# Bloque 1: Introducción

Pablo Rodríguez-Bocca
Departamento de IO, Instituto de Computación
prbocca@fing.edu.uy

4 de agosto de 2021

# Bloque 1: Introducción

(I) Introducción ($\sim$ 3 clases):

- Utilidad del análisis de redes
  - análisis de datos, aprendizaje automático, datos en formato de redes, visualización, uso de las redes en distintas disciplinas
- La representación de grafos
  - nodos, enlace, matriz de adyacencia, grado de un nodo, redes de varios modos, conectividad ...
- [opcional] Introducción a la Inferencia estadística
- Prácticos 0-A, y 0-B (opcional)

# Inferencia estadística (revisión)

Pablo Rodríguez-Bocca
Departamento de IO, Instituto de Computación
prbocca@fing.edu.uy

4 de agosto de 2021

# Teoría de probabilidad

Teoría de probabilidad (repaso)

Inferencia estadística y modelos

Conceptos fundamentales en la inferencia: estimaciones puntuales, intervalos de confianza y test de hipótesis

Tutorial sobre inferencia en la media

Tutorial sobre inferencia con regresión lineal

Ejemplos de inferencia de datos en redes

# Probabilidad

- Un experimento tiene una salida en el espacio de muestreo $\Omega$
- Un evento $A \subseteq \Omega$ tiene probabilidad $P(A)$, donde $P : \Omega \mapsto [0, 1]$ cumple:
  - $P(\Omega) = 1$,
  - $P(A) \geq 0 \ \forall \ A \subseteq \Omega$,
  - si $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
- probabilidad condicional: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

  $\Rightarrow$ regla de Bayes: $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$
- $A$ y $B$ independientes $\Leftrightarrow P(A \cap B) = P(A)P(B) \Leftrightarrow P(A|B) = P(A)$

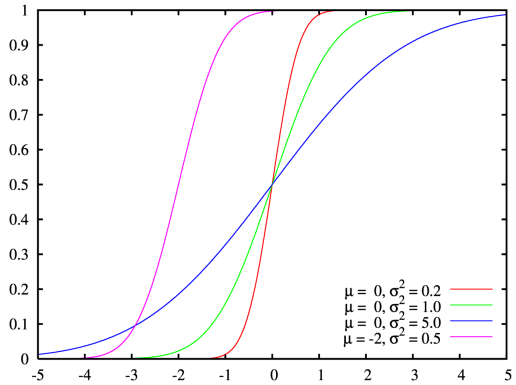
# Variable aleatoria

- variable aleatoria $X$ se usa para formalizar la idea de medidas sobre el experimento, $X : \Omega \mapsto \mathrm{R}$
- *cumulative distribution function (CDF)*:

  $X$ tiene una distribución $F_X$, notación $X \sim F_X$, donde $F_X(x) \equiv P(X \leq x)$

  por lo general se asume una distribución conocida, ejemplo la distribución Normal, $X \sim N(\mu, \sigma)$

  a veces es más útil trabajar con el complemento de la distribución (CCDF), $\overline{F_X}(x) \equiv 1 - F_X(x)$

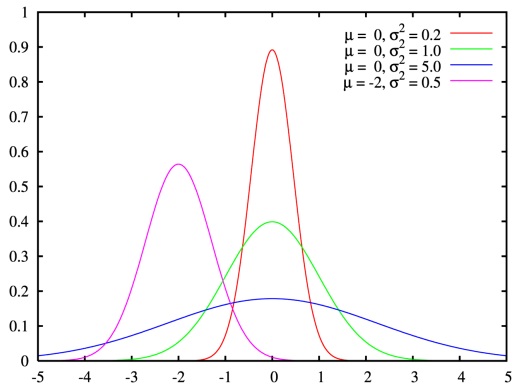



- *probability density function (PDF)*:

  la función de densidad $f_X(x)$ de la variable aleatoria continua $X$

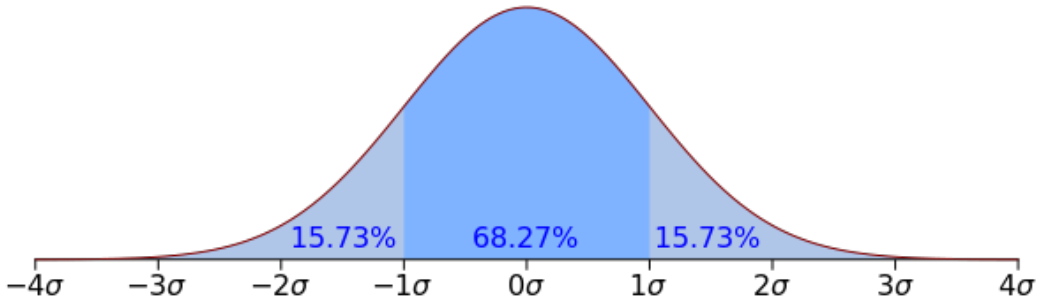
  donde $F_X(x) = \int_{-\text{inf}}^{x} f_X(u)du$ y $f_X(x) = \frac{d}{dx} F_X(x)$

  si $X$ es discreta en realidad se llama *probability mass function (PMF)*



# Estadística de variables aleatorias

- valor esperado de $X$, $\mathrm{E}(X) = \mu_X \equiv \int x f_X(x) dx (= \sum_x x f_X(x))$
- Sea la variable aleatoria $Y = g(X) \Rightarrow \mathrm{E}_Y(Y) = \mathrm{E}_X(g(X))$. Ejemplos:
  - $g(x) = x^k$, $\mathrm{E}(X^k)$ es el momento k-esimo
  - $g(x) = (x - \mu_X)^2$, $\mathrm{E}[(x - \mu_X)^2] = \sigma_X^2$ es la varianza, y $\sigma_X$ la desviación estándar




  - $g(x) = cx + b$, $\mathrm{E}(cx + b) = c\mathrm{E}(x) + b$ y $\sigma^2_{cx+b} = c^2\sigma_X^2$
- También pueden evaluarse estadísticas que combinan variables
  - la covarianza es $cov(X, Y) \equiv \mathrm{E}[(x - \mu_X)(y - \mu_Y)]$ (obs. $cov(X, X) = \sigma_X^2$)
  - la correlación es $\rho_{XY} = corr(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$

# Comportamiento asintótico de variables aleatorias

- Sean $X_1, X_2, \ldots$ una sucesión de variables aleatorias independientes e identicamente distribuidas (i.i.d.)

