



FACULTAD DE  
INGENIERÍA



UNIVERSIDAD  
DE LA REPÚBLICA  
URUGUAY

# Inferencia Estadística

## Breve Repaso

Paola Bermolen

paola@fing.edu.uy

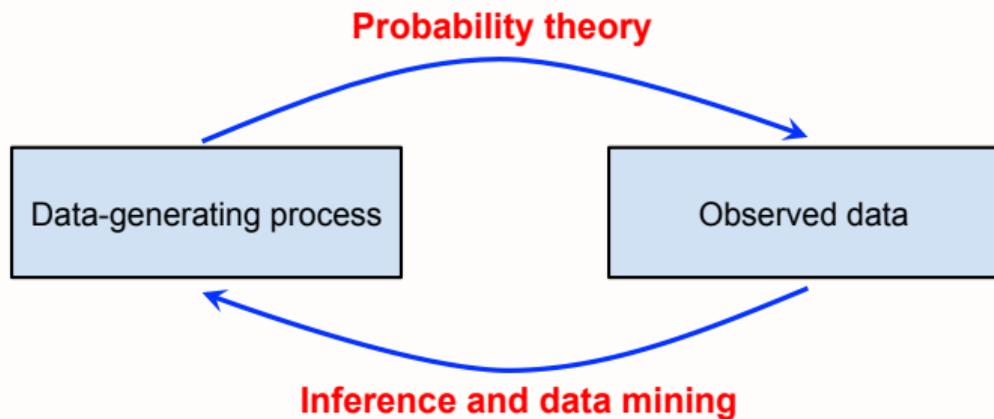


FACULTAD DE  
INGENIERÍA  
UDELAR

# Inferencia estadística y modelos

- 1 Inferencia estadística y modelos
- 2 Estimación puntual, intervalos de confianza y test de hipótesis
- 3 Tutorial 1: Inferencia sobre la media
- 4 Tutorial 2: Inferencia por regresión lineal

# Probabilidad e Inferencia



- **Teoría de la probabilidad** es un formalismo para manejar la incertidumbre
  - Dado un proceso de generación de datos, ¿cuáles son las propiedades de los resultados?
- **Inferencia estadística** trata al problema inverso
  - Dados los resultados ¿qué podemos decir del proceso de generación de dichos resultados?

# Inferencia estadística

- ▶ **Inferencia estadística** refiere al proceso por el cual
  - ⇒ Dadas observaciones  $\mathbf{x} = [x_1, \dots, x_n]^T$  de  $X_1, \dots, X_n \sim F$
  - ⇒ Se busca extraer información sobre la distribución  $F$
- ▶ **Ej:** Inferir un atributo de  $F$  como su media
- ▶ **Ej:** Inferir la CDF  $F$  completa, o la densidad PDF  $f = F'$
- ▶ Muchas veces las observaciones son de la forma  $(y_i, x_i)$ ,  $i = 1, \dots, n$ 
  - ⇒  $Y$  es el resultado o respuesta y  $X$  es el predictor o atributo
- ▶ **P:** ¿Cuál es la relación entre las variables aleatorias (VAs)  $Y$  and  $X$ ?
- ▶ **Ej:** Aprender  $\mathbb{E}[Y | X = x]$  como función de  $x$
- ▶ **Ex:** Predecir un valor no observado  $y_*$  a partir del valor  $X_* = x_*$

# Modelos

- ▶ Un **modelo estadístico** especifica un conjunto  $\mathcal{F}$  de CDFs al cual  $F$  puede pertenecer
- ▶ Un **modelo paramétrico** es de la forma  $\mathcal{F} = \{f(x; \theta) : \theta \in \Theta\}$ 
  - Parámetro(s)  $\theta$  es desconocido y toma valores en un espacio de parámetros  $\Theta$
  - Espacio  $\Theta$  tiene  $\dim(\Theta) < \infty$ , o no creciente con el tamaño de la muestra  $n$
- ▶ **Ej:** Datos vienen de una distribución Gaussiana

$$\mathcal{F}_N = \left\{ f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \mu \in \mathbb{R}, \sigma > 0 \right\}$$

Modelo paramétrico con dos parámetros desconocidos:  $\theta = [\mu, \sigma]^T$  y  $\Theta = \mathbb{R} \times \mathbb{R}_+$

- ▶ Un **modelo no paramétrico** tiene  $\dim(\Theta) = \infty$ , o  $\dim(\Theta)$  crece con  $n$
- ▶ **Ej:**  $\mathcal{F}_{tot} = \{\text{Todas las CDFs } F\}$

# Modelos y problemas de inferencia

- Dada una muestra independiente  $\mathbf{x} = [x_1, \dots, x_n]^T$  de  $X_1, \dots, X_n \sim F$   
⇒ usualmente la inferencia estadística se aplica en contexto de modelos

## Ej: Estimación paramétrica unidimensional

- Sean observaciones Bernoulli con parámetro  $p$
- El problema es estimar  $p$  (i.e., la media o esperanza)

## Ej: Estimación paramétrica bidimensional

- Sean observaciones  $\text{sc}$  con PDF  $f \in \mathcal{F}_N$ , i.e., Gaussianos
- El problema es estimar los parámetros  $\mu$  y  $\sigma$
- Quizás interese solo  $\mu$ , y  $\sigma$  se considera como un **parámetro extra**

