacio paramétrico}}\}.$$

Modelos estadísticos no paramétricos: la forma de P_X es desconocida; la información sobre ella se refiere a propiedades de tipo general (es de tipo continuo, es simétrica,...).

1.2. MUESTRA ALEATORIA SIMPLE

Una muestra aleatoria simple (de tamaño n) de una variable X es un vector, $(X_1, \ldots, X_n)$, formado por n variables aleatorias independientes, todas con la misma distribución que X.

- *Realización muestral*: valor concreto obtenido al observar una muestra aleatoria simple.
- *Espacio muestral*: conjunto de todas las posibles realizaciones muestrales según nuestro conocimiento sobre P_X. Notando $f_\theta(x)$ a la f.m.p. o f. densidad de X bajo P_θ:

  $\chi_\theta = \{x \ / \ f_\theta(x) > 0\} \rightarrow$ valores de X bajo P_θ.

  $\chi = \bigcup_{\theta\in\Theta} \chi_\theta \rightarrow$ valores de X bajo $\mathcal{P}$.

$$\chi^n = \chi \times \cdots \times \chi = \{(x_1, \ldots, x_n); \quad x_1, \ldots, x_n \in \chi\}.$$

Distribución de una muestra aleatoria simple:

- X discreta: $P_\theta\left(X_1 = x_1, \ldots, X_n = x_n\right) = P_\theta\left(X = x_1\right) \cdots P_\theta\left(X = x_n\right), \quad (x_1, \ldots, x_n) \in \chi^n.$
- X continua: $f_\theta^n(x_1, \ldots, x_n) = f_\theta(x_1) \cdots f_\theta(x_n), \quad (x_1, \ldots, x_n) \in \chi^n.$

1.3. FUNCIÓN DE DISTRIBUCIÓN MUESTRAL

Dada una muestra aleatoria simple, $(X_1, \ldots, X_n)$, de una variable X con función de distribución F_X, la función de distribución muestral, $F^*_{X_1,\ldots,X_n}$, es una función sobre $\mathbb{R}$ definida por

$$F^*_{X_1,\ldots,X_n}(x) = \frac{n^{\underline{o}} \text{ de variables } X_i \text{ menores o iguales que } x}{n} = \frac{\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)}{n}, \quad \forall x \in \mathbb{R}.$$

Propiedades:

- *Para cada realización muestral,* $(x_1, \ldots, x_n) \in \chi^n$, $F^*_{x_1,\ldots,x_n}$ *es una función de distribución en* $\mathbb{R}$.
- $\forall x \in \mathbb{R}$, $F^*_{X_1,\ldots,X_n}(x)$ *es una variable aleatoria tal que* $nF^*_{X_1,\ldots,X_n}(x) \to B(n, F_X(x))$:

$$E\left[F^*_{X_1,\ldots,X_n}(x)\right] = F_X(x), \qquad Var\left[F^*_{X_1,\ldots,X_n}(x)\right] = \frac{F_X(x)(1-F_X(x))}{n}.$$

- Para valores grandes de n, en virtud del *Teorema Límite de Lèvy*:

$$F^*_{X_1,\ldots,X_n}(x) \simeq \mathcal{N}\left(F_X(x), \frac{F_X(x)(1-F_X(x))}{n}\right), \quad \forall x \in \mathbb{R}.$$

- $F^*_{X_1,\ldots,X_n}(x) \underset{n\to\infty}{\xrightarrow{P}} F_X(x), \quad \forall x \in \mathbb{R}$ (*Ley débil de Bernoulli*).
- $F^*_{X_1,\ldots,X_n}(x) \underset{n\to\infty}{\xrightarrow{c.s.}} F_X(x), \quad \forall x \in \mathbb{R}$ (*Ley fuerte de Borel*).

Teorema de Glivenko-Cantelli: *Si* $\{X_n\}_{n\in\mathbb{N}}$ *es una sucesión de variables independientes e idénticamente distribuidas con función de distribución* F_X*, las funciones de distribución muestrales* $F^*_{X_1,\ldots,X_n}$ *convergen casi segura y uniformemente a la teórica,* F_X*:*

$$P\left\{\lim_{n\to+\infty} \sup_{x\in\mathbb{R}} \left|F^*_{X_1,\ldots,X_n}(x) - F_X(x)\right| = 0\right\} = 1.$$

$$\Downarrow$$

Con probabilidad 1, al tomar sucesivas observaciones independientes de la variable y considerar las correspondientes funciones de distribución muestrales:

$$\forall \varepsilon > 0 \;\; \exists n_\varepsilon \;/\; n > n_\varepsilon \;\Rightarrow\; F_X(x) \in \left(F^*_{X_1,\ldots,X_n}(x) - \varepsilon,\; F^*_{X_1,\ldots,X_n}(x) + \varepsilon\right), \quad \forall x \in \mathbb{R}.$$

Demostración: Vélez y García, pg. 37.

