Cadenas de Markov
+
Geometría de la Probabilidad
+
Estadística

Clase 18 Curso Propedéutico
2016/06/27

ITAM

Cadenas de Markov

- El estado próximo depende sólo del estado actual.
- Más precisamente, las *probabilidades de transición* están determinadas por el estado actual.

- **Motivación inicial:** un modelo para estudiar el lenguaje.

A. A. Марков (1886).

- Para conocer las probabilidades del futuro solo necesitas conocer el presente!

Matemáticamente…

- Si S son todos los posibles "estados" de la cadena (los valores que pueden tomar las X_t) **la propiedad de Markov dice:**

$$\mathbb{P}(X_t = x \mid X_0 = x_0, \ldots, X_{t-1} = x_{t-1}) = \mathbb{P}(X_t = x \mid X_{t-1} = x_{t-1})$$

Cadenas homogéneas

$$P(X_t = j \mid X_{t-1} = i) = P(X_1 = j \mid X_0 = i)$$

Más general

$$P(X_k = j \mid X_s = i) = P(X_{t+k} = j \mid X_t = i)$$

No ejemplo :

Qué es una cadena de Markov

- Es tres cosas:

1. Un conjunto de estados S
2. Una distribución inicial para X_0
3. Un "Kernel" de transición. Unas probabilidades de transición por cada estado actual.

S finito…. El kernel puede representarse como matriz $P = (P_{ij})$

$$P_{ij} = P(X_{t+1} = j \mid X_t = i)$$

Distribución de X_n dado X_0

- La vez pasado vimos que P^n representa las probabilidades de transición en n pasos.
- Si π es un vector tal que $\pi_j = P(X_0 = j)$ entonces tenemos que

$$P(X_1 = j) = \sum_i \pi_i P_{ij}$$

Por lo tanto $\pi^\top P$ es un vector que representa la distribución de X_1.

En general, $\pi^\top P^n$ es la distribución de X_n

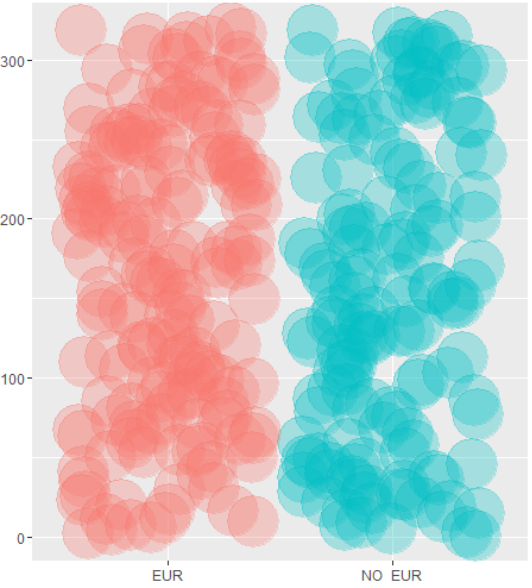
Teorema ergódico

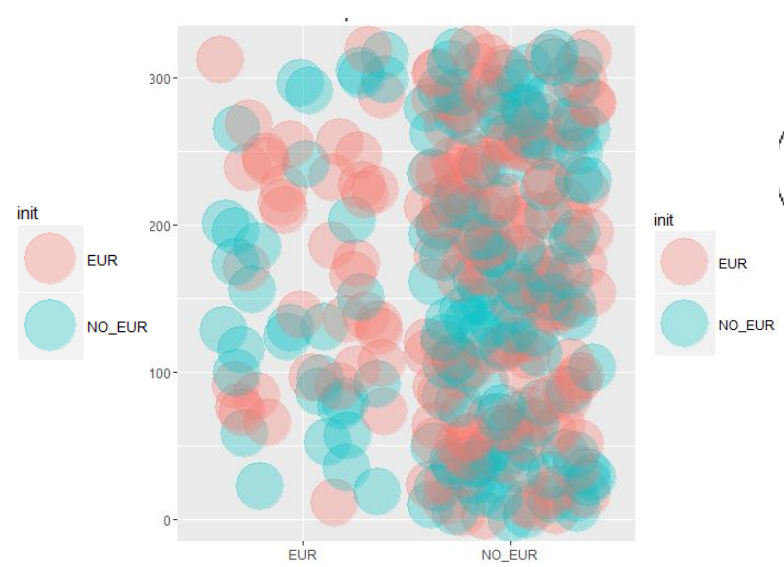
- El resultado más fundamental de las cadenas de Markov establece que si la cadenas es **"regular"** entonces en el largo plazo no importa el estado inicial, la distribución de X_n no depende de X_0.
- Tal distribución se conoce como **distribución estacionaria/límite** y cumple.