## Ej: Estimación no paramétrica de la CDF

- El problema es estimar  $F$  asumiendo solamente que  
 $F \in \mathcal{F}_{Tot} = \{\text{Todas las CDFs } F\}$

# Modelos de regresión

- ▶ Sean observaciones de la forma  $(Y_1, X_1), \dots, (Y_n, X_n) \sim F_{YX}$   
⇒ El objetivo es aprender la relación entre las VAs  $Y$  and  $X$
- ▶ Un enfoque típico es modelar la **función de regresión**

$$r(x) := \mathbb{E} [Y | X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y) dy$$

⇒ Equivalente al **modelo de regresión**  $Y = r(X) + \epsilon$ ,  $\mathbb{E} [\epsilon] = 0$

- ▶ Ej: Modelo de regresión **lineal** paramétrica

$$r \in \mathcal{F}_{Lin} = \{r : r(x) = \beta_0 + \beta_1 x\}$$

- ▶ Ej: Modelo de regresión lineal no-paramétrica asumiendo solo **regularidad**

$$r \in \mathcal{F}_{Sob} = \left\{ r : \int_{-\infty}^{\infty} (r''(x))^2 dx < \infty \right\}$$

# Regresión, predicción y clasificación

- ▶ Dada una muestra  $(y_1, x_1), \dots, (y_n, x_n)$  de  $(Y_1, X_1), \dots, (Y_n, X_n) \sim F_{YX}$ 
  - Ej:  $x_i$  es la presión en sangre del individuo  $i$ ,  $y_i$  años de vida
- ▶ Modelar la relación entre  $Y$  y  $X$  a través de  $r(x) = \mathbb{E}[Y | X = x]$ 
  - ⇒ P: ¿ Problemas clásicos de inferencia en este contexto?

## Ej: Regresión o ajuste de curvas

- El problema es estimar la función de regresión  $r \in \mathcal{F}$

## Ej: Predicción

- El objetivo es predecir  $Y_*$  para un nuevo individuo en base a su  $X_* = x_*$
- Si se tiene una estimación de la regresión  $\hat{r}$ , podemos definir  $y_* := \hat{r}(x_*)$

## Ex: Clasificación

- Las VAs  $Y_i$  pueden ser discretas , e.g. vive o muere codificados como  $\pm 1$
- El problema de predicción en este caso se denomina clasificación

# Conceptos fundamentales en inferencia

- 1 Inferencia estadística y modelos
- 2 Estimación puntual, intervalos de confianza y test de hipótesis
- 3 Tutorial 1: Inferencia sobre la media
- 4 Tutorial 2: Inferencia por regresión lineal

# Estimadores puntuales

- ▶ Estimación puntual refiere a un único “best guess” sobre  $F$
- ▶ Ej: Estimar el parámetro  $\beta$  en un modelo de regresión lineal

$$\mathcal{F}_{Lin} = \{r : r(\mathbf{x}) = \beta^T \mathbf{x}\}$$

- ▶ **Def:** Dada la muestra  $\mathbf{x} = [x_1, \dots, x_n]^T$  de  $X_1, \dots, X_n \sim F$ , un **estimador puntual**  $\hat{\theta}$  del parámetro  $\theta$  es alguna función

$$\hat{\theta} = g(X_1, \dots, X_n)$$

- ⇒ El estimador  $\hat{\theta}$  se calcula a partir de la muestra y es por tanto una VA.
- ⇒ La distribución de  $\hat{\theta}$  se denomina **sampling distribution**
- ▶ La **estimación** es el valor específico del estimador para una muestra dada  $\mathbf{x}$ 
  - ⇒ Se suele escribir  $\hat{\theta}_n$  en referencia explícita al tamaño de la muestra.

## Sesgo, error estándar y error cuadrático medio

- ▶ **Def:** El **sesgo** de un estimador  $\hat{\theta}$  está dado por  $\text{bias}(\hat{\theta}) := \mathbb{E} [\hat{\theta}] - \theta$
- ▶ **Def:** El **error estándar** es el desvío estándar de  $\hat{\theta}$

$$\text{se} = \text{se}(\hat{\theta}) := \sqrt{\text{var} [\hat{\theta}]}$$

⇒ En general, depende de la distribución desconocida  $F$ , pero se puede estimar  $\hat{\text{se}}$

- ▶ **Def:** El **error cuadrático medio (MSE)** es una medida de la calidad de  $\hat{\theta}$

$$\text{MSE} = \mathbb{E} [(\hat{\theta} - \theta)^2]$$

- ▶ Los valores esperados se toman respecto de la distribución

$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

# Descomposición sesgo-varianza del MSE

## Teorema

El  $MSE = \mathbb{E} [(\hat{\theta} - \theta)^2]$  se puede escribir como

$$MSE = \text{bias}^2(\hat{\theta}) + \text{var} [\hat{\theta}]$$

## Demostración.

► Sea  $\bar{\theta} = \mathbb{E} [\hat{\theta}]$ . Entonces

$$\begin{aligned}\mathbb{E} [(\hat{\theta} - \theta)^2] &= \mathbb{E} [(\hat{\theta} - \bar{\theta} + \bar{\theta} - \theta)^2] \\ &= \mathbb{E} [(\hat{\theta} - \bar{\theta})^2] + 2(\bar{\theta} - \theta)\mathbb{E} [\hat{\theta} - \bar{\theta}] + (\bar{\theta} - \theta)^2 \\ &= \text{var} [\hat{\theta}] + \text{bias}^2(\hat{\theta})\end{aligned}$$