1.4. ESTADÍSTICOS MUESTRALES

Dada una muestra aleatoria simple $(X_1, \dots, X_n)$ *de una variable* X*, un estadístico (muestral), es una función de ella,* $T(X_1, \dots, X_n)$*, tal que*

$$T : (\mathbb{R}^n, \mathcal{B}^n) \longrightarrow (\mathbb{R}^k, \mathcal{B}^k)$$

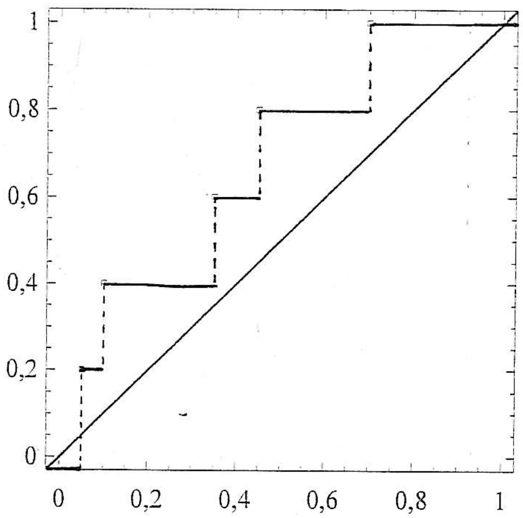
es medible ($T^{-1}(B) \in \mathcal{B}^n,\ \forall B \in \mathcal{B}^k$) *e independiente de cualquier parámetro desconocido.*

ESTADÍSTICOS DE INTERÉS:

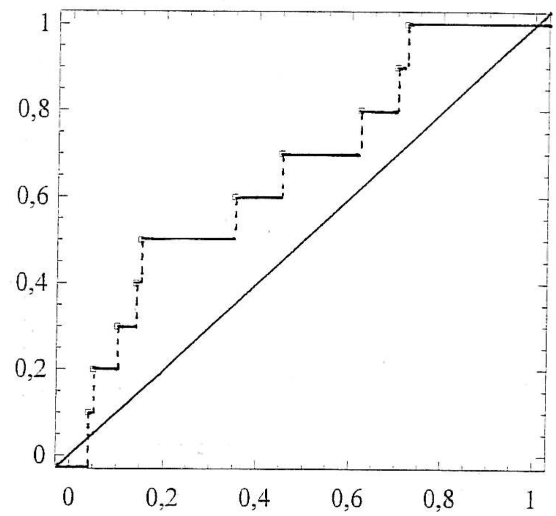
- *Momentos muestrales no centrados*: $A_k = \dfrac{\sum_{i=1}^{n} X_i^k}{n}, \quad k \in \mathbb{N}.$
  - $A_1 = \overline{X} \to$ *media muestral.*
- *Momentos muestrales centrados*: $B_k = \dfrac{\sum_{i=1}^{n} (X_i - \overline{X})^k}{n}, \quad k \in \mathbb{N}.$
  - $B_1 = 0$
  - $B_2 = \dfrac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n} \to$ *varianza muestral.*

    Cuasivarianza muestral: $S^2 = \dfrac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}.$
- $M = \text{máx}(X_1, \dots, X_n), \quad N = \text{mín}(X_1, \dots, X_n).$

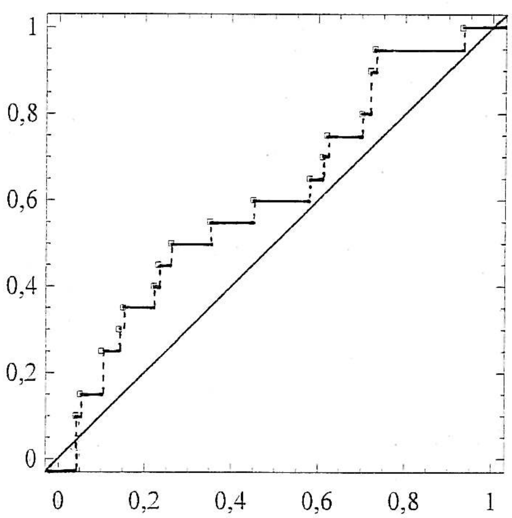
FUNCIÓN DE DISTRIBUCIÓN TEÓRICA Y FUNCIONES DE DISTRIBUCIÓN MUESTRALES ASOCIADAS A OBSERVACIONES DE MUESTRAS DE TAMAÑO n=5, 10, 20 Y 50 DE UNA DISTRIBUCIÓN U(0,1)