  $\Rightarrow$ todas comparten la misma media $\mu$ y la misma varianza $\sigma^2$
- sea la media de la muestra $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- ley débil de los grandes números:

$$\lim_{n\to\infty} \mathrm{P}\left(|\overline{X_n} - \mu| \leq \epsilon\right) = 1$$

- ley fuerte de los grandes números

$$\mathrm{P}\left(\lim_{n\to\infty} \overline{X_n} = \mu\right) = 1$$

- teorema central del límite

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \leq x\right) = N(0,1)$$

# Inferencia estadística y modelos

Teoría de probabilidad (repaso)

Inferencia estadística y modelos

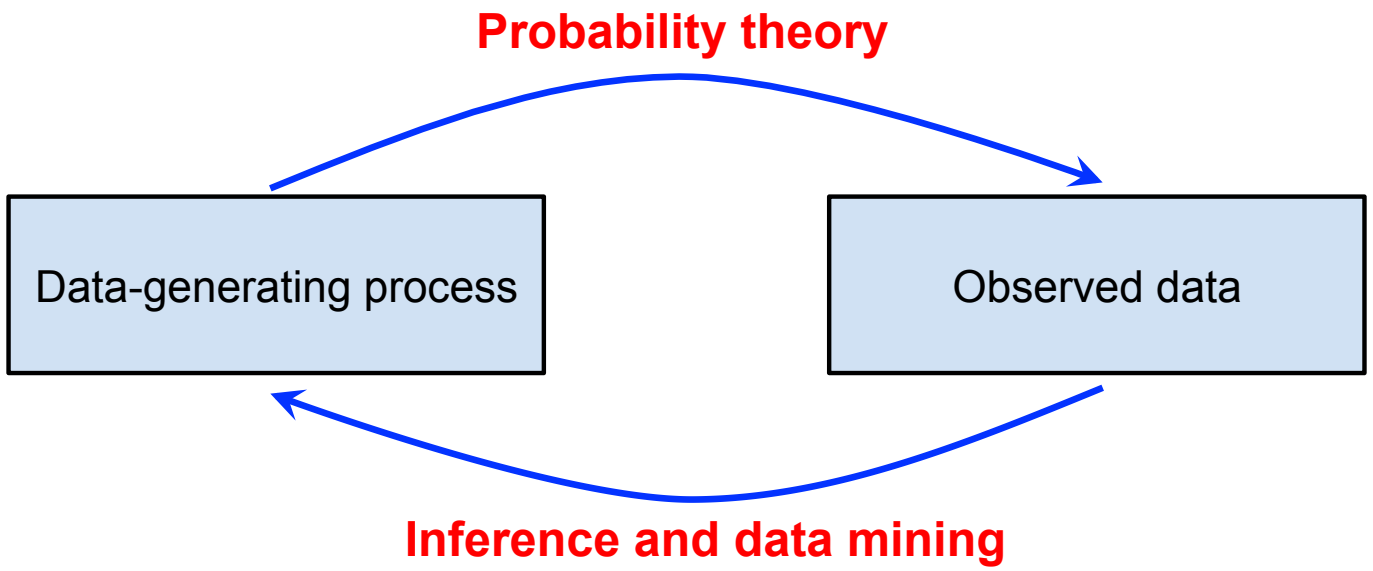
Conceptos fundamentales en la inferencia: estimaciones puntuales, intervalos de confianza y test de hipótesis

Tutorial sobre inferencia en la media

Tutorial sobre inferencia con regresión lineal

Ejemplos de inferencia de datos en redes

# Probabilidad e Inferencia




- Teoría de Probabilidad es un formalismo para trabajar con la incertidumbre
  - Dado un proceso que genera datos, ¿cuáles son las propiedades de los resultados?
- Inferencia estadística trabaja con el problema inverso
  - Dado los resultados, ¿qué se puede decir del proceso generador?

# Inferencia estadística

- Inferencia estadística refiere al proceso por el cual
  - $\Rightarrow$ Dada las observaciones $\mathrm{x} = [x_1, \ldots, x_n]^T$ de $X_1, \ldots, X_n \sim F$
  - $\Rightarrow$ El objetivo es extraer información acerca de la distribución $F$
- Ej: Inferir una propiedad de $F$ como su media
- Ej: Inferir la propia CDF $F$, o la PDF $f = F'$
- Usualmente las observaciones son de la forma $(y_i, x_i)$, $i = 1, \ldots, n$
  - $\Rightarrow$ $Y$ es la respuesta o resultado. $X$ es una predicción o propiedad
- Q: ¿Cuál es la relación ente las variables aleatorias (VAs) $Y$ y $X$?
- Ej: Aprender $\mathbb{E}\left[Y \,\middle|\, X = x\right]$ como una función de $x$
- Ej: Predecir un valor aún no observado $y_*$ a partir de la entrada $X_* = x_*$

# Modelos

- Un modelo estadístico especifica un conjunto $\mathcal{F}$ de CDFs al que $F$ debe pertenecer
- Un modelo paramétrico es de la forma $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$
  - Los parámetro(s) $\theta$ no se conocen, toman valores en el espacio de parámetros $\Theta$
  - El espacio $\Theta$ tiene $\dim(\Theta) < \infty$, y no crece con el tamaño de la muestra $n$
  - Ej: Datos que provienen de una distribución Gausiana

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

$\Rightarrow$ Un modelo de dos parámetros: $\boldsymbol{\theta} = [\mu, \sigma]^T$ y $\Theta = \mathbb{R} \times \mathbb{R}_+$

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

# Modelos y tareas de inferencia

- Dado los datos independientes $\mathrm{x} = [x_1, \ldots, x_n]^T$ de $X_1, \ldots, X_n \sim F$

  $\Rightarrow$ La inferencia estadística es usualmente conducida en el contexto de un modelo

Ej: Estimación paramétrica unidimensional

- Suponga que las observaciones estan distribuidas según Bernoulli con parámetro $p$
- La tarea es estimar el parámetro $p$ (es decir, la media)