► La última igualdad vale ya que  $\mathbb{E} [\hat{\theta} - \bar{\theta}] = \mathbb{E} [\hat{\theta}] - \bar{\theta} = 0$

# Propiedades deseables de los estimadores puntuales

- ▶ ¿Propiedades deseables para un estimador  $\hat{\theta}$  del parámetro  $\theta$ ?
- ▶ **Def:** Un estimador es **insesgado** si  $\text{bias}(\hat{\theta}) = 0$ , i.e., si  $\mathbb{E}[\hat{\theta}] = \theta$   
⇒ Un estimador insesgado es “acertado” en promedio
- ▶ **Def:** Un estimador es **consistente** si  $\hat{\theta}_n \xrightarrow{p} \theta$ , i.e. para todo  $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\hat{\theta}_n - \theta| < \epsilon) = 1$$

⇒ Un estimador consistente converge a  $\theta$  a medida que hay más datos

- ▶ **Def:** Un estimador insesgado es **Normal asintóticamente** si

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{\hat{\theta}_n - \theta}{\text{se}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du$$

⇒ Esto es, para muestras suficientemente grandes  $\hat{\theta}_n \sim \mathcal{N}(\theta, \text{se}^2)$

## Ejemplo: lanzar una moneda

**Ej:** Se lanza la misma moneda  $n$  veces y se registra el resultado

- Se modelan las observaciones como  $X_1, \dots, X_n \sim \text{Ber}(p)$ . ¿Estimación de  $p$ ?
- La opción natural es el estimador dado por la **media muestral/promedio**

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Recordar que para  $X \sim \text{Ber}(p)$ :  $\mathbb{E}[X] = p$  y  $\text{var}[X] = p(1 - p)$
- El estimador  $\hat{p}$  es **insesgado** ya que

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = p$$

## Ejemplo lanzamiento monedas (cont.)

- ▶ El error estándar es

$$\text{se} = \sqrt{\text{var} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right]} = \sqrt{\frac{1}{n^2} \sum_{i=1}^n \text{var} [X_i]} = \sqrt{\frac{p(1-p)}{n}}$$

⇒ Pero  $p$  es desconocido. **Estimación del error estándar**  $\hat{\text{se}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- ▶ Dado que  $\hat{p}_n$  es insesgado, resulta que  $\text{MSE} = \mathbb{E} [(\hat{p}_n - p)^2] = \frac{p(1-p)}{n} \rightarrow 0$ 
  - Entonces  $\hat{p}$  converge en sentido del error cuadrático, de donde  $\hat{p}_n \xrightarrow{P} p$
  - Se deduce que  **$\hat{p}$  es un estimador consistente** del parámetro  $p$
- ▶ Además,  $\hat{p}$  es asintóticamente Normal por el Teorema Central del Límite

# Intervalos de confianza

- ▶ Estimación de regiones de  $\Theta$  donde  $\theta$  pertenece con probabilidad alta
- ▶ **Def:** Dada una muestra i.i.d  $X_1, \dots, X_n \sim F$ , un  $1 - \alpha$  **intervalo de confianza** del parámetro  $\theta$  es un intervalo  $C_n = (a, b)$ , donde  $a = a(X_1, \dots, X_n)$  y  $b = b(X_1, \dots, X_n)$  son funciones de los datos tales que:

$$P(\theta \in C_n) = 1 - \alpha, \text{ para todo } \theta \in \Theta$$

⇒ En palabras,  $C_n = (a, b)$  captura a  $\theta$  con probabilidad  $1 - \alpha$

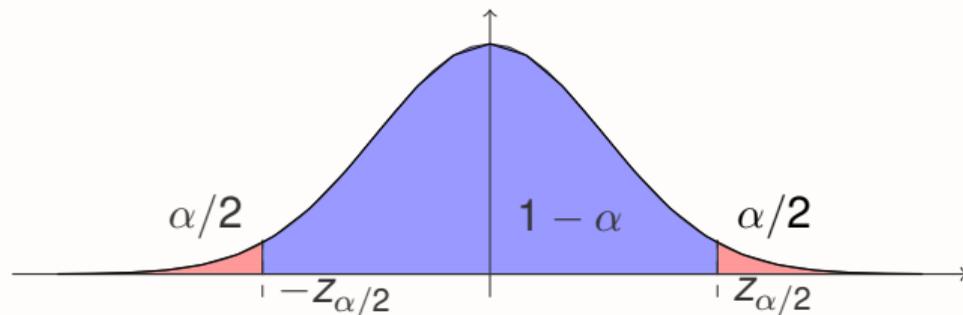
⇒ El intervalo  $C_n$  se calcula a partir de la muestra y por tanto es aleatorio

- ▶ Se llama  $1 - \alpha$  la **cobertura** del intervalo de confianza
- ▶ **Ej:** es común reportar 95 %-intervalos de confianza, i.e.,  $\alpha = 0,05$

## A una lado de la distribución Normal

- ▶ Sea  $X$  una VA Normal estándar, i.e.,  $X \sim \mathcal{N}(0, 1)$  con CDF  $\Phi(x)$

$$\Phi(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du$$



- ▶ Se define  $z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2))$ , i.e., el valor tal que

$$P(X > z_{\alpha/2}) = \alpha/2 \text{ y } P(-z_{\alpha/2} < X < z_{\alpha/2}) = 1 - \alpha$$