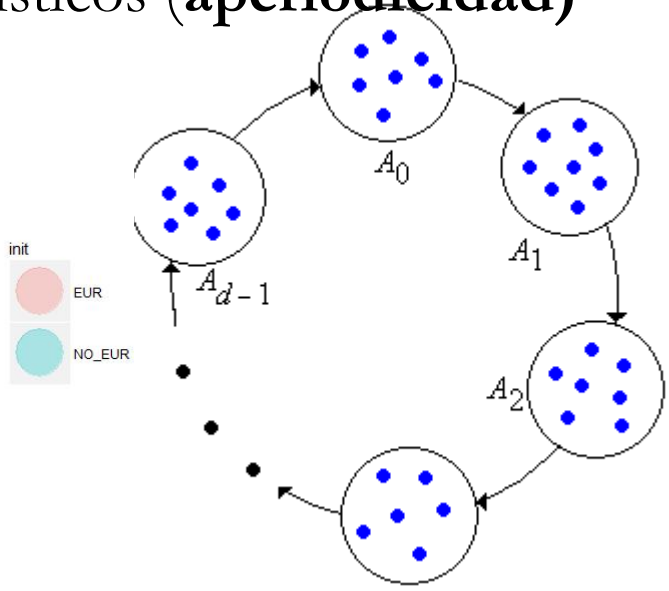
$\pi = P\pi$ $\quad$ $\pi = \lim_{k\to\infty} \pi^{(k)}$ $\quad$ $\pi_j^{(k)} = P(X_j = k)$

Y como consecuencia:

- Aunque no podemos verlo con detalle, hay que recordar un caso particular en el que el *teorema ergódico* se cumple siempre que:
    1. Todos los estados están conectados en el sentido de que existe alguno camino de estados tal que siempre se puede llegar de un estado a otro y de regreso (**irreducibilidad).**
    2. El tiempo esperado de retorno a cada estado es finito (**recurrencia positiva**)
    3. La cadena no tiene *ciclos* determinísticos (**aperiodicidad)**







Google PageRank

- **Motivación:** El problema de un "buscador": un usuario introduce una *búsqueda*, las páginas web se encuentras *indexadas*.
- *S* representa el conjunto de páginas web que responden a la búsqueda
- **Objetivo:** En qué orden listar *S* al usario.

- Necesitamos un método de **rankeo**!

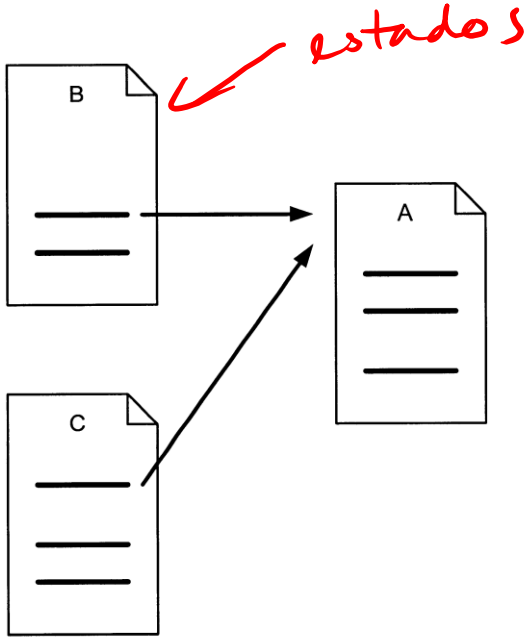


FIG. 1

Approach: Define una cadena de Markov que represente cómo se mueve un "usuario" a través de páginas web y ver que estados son los más visitados en el largo plazo.

¿Cómo construir P?

- Una propuesta:

Estimaremos las probas de transición basados en contar qué porcentaje de los *links* dirigen a cada página.

clicks
historial etc ← construir probas de transición

Posibles Problemas par que no exista la distribución límite?????

Podría no ser irreducible

- Los estados podrían estar desconectados!!!!
- Para eso se agrega un pequeño "ruido" que corresponde a moverse completamente aleatoriamente entre cada página web.

- Se escoge un número $\alpha \in (0,1)$ (e.g $\alpha \approx .85$) y se pone

$$p_{ij} = \alpha \frac{l_{ij}}{\sum_k l_{ik}} + (1-\alpha)\frac{1}{n}$$

Donde n es el # total de páginas web y l_{ij} el total de link en la página i con destino j.