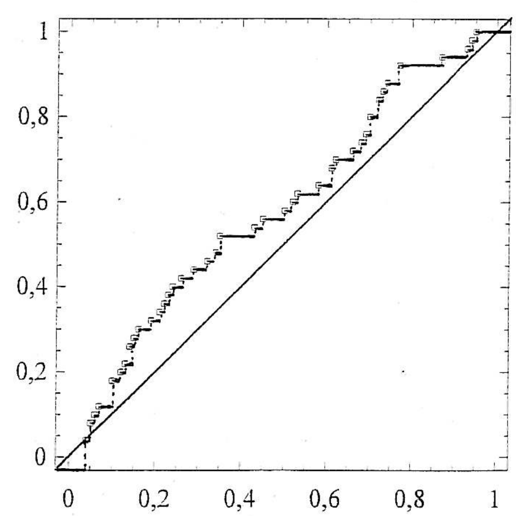
,05	,10	,35	,45	,70



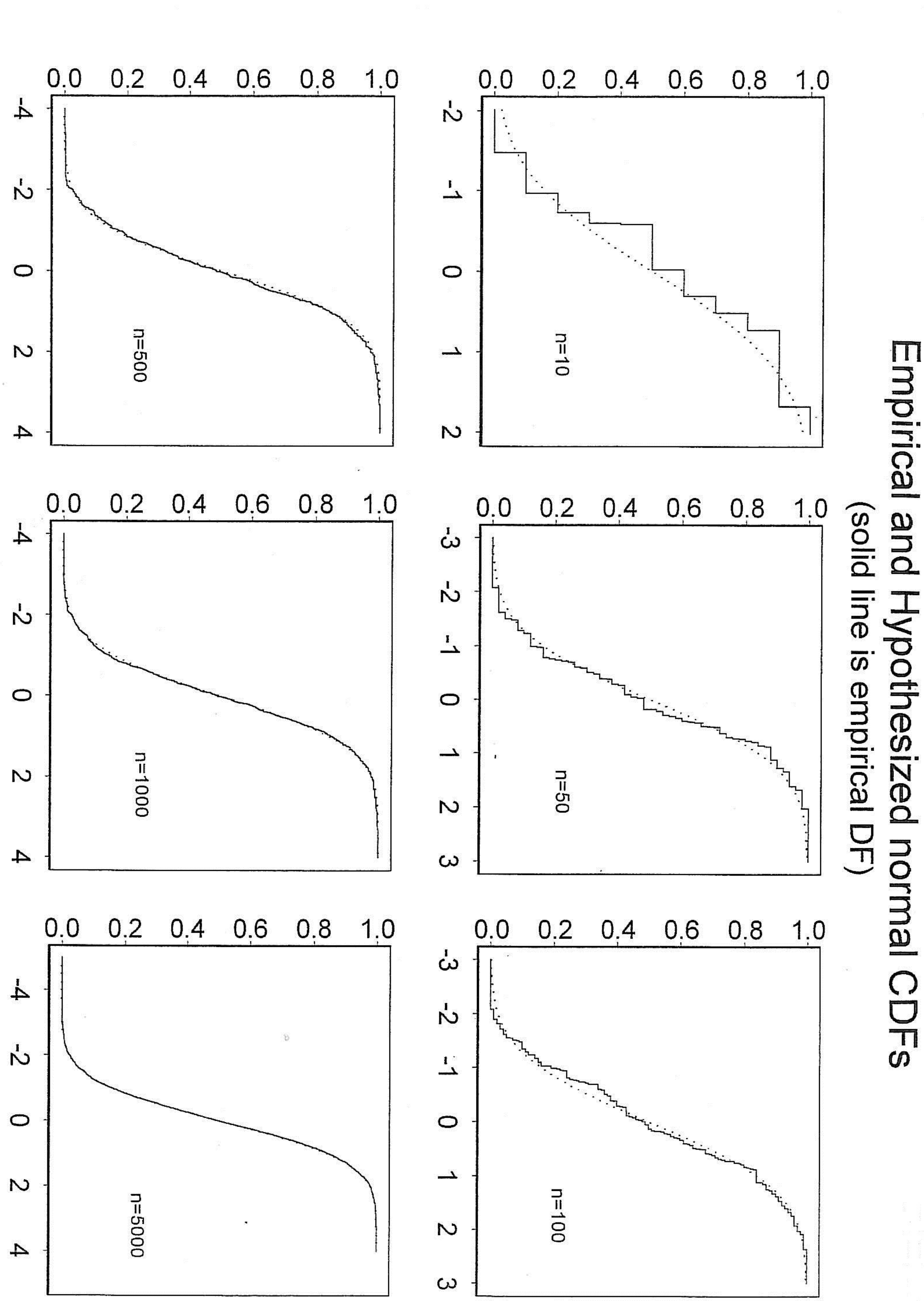
,04	,05	,10	,14	,15	,35	,45	,62	,70	,72



,04	,04	,05	,10	,10	,14	,15	,22	,23	,26
,35	,45	,58	,61	,62	,70	,72	,72	,73	,93



,04	,04	,05	,05	,06	,07	,10	,10	,10	,12
,13	,14	,14	,15	,16	,19	,21	,22	,23	,24
,26	,29	,32	,34	,35	,35	,43	,45	,50	,52
,53	,58	,61	,61	,62	,66	,68	,69	,70	,70
,72	,72	,73	,74	,77	,77	,87	,93	,94	,95


Empirical and Hypothesized normal CDFs
(solid line is empirical DF)
n=10
n=50
n=100
n=500
n=1000
n=5000