# Intervalos de confianza basados en la Normal

- ▶ Los buenos estimadores puntuales  $\hat{\theta}_n$  son normales cuando  $n \rightarrow \infty$ , i.e.,  
 $\hat{\theta}_n \sim \mathcal{N}(\theta, \hat{s}e^2)$   
⇒ Propiedad útil para construir intervalos de confianza de  $\theta$

## Teorema

Supongamos que  $\hat{\theta}_n \sim \mathcal{N}(\theta, \hat{s}e^2)$  cuando  $n \rightarrow \infty$ . Sea  $\Phi$  la CDF de una Normal estándar y sea  $z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2))$ . Se define el intervalo

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\hat{s}e, \hat{\theta}_n + z_{\alpha/2}\hat{s}e).$$

Entonces  $P(\theta \in C_n) \rightarrow 1 - \alpha$ , cuando  $n \rightarrow \infty$

- ▶ Estos intervalos tienen la cobertura correcta solo aproximadamente (para valores grandes de  $n$ )

## Demostración.

- ▶ Sea la VA normal centrada y escalada

$$X_n = \frac{\hat{\theta}_n - \theta}{\hat{s}e}$$

- ▶ Por hipótesis  $X_n \rightarrow X \sim \mathcal{N}(0, 1)$  as  $n \rightarrow \infty$ . Entonces,

$$\begin{aligned} P(\theta \in C_n) &= P\left(\hat{\theta}_n - z_{\alpha/2}\hat{s}e < \theta < \hat{\theta}_n + z_{\alpha/2}\hat{s}e\right) \\ &= P\left(-z_{\alpha/2} < \frac{\hat{\theta}_n - \theta}{\hat{s}e} < z_{\alpha/2}\right) \\ &\rightarrow P\left(-z_{\alpha/2} < X < z_{\alpha/2}\right) = 1 - \alpha \end{aligned}$$

- ▶ La última igualdad vale por definición de  $z_{\alpha/2}$



## Ejemplo lanzamiento de moneda (una vez más...)

**Ej:** Dadas las observaciones  $X_1, \dots, X_n \sim \text{Ber}(p)$ . ¿Estimación de  $p$ ?

- ▶ Estudiamos propiedades del estimador **media muestral/promedio**

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

- ▶ Por el Teorema Central del Límite, se tiene que

$$\hat{p} \sim \mathcal{N} \left( p, \frac{\hat{p}(1 - \hat{p})}{n} \right) \text{ con } n \rightarrow \infty$$

- ▶ Entonces, un  $1 - \alpha$  intervalo de confianza aproximado para  $p$  es

$$C_n = \left( \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right)$$

# Test de Hipótesis

- ▶ Para **test de hipótesis** empezamos con alguna teoría de base que creemos es válida
  - **Ej:** Los datos proviene de una distribución Normal centrada
- ▶ **P:** ¿Los datos brindan evidencia suficiente para rechazar dicha teoría?
- ▶ La teoría a testear se llama **hipótesis nula**, y se escribe  $H_0$ 
  - ⇒ Se plantea también una **hipótesis alternativa** a la nula,  $H_1$
- ▶ Formalmente, dada una muestra i.i.d.  $\mathbf{x} = [x_1, \dots, x_n]^T$  from  $X_1, \dots, X_n \sim F$ 
  - Se define un estadístico del test  $T(\mathbf{x})$ , i.e., una función de la muestra
  - Se define una región crítica (de rechazo)  $\mathcal{R}$  de la forma

$$\mathcal{R} = \{\mathbf{x} : T(\mathbf{x}) > c\}$$

- ▶ Si los datos  $\mathbf{x} \in \mathcal{R}$  se rechaza  $H_0$ , en otro caso decimos que **no hay evidencia suficiente para rechazar  $H_0$**
- ▶ El problema es seleccionar un estadístico del test  $T$  y el **valor crítico  $c$**

## Testeando si una moneda es justa

- Ej:** Se considera el lanzamiento de una misma moneda  $n$  veces y se registran los resultados
- ▶ Modelamos los resultados como  $X_1, \dots, X_n \sim \text{Ber}(p)$ . ¿Podemos saber si es justa la moneda?
  - ▶ Sea  $H_0$  la hipótesis nula de que la moneda es justa, y  $H_1$  la alternativa  
⇒ Podemos escribir las hipótesis como  $H_0 : p = 1/2$  versus  $H_1 : p \neq 1/2$
  - ▶ Consideramos el **estadístico del test** dado por

$$T(X_1, \dots, X_n) = \left| \hat{p}_n - \frac{1}{2} \right| = \left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{2} \right|$$

- ▶ Parece razonable rechazar  $H_0$  si  $(X_1, \dots, X_n) \in \mathcal{R}$ , donde

$$\mathcal{R} = \{(X_1, \dots, X_n) : T(X_1, \dots, X_n) > c\}$$

- ▶ Veremos que esto es un Wald test, de donde  $c = z_{\alpha/2} \hat{s}e$ .

