

Tema 1. Introducción a la estadística noparamétrica

Rosa M. Crujeiras
Alberto Rodríguez



Dpto. de Estadística e Investigación Operativa
Máster en Técnicas Estadísticas
Curso 2009-2010

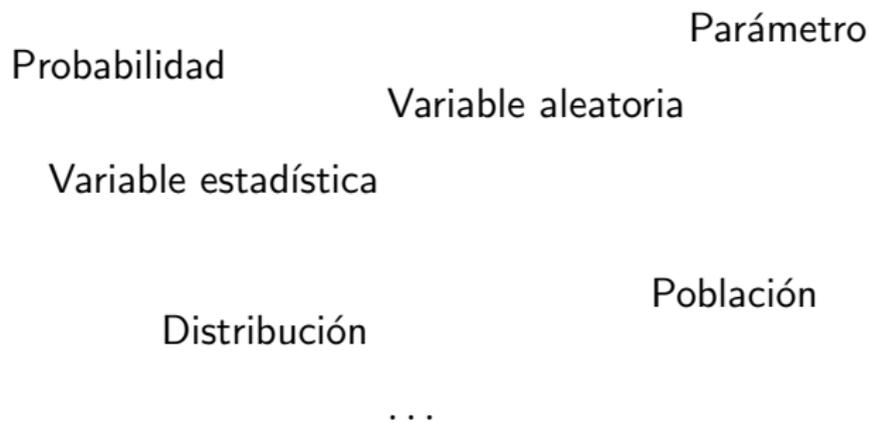
Objetivo de las Técnicas Estadísticas

Modelar situaciones en las que interviene el azar, es decir, los experimentos aleatorios.

Objetivo de las Técnicas Estadísticas

Modelar situaciones en las que interviene el azar, es decir, los experimentos aleatorios.

¿Qué es un experimento aleatorio?



Descriptiva

- X variable estadística
- x_1, \dots, x_n muestra de datos
- Representación de datos
 - Frecuencias relativas
 - Frecuencias rel. acumuladas
 - Histograma
 - Medidas características

$$\bar{x}, s^2$$

- (X, Y) bidimensional

$$y = a + bx$$

Inferencia

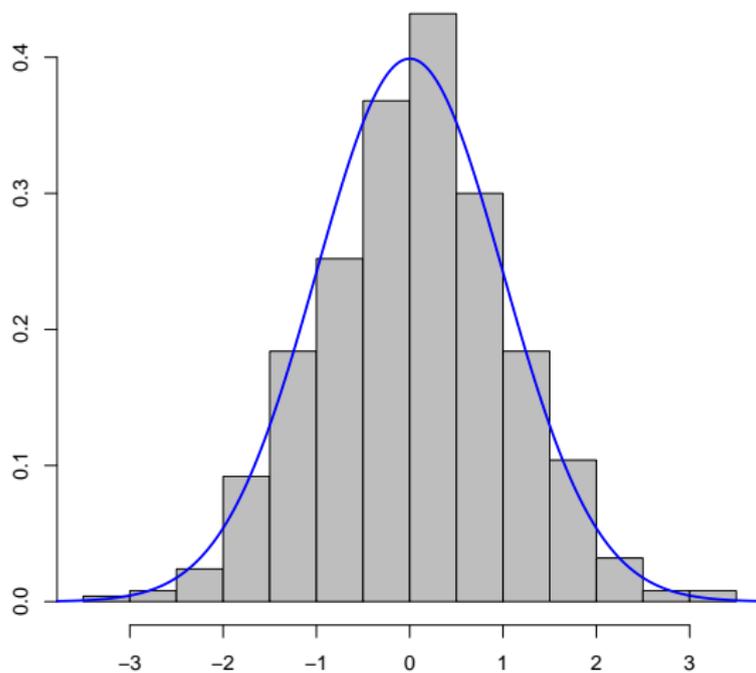
- X variable aleatoria
- Población
- Caracterización de X
 - Probabilidad \mathbb{P}
 - Distribución F ($F_\theta, \theta \in \Theta$)
 - Densidad f ($f_\theta, \theta \in \Theta$)
 - Medidas características

$$\mu = \int x dF(x), \sigma^2 = \int (x - \mu)^2 dF(x)$$

- (X, Y) bidimensional

$$Y = m(X) + \varepsilon$$

Histograma



Sea X la variable de interés. Tratamos de dotarla de un modelo de distribución (paramétrico) para su estudio. Por ejemplo, consideremos las variables X :

- N° de alumnos que aprobarán la asignatura
- Llegada de clientes a una oficina
- Tiempo de funcionamiento de una bombilla
- Tiempo de funcionamiento de las componentes de un circuito
- Nivel de albúmina en sangre
- ...