Ej: Estimación paramétrica bidimensional

- Suponga que la PDF $f \in \mathcal{F}_N$, es decir, datos gausianos
- El problema es estimar los parámetros $\mu$ y $\sigma$
- Solo se puede estimar $\mu$, y considerar $\sigma$ como un parámetro de ruido

Ej: Estimación no-paramétrica de la CDF

- El objetivo es estimar $F$ asumiendo solo que $F \in \mathcal{F}_{All} = \{\text{All CDFs } F\}$

# Modelos de regresión

- Suponga observaciones de la forma $(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{YX}$

  $\Rightarrow$ Goal: aprender la relación entre las VAs $Y$ y $X$

- Una propuesta típica (y básica) es modelar la función de regresión

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

  $\Rightarrow$ 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 no-paramétrica, asumiendo solo suavidad

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

# Regresión, predicción y clasificación

- Dados los datos $(y_1, x_1), \ldots, (y_n, x_n)$ de $(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{YX}$
  - Ej: $x_i$ es la presión sanguínea de la persona $i$, $y_i$ cuanto tiempo vive
- Modelar la relación entre $Y$ y $X$ usando $r(x) = \mathbb{E}\left[Y \,\middle|\, X = x\right]$

  $\Rightarrow$ Q: ¿Cuáles son las tareas de inferencia clásicas en este contexto?

Ej: Regresión o ajuste de curva

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

Ej: Predicción

- El objetivo es predecir $Y_*$ para una nueva persona basada en su $X_* = x_*$
- Si una estimación de regresión $\hat{r}$ existe, entonces $y_* := \hat{r}(x_*)$

Ej: Clasificación

- Si las VAs $Y_i$ son discretas (ej vivir o morir codificado según $\pm 1$)
- El problema de predicción anterior se denomina clasificación

# Conceptos fundamentales en la inferencia

Teoría de probabilidad (repaso)

Inferencia estadística y modelos

Conceptos fundamentales en la inferencia: estimaciones puntuales, intervalos de confianza y test de hipótesis

Tutorial sobre inferencia en la media

Tutorial sobre inferencia con regresión lineal

Ejemplos de inferencia de datos en redes

# Estimación puntual

- La estimación puntual refiere a hacer una única "mejor estimación" sobre $F$
- **Def:** Dados los datos $\mathrm{x} = [x_1, \ldots, x_n]^T$ de $X_1, \ldots, X_n \sim F$, una estimación puntual $\hat{\theta}$ del parámetro $\theta$ es alguna función

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

  $\Rightarrow$ El estimador $\hat{\theta}$ es calculado de los datos, por lo tanto es una VA

  $\Rightarrow$ La distribución de $\hat{\theta}$ es llamada la distribución de la muestra

- La estimación es el valor específico para una muestra de datos x

  $\Rightarrow$ Puede escribirse $\hat{\theta}_n$ para hacer referencia explícita al tamaño de la muestra

# Sesgo (*bias*), error estándar y el error cuadrático

- **Def:** El sesgo de un estimador $\hat{\theta}$ es $\text{bias}(\hat{\theta}) := \mathbb{E}\left[\hat{\theta}\right] - \theta$
- **Def:** El error estándar es la desviación estándar de $\hat{\theta}$

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

$\Rightarrow$ Usualmente, depende de la $F$ desconocida. Puede hacerse un estimador $\hat{\text{se}}$

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

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

- Los valores esperados $\mathbb{E}[]$ son respecto a la distribución de los datos

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

# La descomposición del MSE en sesgo y varianza

**Theorem**
*El* $MSE = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$ *puede ser escrito:*

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

**Demostración.**

- Sea $\bar{\theta} = \mathbb{E}\left[\hat{\theta}\right]$. Entocnes

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

- La última igualdad es porque $\mathbb{E}\left[\hat{\theta} - \bar{\theta}\right] = \mathbb{E}\left[\hat{\theta}\right] - \bar{\theta} = 0$

□

# Propiedades deseables de los estimadores puntuales

- Q: ¿Qué propiedades son deseables para el estimador $\hat{\theta}$ del parámetro $\theta$?
- **Def:** Un estimador es insesgado si $\text{bias}(\hat{\theta}) = 0$, es decir, si $\mathbb{E}\left[\hat{\theta}\right] = \theta$

  $\Rightarrow$ Un estimador insesgado da "en el blanco" del promedio
- **Def:** Un estimador es consistente si $\hat{\theta}_n \xrightarrow{p} \theta$, es decir, para cualquier $\epsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}\left(|\hat{\theta}_n - \theta| < \epsilon\right) = 1$$

  $\Rightarrow$ Un estimador consistente converge a $\theta$ a medida que se recolectan más datos
- **Def:** Un estimador es asintóticamente Normal si

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

  $\Rightarrow$ Es decir, para una muestra lo suficientemente grande $\hat{\theta}_n \sim \mathcal{N}(\theta, \text{se}^2)$

# Ejemplo: lanzar de moneda

Ej: Se lanza la misma moneda $n$ veces y se registran los resultados

- Se modelan las observaciones como $X_1, \ldots, X_n \sim \text{Ber}(p)$. ¿Cómo estimar $p$?
- La elección natural es el estimador promedio de las muestras

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

- Recordamos, para $X \sim \text{Ber}(p)$, se tiene $\mathbb{E}[X] = p$ y $\text{var}[X] = p(1-p)$
- El estimador $\hat{p}$ es insesgado porque

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

$\Rightarrow$ También se usa que el valor esperado es un operador lineal

# Ejemplo: lanzar de moneda (2)

- El error estándar es

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

  $\Rightarrow$ $p$ desconocido. El error estándar estimado es $\hat{\text{se}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- Dado que $\hat{p}_n$ es insesgado, entonces
  $\text{MSE} = \mathbb{E}\left[(\hat{p}_n - p)^2\right] = \frac{p(1-p)}{n} \to 0$
  - Entonces $\hat{p}$ converge en el sentido de media cuadrática, y por tanto $\hat{p}_n \xrightarrow{p} p$
  - Es decir, $\hat{p}$ es un estimador consistente del parámetro $p$
- También, $\hat{p}$ es asintóticamente Normal por el teorema Central del Límite