# Tutorial 1: Inferencia sobre la media

- 1 Inferencia estadística y modelos
- 2 Estimación puntual, intervalos de confianza y test de hipótesis
- 3 Tutorial 1: Inferencia sobre la media
- 4 Tutorial 2: Inferencia por regresión lineal

# Inferencia sobre la media

- ▶ Sea una muestra i.i.d de  $n$  observaciones  $X_1, \dots, X_n \sim F$
- ▶ **P:** ¿Cómo podemos hacer **inferencia sobre la media**  $\mu = \mathbb{E}[X_1]$ ?  
⇒ **Problema práctico y canónico en inferencia estadística**
- ▶ Un estimador natural de  $\mu$  es el **promedio/ media muestral**

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

⇒ Justificado ampliamente por la **ley (fuerte) de los grandes números**

$$\lim_{n \rightarrow \infty} \hat{\mu}_n = \mu \quad \text{casi seguramente}$$

- ▶ Es un ejemplo simple de **estimador por método de los momentos (MM)**...
- ▶ ...y también del **estimador por máxima verosimilitud (MLE)**

# Método de los momentos

- ▶ En inferencia paramétrica queremos estimar  $\theta \in \Theta \subseteq \mathbb{R}^p$  en

$$\mathcal{F} = \{f(x; \theta) : \theta \in \Theta\}$$

- ▶ Para  $1 \leq j \leq p$ , se define el  **$j$ -ésimo momento** de  $X \sim F$  como

$$\alpha_j \equiv \alpha_j(\theta) = \mathbb{E} [X^j] = \int_{-\infty}^{\infty} x^j f(x; \theta) dx$$

- ▶ De la misma manera, el  **$j$ -ésimo momento muestral** es un estimador de  $\alpha_j$ , esto es

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^n X_i^j$$

⇒ El  $j$ -ésimo momento  $\alpha_j(\theta)$  depende del parámetro desconocido  $\theta$

⇒ Pero  $\hat{\alpha}_j$  no depende de  $\theta$  ya que es, **una función solo de la muestra**

# Estimador por método de los momentos

- ▶ El primer método para estimación paramétrica es el **método de los momentos**  
⇒ EMM no son óptimos pero son típicamente muy sencillos de calcular
- ▶ **Def:** El **estimador por método de los momentos (EMM)**  $\hat{\theta}_n$  es la solución al sistema

$$\begin{aligned}\alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 \\ &\vdots \\ \alpha_p(\hat{\theta}_n) &= \hat{\alpha}_p\end{aligned}$$

⇒ Es un sistema de  $p$  ecuaciones (no-lineales) con  $p$  incógnitas

- ▶ **Ej:** Volviendo a estimar la media  $\mu$ ,  $p = 1$  y  $\mu = \theta = \alpha_1(\theta)$  de donde

$$\hat{\mu}_n^{MM} = \hat{\alpha}_1 = \frac{1}{n} \sum_{i=1}^n X_i$$

## Ejemplo: modelo de datos gaussianos

**Ej:** Supongamos que  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ , i.e., el modelo es  $F \in \mathcal{F}_N$

► **P:** ¿Cuál es el EMM del vector de parámetros  $\theta = [\mu, \sigma^2]^T$ ?

► Los primeros  $p = 2$  momentos están dados por

$$\alpha_1(\theta) = \mathbb{E}[X_1] = \mu, \quad \alpha_2(\theta) = \mathbb{E}[X_1^2] = \sigma^2 + \mu^2$$

► El EMM  $\hat{\theta}_n$  es la solución al siguiente sistema de ecuaciones

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\hat{\sigma}_n^2 + \hat{\mu}_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$$

► La solución es:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_n)^2$$

# Estimador de máxima verosimilitud

- ▶ A veces “el” método en estimación paramétrica es **máxima verosimilitud**
- ▶ Consideremos una muestra i.i.d.  $X_1, \dots, X_n$  fde una PDF  $f(x; \theta)$
- ▶ La **función de verosimilitud**  $\mathcal{L}_n(\theta) : \Theta \rightarrow \mathbb{R}_+$  se define como

$$\mathcal{L}_n(\theta) := \prod_{i=1}^n f(X_i; \theta)$$

⇒  $\mathcal{L}_n(\theta)$  es la PDF conjunta de la muestra, vista como función de  $\theta$

⇒ La **log-verosimilitud** es  $\ell_n(\theta) := \log \mathcal{L}_n(\theta)$

- ▶ **Def:** El **estimador por máxima verosimilitud (MLE)**  $\hat{\theta}_n$  está dado por

$$\hat{\theta}_n = \arg \max_{\theta} \mathcal{L}_n(\theta)$$

- ▶ **Muy útil:**  $\mathcal{L}_n(\theta)$  y  $\ell_n(\theta)$  alcanzan su máximo en el mismo valor

## Ejemplo: muestra de datos Bernoulli

- ▶ Supongamos  $X_1, \dots, X_n \sim \text{Ber}(p)$ . ¿Cuál es el MLE de  $\mu = p$ ?  
⇒ La densidad conjunta es  $f(x; p) = p^x(1-p)^{1-x}$ ,  $x \in \{0, 1\}$
- ▶ La función de verosimilitud es (definimos  $S_n = \sum_{i=1}^n X_i$ )

$$\mathcal{L}_n(p) = \prod_{i=1}^n f(X_i; p) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i} = p^{S_n} (1-p)^{n-S_n}$$

⇒ La log-verosimilitud es  $\ell_n(p) = S_n \log(p) + (n - S_n) \log(1 - p)$

- ▶ El MLE  $\hat{p}_n$  es la solución a la ecuación

$$\left. \frac{\partial \ell_n(p)}{\partial p} \right|_{p=\hat{p}_n} = \frac{S_n}{\hat{p}_n} - \frac{n - S_n}{1 - \hat{p}_n} = 0$$