Sea X el nivel de albúmina en sangre y supongamos que su comportamiento aleatorio puede modelarse a través de $X \sim N(\mu, \sigma^2)$.

- ¿Cuál es el nivel esperado de albúmina en sangre?
- ¿Entre qué dos valores está este nivel, con una probabilidad $(1 - \alpha)\%$?
- ¿Podemos admitir que el nivel medio de albúmina es como mínimo de 10 g/dl?

Supongamos que queremos estimar:

$$\eta = \mathbb{P}(X > t),$$

para un cierto t fijado, y disponemos de una m.a.s. X_1, \dots, X_n .

Supongamos que queremos estimar:

$$\eta = \mathbb{P}(X > t),$$

para un cierto t fijado, y disponemos de una m.a.s. X_1, \dots, X_n .
Si suponemos un modelo Normal:

$$\eta = \mathbb{P}(X > t) = 1 - \mathbb{P}(X \leq t) = 1 - \Phi(t)$$

¿Y ahora?

Pero si X mide el tiempo de duración de una bombilla, parece más adecuado $X \sim Exp(1/\theta)$, con

$$f_{\theta}(x) = \frac{1}{\theta}e^{-x/\theta}, \quad F_{\theta}(x) = 1 - e^{-x/\theta}$$

Entonces

$$\hat{\eta}_P = 1 - F_{\hat{\theta}}(t) = e^{-t/\hat{\theta}} = e^{-t/\bar{X}}$$

Una alternativa que aborda ambas situaciones es la estimación noparamétrica:

$$\hat{\eta}_{NP} = \frac{\#\{X_i > t\}}{n} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i > t)$$

Para cada test paramétrico hay al menos un test noparamétrico *equivalente*.

Queremos comparar las medias de X e Y a partir de m.a.s.
 $X_1, \dots, X_n, Y_1, \dots, Y_m$:

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{1/n + 1/m}} \sim t_{n+m-2}$$

Queremos comparar las medias de X e Y a partir de m.a.s.
 $X_1, \dots, X_n, Y_1, \dots, Y_m$:

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{1/n + 1/m}} \sim t_{n+m-2}$$

Hipótesis:

- X, Y v.a. independientes
- $X \sim N(\mu_X, \sigma^2), Y \sim N(\mu_Y, \sigma^2)$

El estadístico U de Mann-Whitney se define como el número de veces que Y precede a X o viceversa:

$$X_1, \dots, X_n, \quad Y_1, \dots, Y_m$$

$$D_{ij} = \begin{cases} 1 & \text{si } Y_j < X_i \\ 0 & \text{si } Y_j > X_i \end{cases}$$

$$U = \sum_{i=1}^n \sum_{j=1}^m D_{ij}$$

Gibbons and Chakraborti (1992)

Paramétrico	Noparamétrico
<i>t</i> -test	<i>U</i> -Mann Whitney Wald-Wolfowitz Kolmogorov-Smirnov
ANOVA	Kruskal-Wallis
<i>t</i> -test depend.	Test de signos Test de Wilcoxon
Coef. correlación	ρ -Spearman τ -Kendall γ -coefficient

- Los estimadores paramétricos son más eficientes si el modelo es adecuado... pero si no lo es, pueden ser inconsistentes.
- Los estimadores noparamétricos proporcionan resultados consistentes, pero a costa de perder eficiencia en muestras pequeñas o moderadas.
- ¿Cuándo utilizar una técnica noparamétrica?
 - Cuando no tenemos claro el modelo paramétrico
 - Para validar las hipótesis paramétricas

En esta asignatura veremos métodos de estimación noparamétricos para:

- La función de distribución.
- La función de densidad.
- La función de regresión.