# Intervalos de confianza

- Especificar una región de $\Theta$ donde es más probable se encuentre $\theta$
- **Def:** Dados los datos i.i.d. $X_1, \ldots, X_n \sim F$, el $1-\alpha$ intervalo de confianza del parámetro $\theta$ es un intervalo $C_n = (a, b)$, donde $a = a(X_1, \ldots, X_n)$ y $b = b(X_1, \ldots, X_n)$ son funciones de los datos tal que

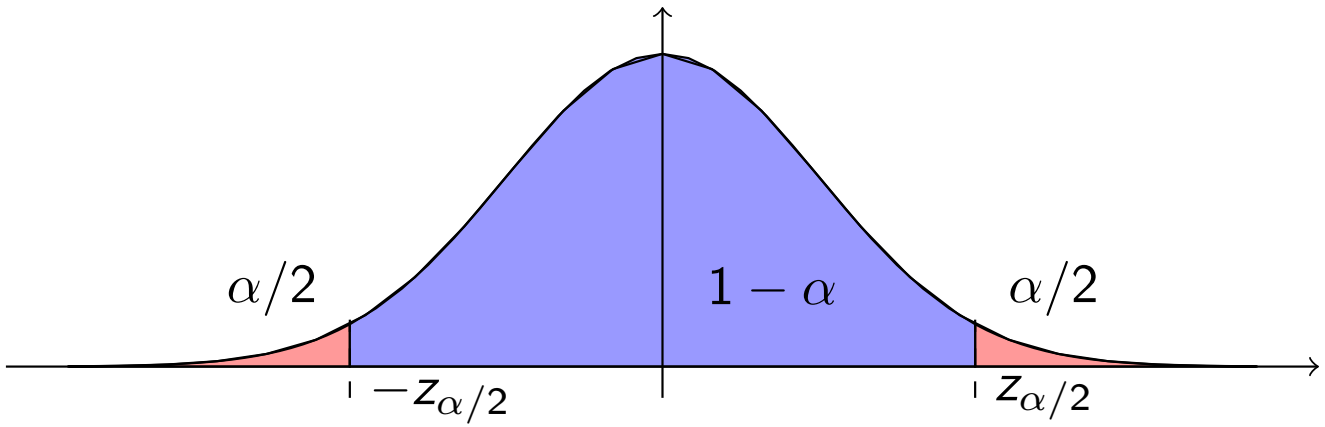
$$\mathrm{P}(\theta \in C_n) = 1 - \alpha, \quad \text{para todo } \theta \in \Theta$$

  - $\Rightarrow$ En palabras, $C_n = (a, b)$ atrapa a $\theta$ con probabilidad $1 - \alpha$
  - $\Rightarrow$ El intervalo $C_n$ es calculado de los datos, entonces es aleatorio
- Llamamos a $1 - \alpha$, la cobertura del intervalo de confianza
- Ej: Usualmente se busca un intevalo de confianza de 95 %, es decir, $\alpha = 0{,}05$

# En la distribución Normal estándar...

- Sea $X$ una VA con distribución Normal estándar, es decir, $X \sim \mathcal{N}(0,1)$ con CDF $\Phi(x)$

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




- Sea $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, es decir, el valor que cumple:

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

# Intervalos de confianza basados en la Normal

- Los buenos estimadores $\hat{\theta}_n$ son Normales cuando $n \to \infty$, es decir, $\hat{\theta}_n \sim \mathcal{N}(\theta, \hat{\text{se}}^2)$

  $\Rightarrow$ Propiedad útil para aproximar los intervalos de confianza de $\theta$

**Theorem**

*Suponga que* $\hat{\theta}_n \sim \mathcal{N}(\theta, \hat{\text{se}}^2)$ *cuando* $n \to \infty$. *Sea* $\Phi$ *la CDF de la Normal estándar y sea* $z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2))$. *Definimos el intervalo*

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

*Entonces* $\mathrm{P}(\theta \in C_n) \to 1 - \alpha$, *cuando* $n \to \infty$

- Estos intervalos solo tienen una cobertura aproximada (para $n$ grande)

# Ejemplo: lanzar de moneda (3)

Ej: Dadas las observaciones $X_1, \ldots, X_n \sim \text{Ber}(p)$. ¿Cómo estimar $p$?

- Estudiamos propiedades del estimador promedio de muestras

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

- Por el teorema Central del Límite, se sabe que

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

- Entonces, una aproximación del intervalo de confianza $1-\alpha$ 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

- En un test de hipótesis se comienza con una teoría (modelo) por defecto
  - Ej: Los datos provienen de una distribución Gausiana con cero de media
- Q: ¿Los datos ofrecen suficiente evidencia para rechazar la teoría?
- La teoría por defecto se llama hipótesis nula, denotada por $H_0$

  $\Rightarrow$ Es necesario definir una hipótesis alternativa a la nula, $H_1$
- Formalmente, dados los datos i.i.d. $\mathrm{x} = [x_1, \ldots, x_n]^T$ de $X_1, \ldots, X_n \sim F$
  (i) Crear una estadística de prueba $T(\mathrm{x})$, es decir, una función de los datos
  (ii) Definir la región de rechazo $\mathcal{R}$

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

- Si los datos $\mathrm{x} \in \mathcal{R}$ entonces rechazamos $H_0$, en otro caso mantenemos (no rechazamos) $H_0$
- El problema es seleccionar la estadística de prueba $T$ y el valor crítico $c$

# Testear si una moneda es justa

Ej: Se lanza la misma moneda $n$ veces y se recolectan los resultados

- Se modelan las observaciones como $X_1, \ldots, X_n \sim \text{Ber}(p)$. ¿La moneda es justa?
- Sea $H_0$ la hipótesis donde la moneda es justa, y $H_1$ la alternativa
  $\Rightarrow$ Podemos escribir las hipótesis:

$$H_0 : p = 1/2 \quad \text{versus} \quad H_1 : p \neq 1/2$$

- Sea la estadística de prueba dada por

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

- Es razonable rechazar $H_0$ si $(X_1, \ldots, X_n) \in \mathcal{R}$, donde

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

- Veremos que este es el test de Wald, y $c = z_{\alpha/2}\hat{\text{se}}$. Más próximamente

# Tutorial sobre inferencia en la media

Teoría de probabilidad (repaso)

Inferencia estadística y modelos

Conceptos fundamentales en la inferencia: estimaciones puntuales, intervalos de confianza y test de hipótesis

Tutorial sobre inferencia en la media

Tutorial sobre inferencia con regresión lineal

Ejemplos de inferencia de datos en redes

# Inferencia en la media

- Sea la muestra de $n$ observaciones i.i.d. de $X_1, \ldots, X_n \sim F$
- Q: ¿Cómo se puede hacer inferencia en la media $\mu = \mathbb{E}[X_1]$?