- ▶ La solución es

$$\hat{\mu}_n^{ML} = \hat{p}_n = \frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

## Ejemplo: modelo de datos Gaussianos

- ▶ Supongamos  $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$ . ¿Cuál es el MLE de  $\mu$ ?  
⇒ La PDF conjunta es  $f(x; \mu) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x-\mu)^2}{2} \right\}$ ,  $x \in \mathbb{R}$
- ▶ La función de verosimilitud es (a menos de constantes independientes de  $\mu$ )

$$\mathcal{L}_n(\mu) = \prod_{i=1}^n f(X_i; \mu) \propto \exp \left\{ -\sum_{i=1}^n \frac{(X_i - \mu)^2}{2} \right\}$$

⇒ La log-verosimilitud es  $\ell_n(\mu) \propto -\sum_{i=1}^n (X_i - \mu)^2$

- ▶ TEI  $\hat{\mu}_n$  es la solución a la ecuación

$$\left. \frac{\partial \ell_n(\mu)}{\partial \mu} \right|_{\mu=\hat{\mu}_n} = 2 \sum_{i=1}^n (X_i - \hat{\mu}_n) = 0$$

- ▶ La solución es, una vez más, la media muestral

$$\hat{\mu}_n^{ML} = \frac{1}{n} \sum_{i=1}^n X_i$$

# Propiedades del MLE

- ▶ MLE tiene propiedades deseables bajo condiciones generales de la densidad  $f$

**P1) Consistencia:**  $\hat{\theta}_n \xrightarrow{P} \theta$  cuando el tamaño de la muestra  $n$  crece

**P2) Equivariancia:** Si  $\hat{\theta}_n$  es el MLE de  $\theta$ , entonces  $g(\hat{\theta}_n)$  es el MLE de  $g(\theta)$

**P3) Normalidad asintótica:** Para  $n$  grande, se tiene que  $\hat{\theta}_n \sim \mathcal{N}(\theta, \hat{s}\hat{e}^2)$

**P4) Eficiencia:** Para  $n$  grande,  $\hat{\theta}_n$  alcanza la cota inferior de Cramér-Rao

- ▶ Eficiencia significa estimador insesgado con varianza mínima

- ▶ Ej: Se puede usar el MLE para construir un intervalo de confianza para  $\mu$ , i.e.,

$$C_n = \left( \hat{\mu}_n^{ML} - z_{\alpha/2} \hat{s}\hat{e}, \hat{\mu}_n^{ML} + z_{\alpha/2} \hat{s}\hat{e} \right)$$

⇒ Por Normalidad asintótica,  $P(\mu \in C_n) \approx 1 - \alpha$  para  $n$  grande

⇒ Para el modelo  $\mathcal{N}(\mu, 1)$ ,  $\hat{\mu}_n^{ML} \pm \frac{z_{\alpha/2}}{\sqrt{n}}$  tiene cobertura exacta

# El test Wald

- ▶ Considero el siguiente **test de hipótesis** sobre la media  $\mu$

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

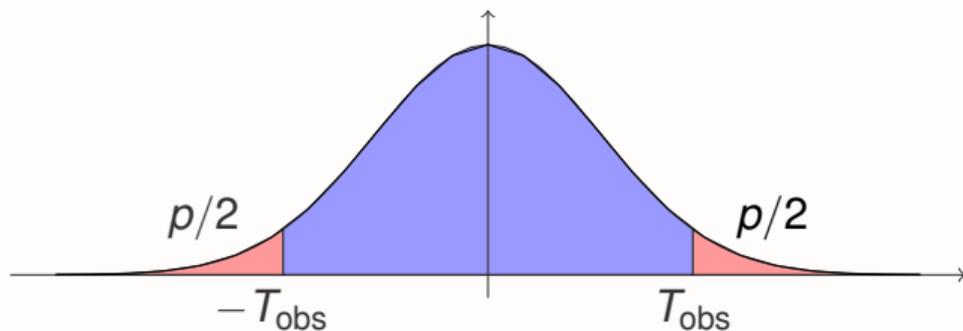
- ▶ Sea  $\hat{\mu}_n$  la media muestral, con error estándar estimado  $\hat{s}_e$
- ▶ **Def:** Dado  $\alpha \in (0, 1)$ , el **test Wald** rechaza  $H_0$  cuando

$$T(X_1, \dots, X_n) := \left| \frac{\hat{\mu}_n - \mu_0}{\hat{s}_e} \right| > z_{\alpha/2}$$

- ▶ Si  $H_0$  es verdadera,  $\frac{\hat{\mu}_n - \mu_0}{\hat{s}_e} \sim \mathcal{N}(0, 1)$  por el TCL
  - ⇒ Probabilidad de rechazar  $H_0$  erróneamente no es más grande que  $\alpha$
- ▶ El valor de  $\alpha$  se llama **nivel de significancia** del test

## El $p$ -valor

- ▶ Reportar “rechazo  $H_0$ ” o “no rechazo  $H_0$ ” para tal  $\alpha$  no es tan informativo  
⇒ Podemos preguntar para cada  $\alpha$ , si se rechaza o no a ese nivel
- ▶ Sea  $T_{\text{obs}} := T(\mathbf{x})$  el valor del estadístico del test para la muestra observada