  $\Rightarrow$ Problema práctico y clásico de la inferencia estadística
- Un estimador natural de $\mu$ es el estimador promedio de la muestra

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

- Bien motivado, ya que por la ley fuerte de los grandes números

$$\lim_{n \to \infty} \hat{\mu}_n = \mu \quad \text{con prácticamente toda seguridad}$$

- Es un ejemplo simple del estimador con el método de los momentos (MME)...

  ...y también del estimador de máxima verosimilitud (MLE)

# Momentos y momentos de la muestra

- En inferencia paramétrica, deseamos estimar $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ en

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

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

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

- Igualmente, el *j*-simo momento de la muestra es una estimación de $\alpha_j$, es decir:

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

  $\Rightarrow$ El *j*-simo momento $\alpha_j(\boldsymbol{\theta})$ depende del $\boldsymbol{\theta}$ desconcido

  $\Rightarrow$ Pero $\hat{\alpha}_j$ no depende, es solo función de los datos

# Estimador con el Método de los Momentos (MME)

- Un primer método para la estimación paramétrica es el método de los momentos

  $\Rightarrow$ MMEs no son óptimos, pero usualmente fácil de calcular

- **Def:** El estimador del método de los momentos (MME) $\hat{\boldsymbol{\theta}}_n$ es la solución de

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

  $\Rightarrow$ Un sistema de $p$ ecuaciones (no lineales) con $p$ incógnitas

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

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

# Ejemplo: modelo de datos Gausianos

Ej: Suponga que $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, es decir, el modelo es $F \in \mathcal{F}_N$

- Q: ¿Cúal es el MME del vector de parámetros $\boldsymbol{\theta} = [\mu, \sigma^2]^T$?
- Los primeros $p = 2$ momentos son

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

- El MME $\hat{\boldsymbol{\theta}}_n$ es la solución del siguiente sistema de ecuaciones

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

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

- La solución es

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

# Estimador de Máxima Verosimilitud (MLE)

- Usualmente "el" método para la estimación paramétrica es la máxima verosimilitud
- Sean los datos i.i.d. $X_1, \ldots, X_n$ de una PDF $f(x; \theta)$
- La función de verosimilitud $\mathcal{L}_n(\theta) : \Theta \to \mathbb{R}_+$ se define como

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

$\Rightarrow \mathcal{L}_n(\theta)$ es la PDF conjunta de los datos, considerados como función de $\theta$

$\Rightarrow$ La función de verosimilitud logarítmica es $\ell_n(\theta) := \log \mathcal{L}_n(\theta)$

- **Def:** El estimador de máxima verosimilitud (MLE) $\hat{\boldsymbol{\theta}}_n$ es

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

- Muy útil: El maximizador de $\mathcal{L}_n(\theta)$ coincide con el de $\ell_n(\theta)$

# Ejemplo: Modelo de datos Bernoulli

- Sea $X_1, \ldots, X_n \sim \text{Ber}(p)$. ¿Cuál es el MLE de $\mu = p$?

  $\Rightarrow$ El PMF de los datos es $f(x; p) = p^x (1-p)^{1-x}$, $x \in \{0, 1\}$

- La función de verosimilitud es (donde $S_n = \sum_{i=1}^n X_i$)

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

  $\Rightarrow$ La verosimilitud logarítmica es

$$\ell_n(p) = S_n \log(p) + (n - S_n)\log(1-p)$$

- El MLE $\hat{p}_n$ es la solución de 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 Gausiano

- Sea $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$. ¿Cuál es el MLE de $\mu$?

  $\Rightarrow$ El PDF de los datos 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 (sin considerar constantes independientes de $\mu$)

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

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

- El MLE $\hat{\mu}_n$ es la solución de 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, nuevamente el estimador promedio de la muestra

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

# Propiedades del MLE

- MLEs tienen propiedades deseables con bajas condiciones sobre $f(x;\theta)$

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

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

P3) Asintóticamente Normal: Para $n$ grandes, se tiene $\hat{\theta}_n \sim \mathcal{N}(\theta, \hat{\text{se}}^2)$

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

- La eficiencia significa que no hay otro estimador insesgado con menor varianza

- Ej: Se puede usar el MLE para crear un intervalo de confianza para $\mu$, es decir,

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

$\Rightarrow$ Por ser asintóticamente Normal, $\mathrm{P}\left(\mu \in C_n\right) \approx 1-\alpha$ para $n$ grandes

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

# El test de Wald

- Considerar el siguiente test de hipótesis respecto a la media $\mu$

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

- Sea $\hat{\mu}_n$ la media de la muestra, con error estándar estimado $\hat{\text{se}}$
- **Def:** Sea $\alpha \in (0, 1)$, el test de Wald rechaza $H_0$ cuando

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

- Si $H_0$ es verdadero, $T(X_1, \ldots, X_n) \sim \mathcal{N}(0, 1)$ por el teorema Central del Límite