- ▶ La probabilidad  $p := P_{H_0}(|T(\mathbf{X})| \geq T_{\text{obs}})$  se llama  **$p$ -valor**  
⇒ Valor más chico para el cual el test rechaza  $H_0$
- ▶ Un  $p$ -valor chico ( $< 0,05$ ) indica poca evidencia para soportar la hip. nula  $H_0$

# Tutorial 2: Inferencia por regresión lineal

- 1 Inferencia estadística y modelos
- 2 Estimación puntual, intervalos de confianza y test de hipótesis
- 3 Tutorial 1: Inferencia sobre la media
- 4 Tutorial 2: Inferencia por regresión lineal

# Regresión Lineal

- ▶ Sean observaciones de la forma  $(Y_1, X_1), \dots, (Y_n, X_n) \sim F_{YX}$   
⇒ El objetivo es aprender la relación entre las VAs  $Y$  and  $X$
- ▶ Un primer enfoque es modelar la **función de regresión**

$$r(x) = \mathbb{E} [Y | X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$$

- ▶ El **modelo simple de regresión lineal** especifica que dado  $X_i = x_i$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n$$

- Las  $y_i$ 's se modelan como muestras ruidosas de la recta  $r(x) = \beta_0 + \beta_1 x$
  - Los errores  $\epsilon_i$  son i.i.d., con  $\mathbb{E} [\epsilon_i | X_i = x_i] = 0$  y  $\text{var} [\epsilon_i | X_i = x_i] = \sigma^2$
- ▶ Bajo el modelo lineal, la regresión equivale a inferencia paramétrica

$$\hat{r}(x) \Leftrightarrow [\hat{\beta}_0, \hat{\beta}_1]^T$$

# Regresión Lineal Múltiple

- ▶ Más general, supongamos que se observan datos  $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$   
⇒ cada input  $\mathbf{x}_i = [x_{i1}, \dots, x_{ip}]^T$  es un vector  $p \times 1$  de atributos
- ▶ El **modelo de regresión lineal múltiple** especifica que

$$y_i = \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i, \quad i = 1, \dots, n$$

- Típicamente  $x_{i1} = 1$  para todo  $i$ , dado un término de corte
  - Los errores  $\epsilon_i$  son i.i.d., con  $\mathbb{E}[\epsilon_i | \mathbf{X}_i = \mathbf{x}_i] = 0$  y  $\text{var}[\epsilon_i | \mathbf{X}_i = \mathbf{x}_i] = \sigma^2$
- ▶ De manera compacta(matricial) se representa como  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , definiendo

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

# Estimador de Mínimos Cuadrados

- ▶ Un estimador clásico  $\hat{\beta}$  minimiza la **suma cuadrática de los residuos (RSS)**

$$\text{RSS}(\beta) = \sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2 = \|\mathbf{y} - \mathbf{X}\beta\|^2$$

⇒ Residuos (errores) son las distancias de  $y_i$  al hiperplano  $r(\mathbf{x}) = \beta^T \mathbf{x}$

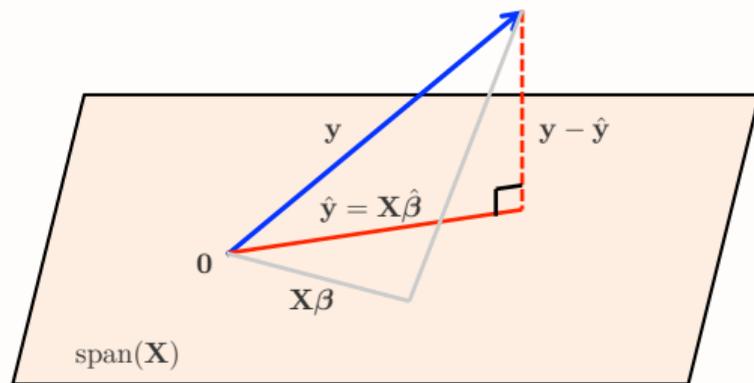
- ▶ **Def:** El **estimador de mínimos cuadrados (LSE)**  $\hat{\beta}_n$  es la solución a

$$\hat{\beta}_n = \arg \min_{\beta} \text{RSS}(\beta)$$

- ▶ Resolviendo la optimización resultad que, el LSE es  $\hat{\beta}_n = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$   
⇒ Definido solo si  $\mathbf{X}^T \mathbf{X}$  es invertible  $\Leftrightarrow \mathbf{X}$  tiene rango completo  $p$

# Geometría del LSE

- ▶ En mínimos cuadrados se busca el vector  $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \in \text{span}(\mathbf{X})$  más cercano a  $\mathbf{y}$



- ▶ Solución: **Proyección ortogonal** de  $\mathbf{y}$  sobre  $\text{span}(\mathbf{X})$ , i.e., (let  $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ )

$$\hat{\mathbf{y}} = P_{\mathbf{X}}(\mathbf{y}) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{U}\mathbf{U}^T\mathbf{y}$$

- ▶ El residuo  $\mathbf{y} - \hat{\mathbf{y}}$  pertenece al complemento ortogonal  $(\text{span}(\mathbf{X}))^\perp$

# Propiedades del LSE

► LSE  $\hat{\beta}_n = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  es una combinación lineal de las VAs  $\mathbf{y}$

**P1) Insesgado:**  $\mathbb{E} [\hat{\beta}_n | \mathbf{X}] = \beta$  con  $\text{var} [\hat{\beta}_n | \mathbf{X}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$