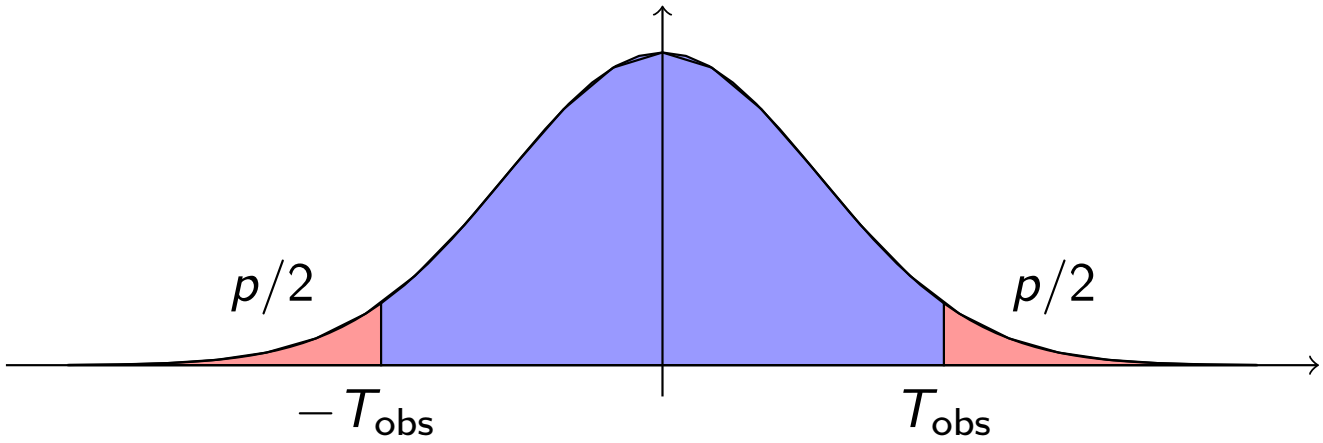
  $\Rightarrow$ La probabilidad de rechazar incorrectamente $H_0$ es a lo sumo $\alpha$

- El valor de $\alpha$ se llama nivel de significación del test

# El *p-value*

- Reportar que "se rechaza $H_0$" o "se mantiene $H_0$" no es muy informativo

  $\Rightarrow$ Se puede pedir, para cada $\alpha$, si la prueba se rechaza en ese nivel
- Sea $T_{\text{obs}} := T(\text{x})$ el valor de la estadística de prueba para la muestra observada




- La probabilidad $p := \text{P}_{H_0}(|T(\text{X})| \geq T_{\text{obs}})$ se llama *p-value*

  $\Rightarrow$ El valor más pequeño en el que rechazaríamos $H_0$
- Un pequeño *p*-value ($< 0{,}05$) indica una reducida evidencia para soportar $H_0$

# Inferencia Bayesiana

- Los métodos discutidos hasta ahora se denominan frecuentistas, donde
  - F1: La probabilidad se refiere a las frecuencias relativas límite
  - F2: Los parámetros son fijos, constantes desconocidas
  - F3: Los procedimientos ofrecen garantías sobre el rendimiento a largo plazo
- Alternativamente, la inferencia bayesiana se basa en
  - B1: La probabilidad describe el grado de creencia, no la frecuencia límite
  - B2: Podemos hacer declaraciones de probabilidad sobre los parámetros
  - B3: Una distribución de probabilidad para $\theta$ se crea para hacer inferencias
- ¿Polémico? Intrínsecamente abarca una noción subjetiva de probabilidad
  - Los métodos bayesianos no ofrecen garantías de rendimiento a largo plazo
  - Muy útil para combinar convicciones previas (*prior*) con datos

# Inferencia Bayesiana: idea de base

- De la probabilidad condicional $P(A|B) = \frac{P(A \cap B)}{P(B)}$, se obtiene la regla bayesiana

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

.

- Si usamos la regla como inferencia

$$P(modelo|datos) = \frac{P(datos|modelo)P(modelo)}{P(datos)}$$

donde

  - los datos es lo que se obtiene de un experimento
  - el modelo es nuestra creencia de la realidad, nuestro conocimiento a priori

- Finalmente, en la inferencia bayesiana se aplica

$$P(modelo|datos) \propto P(datos|modelo)P(modelo)$$

# Ejemplo: Clicks en banner

- Los datos son: $L = \#$ impresiones de un banner, y $C = \#$ clicks
- Paso 1: Suponemos un modelo, en nuestro caso una distribución de CTR (*Click Through Rate*). Sea $\theta$ la VA que representa los valores posibles de CTR, y su distribución a priori es $\theta \sim Beta(\alpha, \beta)$. Por tanto $P(modelo)$ es

$$P(\theta) = f_{\alpha,\beta}(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

- Paso 2: Si el modelo es correcto, la realidad del experimento permite calcular la probabilidad de verosimilitud $P(datos \mid modelo)$. Si las impresiones son independientes $X_1, \ldots, X_n \sim \text{Ber}(\theta)$, la probabilidad de tener $C$ clicks en $L$ impresiones es la distribución binomial:

$$P(L \text{ impresiones, } C \text{ clicks } \mid \theta) = \binom{L}{C} \theta^C (1-\theta)^{L-C}$$

- Paso 3: Cuentas y obtenemos la distribución a posteriori: $P(modelo \mid datos)$ (en este ejemplo se simplifica mucho):

$$P(\theta \mid L \text{ impresiones, } C \text{ clicks }) = f_{\alpha+C,\beta+L-C}(\theta)$$

# Ejemplo: Clicks en banner (cont)

Esta inferencia permite responder muchas preguntas:

- Podemos acotar el CTR, $\theta \in [a, b]$ con un 95 % de certidumbre si:

$$P(a < \theta < b) = \int_a^b f_{\alpha+C,\beta+L-C}(\theta)d\theta > 0{,}95$$

- Dada la cantidad de impresiones $L'$, tenemos la cantidad esperada de clicks $C'$ para esa certidumbre:

$$C' = L' \int_a^b f_{\alpha+C,\beta+L-C}(\theta)d\theta$$

- Si estamos decidiendo entre dos banners (testing A/B)
  - $L_A, C_A$ son las impresiones y clicks para el banner A de control
  - $L_B, C_B$ son las impresiones y clicks para el banner B de prueba
  - No se conoce a priori nada de los CTRs de ambos tests. Asumimos $\alpha_A = \alpha_B = \beta_A = \beta_B = 1$, e inferimos un CTR para cada test: $\theta_A$ y $\theta_B$
  - Entonces la probabilidad de que la opción B sea mejor que la opción A es:

$$P(\theta_B > \theta_A) = \sum_{i=0}^{C_B} \frac{B(1 + C_A + i, 1 + L_B - C_B + 1 + L_A - C_A)}{(1 + L_B - C_B + i)B(1 + i, 1 + L_B - C_B)B(1 + C_A, 1 + L_A - C_A)}$$

# Pasos usuales de la inferencia bayesiana

Paso 1: Elegir la densidad de probabilidad $f(\theta)$ llamada distribución a priori

- A priori expresa nuestra creencia sobre $\theta$, antes de ver los datos

Paso 2: Elegir un modelo estadístico $f(x \mid \theta)$ (comparar con $f(x; \theta)$)

- Refleja nuestras creencias sobre el proceso de generación de los datos, es decir, $X$ dado $\theta$

Paso 3: Dado los datos $\mathbf{X} = [X_1, \ldots, X_n]^T$, actualizamos nuestras creencias y calculamos la distribución a posteriori $f(\theta|\mathbf{X})$ usando la regla de Bayes

$$f(\theta|\mathbf{X}) \propto \prod_{i=1}^{n} f(X_i \mid \theta) f(\theta) = \mathcal{L}_n(\theta) f(\theta)$$

$\Rightarrow$ Usando $f(\theta|\mathbf{X})$ podemos obtener estimaciones puntuales, intervalos de confianza, etc.

# Tutorial sobre inferencia con regresión lineal

Teoría de probabilidad (repaso)

Inferencia estadística y modelos

Conceptos fundamentales en la inferencia: estimaciones puntuales, intervalos de confianza y test de hipótesis

Tutorial sobre inferencia en la media

Tutorial sobre inferencia con regresión lineal

Ejemplos de inferencia de datos en redes

# Regresión lineal

- Suponga las observaciones son de $(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{YX}$
  $\Rightarrow$ Goal: aprender la relación entre las VAs $Y$ y $X$
- Una opción típica (y básica) es modelar la función de regresión

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

- El modelo simple de regresión lineal especifica que dados $X_i = x_i$

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

  - Los $y_i$ son modelados como muestras con ruido 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 $\text{var}[\epsilon_i | X_i = x_i] = \sigma^2$
- Con el modelo lineal, la regresión equivale a la inferencia paramétrica

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

# Regresión lineal múltiple

- Más general, supongo las observaciones son $(y_1, \mathrm{x}_1), \ldots, (y_n, \mathrm{x}_n)$

  $\Rightarrow$ Cada valor $\mathrm{x}_i = [x_{i1}, \ldots, x_{ip}]^T$ es un vector $p \times 1$ de características

- La regresión lineal múltiple especifica

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

  - Típicamente $x_{i1} = 1$ para todo $i$, proporcionando un término de intercepción
  - Los errores $\epsilon_i$ son i.i.d., con $\mathbb{E}[\epsilon_i | X_i = x_i] = 0$ y $\mathrm{var}[\epsilon_i | X_i = x_i] = \sigma^2$

- Se puede representar matricialmente, $\mathrm{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, donde

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

# Estimadores y uso en predicción

- Hay varias formas de estimar los parámetros $\hat{\beta}$ de la regresión

  $\Rightarrow$ La más común es con mínimos cuadrados, minimizando el residuo cuadrático (RSS)

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

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

- Se puede predecir un valor no observado $Y_* = y_*$ a partir de $\text{x}_*$ usando

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

- En las redes: podríamos aplicar regresión lineal para estimar alguna característica de los vértices en función de las otras características. Pero no incorporaría la información de red...

# Ejemplos de inferencia de datos en redes

Teoría de probabilidad (repaso)

Inferencia estadística y modelos

Conceptos fundamentales en la inferencia: estimaciones puntuales, intervalos de confianza y test de hipótesis

Tutorial sobre inferencia en la media

Tutorial sobre inferencia con regresión lineal

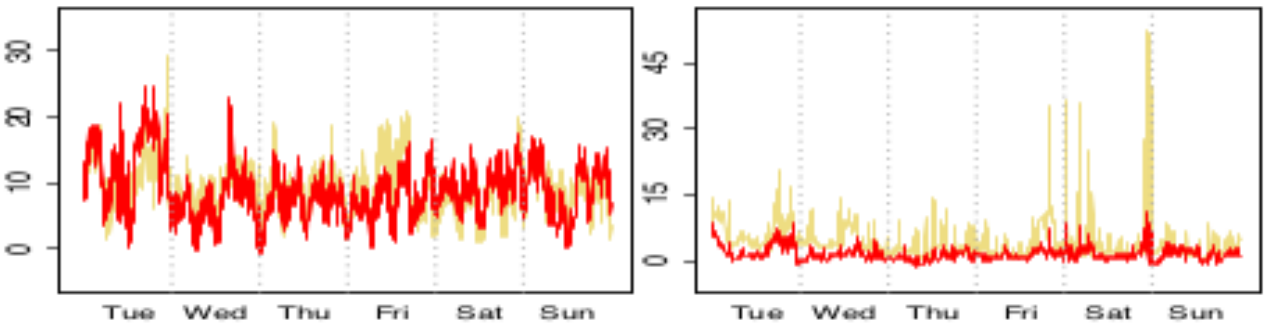
Ejemplos de inferencia de datos en redes

# Ejemplo: Inferencia de la matríz de tráfico de Internet

- Q: ¿Porqué los ISPs monitorean sus redes?
  - R1) Identificar fallas en la red, su impacto y razones
  - R2) Ajustar el ruteo (corto plazo) $\rightarrow$ controlar congestión $\rightarrow$ optimizar QoS
  - R3) Ingeniería de tráfico (mediano plazo) $\rightarrow$ planificación de capacidad
  - R4) Políticas de seguridad contra ciber-ataques (worms, DoS)
- Si la red del ISP se representa con el grafo $G = (V, E)$, entonces la matriz de tráfico tiene $N_v^2$ flujos, los cuales no se monitorean, sino que se monitorea el tráfico agregado en cada enlace $N_e$
- Ejemplo, backbone Abilene: $N_v = 11$ PoPs, $N_e = 30$ enlaces, $N_f = 110$ flujos