**P2) Consistencia:**  $\hat{\beta}_n \xrightarrow{P} \beta$  cuando el tamaño de la muestra  $n$  crece

**P3) Normalidad asintótica:** Para valores grandes de  $n$ , se tiene que  
 $\hat{\beta}_n \sim \mathcal{N}(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$

**P4)** Si los errores  $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ , entonces  $\hat{\beta}_n \sim \mathcal{N}(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$  exactamente; y

**Eficiencia:** No hay otro estimador insesgado de  $\beta$  con varianza más chica

► **Ej:** Se puede usar el LSE para construir intervalos de confianza para cada  $\beta_j$ :

$$C_n = \left( \hat{\beta}_j - z_{\alpha/2} \hat{\text{se}}(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2} \hat{\text{se}}(\hat{\beta}_j) \right)$$

⇒ Dada la normalidad asintótica (o exacta),  $P(\beta_j \in C_n) \approx 1 - \alpha$

⇒ Observar que  $\hat{\text{se}}(\hat{\beta}_j) = \hat{\sigma} \sqrt{[(\mathbf{X}^T \mathbf{X})^{-1}]_{jj}}$ , donde  $\hat{\sigma}^2 = \frac{\text{RSS}(\hat{\beta})}{n-p}$

# Test de hipótesis y predicción

**Ej:** Sea el **test de hipótesis** respecto del parámetro  $\beta_j$

$$H_0 : \beta_j = \beta_j^{(0)} \quad \text{versus} \quad H_1 : \beta_j \neq \beta_j^{(0)}$$

► Dada la normalidad asintótica (o exacta) del LSE, un test al nivel  $\alpha$  es

$$\text{Rechazar } H_0 \text{ si } T_j := \left| \frac{\hat{\beta}_j - \beta_j^{(0)}}{\hat{\text{se}}(\hat{\beta}_j)} \right| > z_{\alpha/2}$$

**Ej:** Podemos **predecir** un valor no observado  $Y_* = y_*$  a partir del valor  $\mathbf{x}_*$  como

$$y_* = \mathbf{x}_*^T \hat{\boldsymbol{\beta}}$$

► Se puede definir una noción de error estándar para  $y_*$ , e intervalos de confianza

⇒ **debe tener en cuenta la variabilidad en la estimación de  $\boldsymbol{\beta}$  y de  $\epsilon_*$**

## LSE como MLE

- ▶ Supongamos que condicionado a  $\mathbf{X}_i = \mathbf{x}_i$ , los errores  $\epsilon_i$  son Normales i.i.d.  
⇒ La PDF condicional es  $f(\epsilon_i | \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{\epsilon_i^2}{2\sigma^2} \right\}$
- ▶ Asumiendo que  $\sigma^2$  es conocida, la función de verosimilitud (condicional) es

$$\mathcal{L}_n(\beta) = \prod_{i=1}^n f(y_i | \mathbf{x}_i; \beta) \propto \exp \left\{ -\sum_{i=1}^n \frac{(y_i - \beta^T \mathbf{x}_i)^2}{2\sigma^2} \right\}$$

⇒ La log-verosimilitud es  $\ell_n(\beta) \propto -\text{RSS}(\beta)$

- ▶ El MLE  $\hat{\beta}_n^{ML}$  maximiza la log-verosimilitud, entonces

$$\hat{\beta}_n^{ML} = \arg \max_{\beta} \ell_n(\beta) = \arg \min_{\beta} \text{RSS}(\beta) = \hat{\beta}_n^{LS}$$

- ▶ **Ejercicio:** Bajo el modelo Gaussiano lineal, el LSE es también un MLE

# Ridge regression

- LSE penalizado con  $\ell_2$  se conoce como **ridge regression**

$$\hat{\beta}^{ridge} = \arg \min_{\beta} \text{RSS}(\beta) + \lambda \|\beta\|_2^2$$

- Para  $\lambda > 0$ , el estimado "ridge"  $\hat{\beta}^{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$

- Diferente del LSE  $\hat{\beta}^{LS} := \arg \min_{\beta} \text{RSS}(\beta)$
- Es sesgado y sesgo( $\hat{\beta}^{ridge}$ ) crece con  $\lambda$
- Está bien definido incluso si  $\mathbf{X}$  no tiene rango completo

- La pérdida de sesgo se gana en potencial de reducir varianza  $< \text{var} [\hat{\beta}^{LS}]$

- Ej:  $\text{var} [\hat{\beta}^{LS}]$  es grande cuando  $\mathbf{X}$  es cercana a déficit de rango,  $(\mathbf{X}^T \mathbf{X})^{-1}$  inestable

- De la descomposición sesgo-varianza del MSE, compromiso( $\lambda$ ) puede dar

$$\text{MSE}(\hat{\beta}^{ridge}) < \text{MSE}(\hat{\beta}^{LS})$$

# LSE penalizado por complejidad

- Ridge es un ejemplo de una clase más grande de **LSE penalizado por complejidad**

$$\hat{\beta}^J = \arg \min_{\beta} \text{RSS}(\beta) + \lambda J(\beta)$$

- Función  $J(\cdot)$  penaliza (i.e., restringe) los parámetros en  $\beta$
  - El espacio de parámetros restringido  $\Theta$  resulta en modelos 'menos complejos'
  - **Tuneo de  $\lambda$  balancea bondad de ajuste y complejidad**
- **Ej:** penalización con norma  $\ell_1$  resulta en **sparsity**, i.e., selección de variables