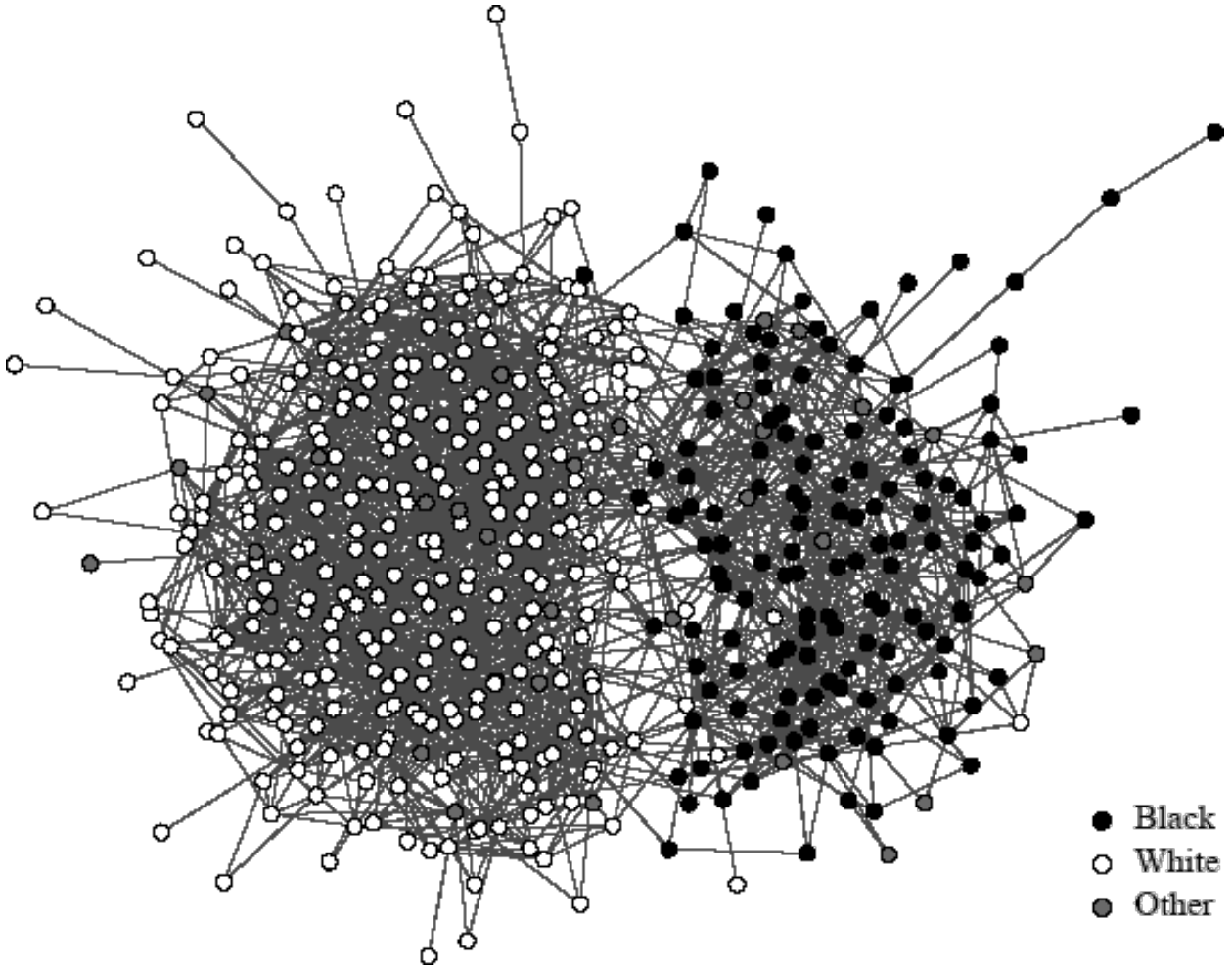
# Ejemplo: Inferencia de la matríz de tráfico de Internet (cont)

- Formalizando, sean:
  - $x \in R^{N_v^2}$: vector de flujo de tráfico a estimar
  - $y \in R^{N_e}$: vector de tráfico en enlaces
  - $B \in R^{N_e x N_v^2}$: matriz binaria de ruteo (si el enlace es usado al ir entre dos vértices)
- Sabemos que los flujos se agregan en cada enlace: $E(Y|X = x) = Bx$
  - no se puede resolver como una simple regresión lineal $Y = Bx + \epsilon$ porque $N_v^2 \gg N_e$
  - se puede resolver usando $\hat{x} = \text{argmax}_x [(y - Bx)^T (y - Bx) + \lambda J(x)]$ donde $J(.)$ es una función de penalización
- Se obtienen buenos resultados: flujos reales y su estimación KF



# Ejemplo: Inferencia y mezcla selectiva

- Las personas tienen fuerte tendencia de asociarse con iguales → llamada homofilia o mezcla selectiva
  - Ej: relaciones románticas en secundaria de EE.UU. . .



# Ejemplo: Inferencia y mezcla selectiva (cont)

- Sea $G = (V, E)$ un grafo de relaciones de amistad. Sea Y su matriz de adyacencia
- mediante regresión lineal, se puede estimar una relación en base a medidas en las personas ($x \sim$ genero, edad,...):

$$P(Y_{ij}|X_i = x_i, X_j = x_j)$$

- o mejor aún, agregar información de otras amistades en la red (ya no es regresión lineal):

$$P(Y_{ij}|Y_{(-ij)}, X_i = x_i, X_j = x_j)$$

- Más adelante veremos como incorporar la información de red en las predicciones...

# Glosario

- Statistical inference
- Outcome or response
- Predictor, feature or regressor
- (Non) parametric model
- Nuisance parameter
- Regression function
- Prediction
- Classification
- Point and set estimation
- Estimator and estimate
- Standard error
- Consistent estimator
- Confidence interval
- Hypothesis test
- Null hypothesis
- Test statistic and critical value
- Method of moments estimator
- Maximum likelihood estimator
- Likelihood function
- Significance level and $p - value$
- Prior and posterior distribution

# Lectura guiada y opcional

- Lectura guiada:

  [SAND] Capítulo 2, Sección 2.2.

- No se cubre este contenido introductorio en [NE]
- Se recomienda como opcional realizar un curso introductorio a la inferencia. Por ejemplo:

  `https://es.coursera.org/learn/statistical-inference`