LA GEOMETRÍA (Y ÁLGEBRA LINEAL) DE LA PROBABILIDAD

Recordatorios y Observaciones

1. Espacio vectorial: conjunto cuyos elementos sabemos sumar y multiplicar por un escalar, y es cerrado bajo estas operaciones

Obs 1: Sabemos sumar variables aleatorias

$X, Y : \Omega \to \mathbb{R}$

$(X+Y)(\omega) = X(\omega) + Y(\omega)$

También se pueden multiplicar por escalar

$X : \Omega \to \mathbb{R}$ $\quad a \in \mathbb{R}$

$(aX)(\omega) = a(X(\omega))$

(Esto es cierto para funciones, no solo v.a.s)

Las variables aleatorias son un espacio vectorial (de dimensión infinita no tiene base)

2. Propiedades del producto punto en $\mathbb{R}^n$

$\langle x, y \rangle = x^T y = x \cdot y = \sum_{i=1}^{n} x_i y_i$

operador bilineal

- $\langle x, x \rangle \geq 0$ $\quad \langle x, x \rangle = 0 \iff x = 0$
- $\langle x, y \rangle = \langle y, x \rangle$
- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ y $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$

Obs 2: En variables aleatorias $\langle X, Y \rangle = \mathbb{E}(XY)$ cumple las mismas

propiedades salvo $\langle X,X\rangle = \mathbb{E}(X^2)=0 \iff X=0$. En probabilidad $\mathbb{E}(X^2=0) \Rightarrow \mathbb{P}(X=0)=1$ pero podría existir ω con $X(\omega)\neq 0$ pero $\mathbb{P}(X\neq 0)=0$, entonces es "casi" como $X=0$. Formalmente se dice $X=0$ casi seguramente. Por lo tanto, $\langle X,Y\rangle = E(XY)$ es un producto punto.

$$\cos(\theta(\vec{x},\vec{y})) = \frac{\langle x,y\rangle}{\sqrt{\langle x,x\rangle\langle y,y\rangle}} \Rightarrow \text{En v.a.s} \quad \theta(X,Y) = \cos^{-1}\left(\frac{\mathbb{E}(XY)}{\sqrt{\mathbb{E}(X^2)\,\mathbb{E}(Y^2)}}\right)$$

¡Existen ángulos entre variables aleatorias! (Las vas son espacios de Hilbert)

3. Relación entre norma y producto punto

$$\langle X,X\rangle = \|X\|^2$$

En v.a.s $\|X\| = \sqrt{\mathbb{E}(X^2)} = \sqrt{\langle X,X\rangle}$

En $\mathbb{R}^n$ $\|x\|_2 = \sqrt{\sum x_i^2} = \sqrt{\langle x,x\rangle}$

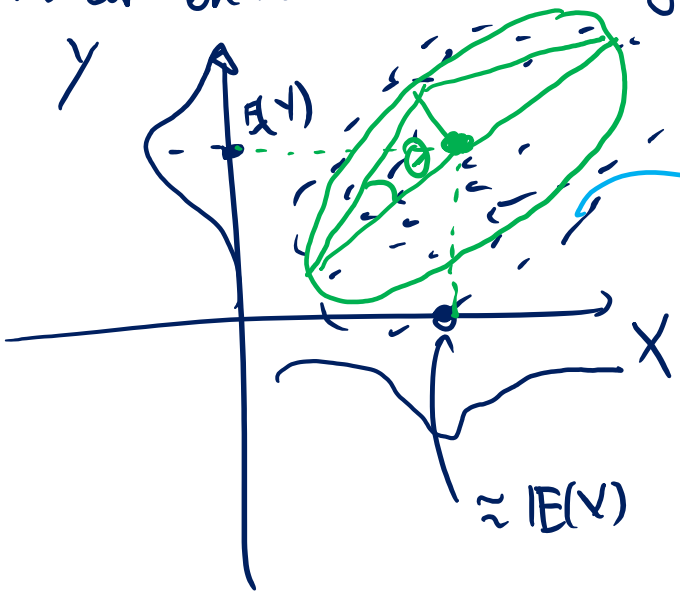
4. Este approach nos da una nueva visión de la covarianza y la varianza

$$Cov(X,Y) = \mathbb{E}\left((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right) = \langle X-\mathbb{E}(X), Y-\mathbb{E}(Y)\rangle$$

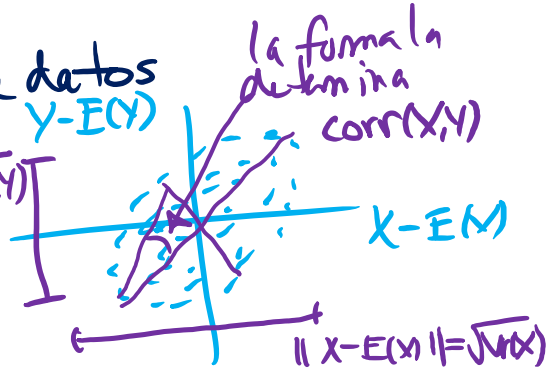
$$Var(X) = \mathbb{E}\left((X-\mathbb{E}(X))^2\right) = \|X-\mathbb{E}(X)\|^2$$

Podemos entonces pensar las covarianzas y varianzas en términos de "ángulos" y "normas"

Pensar en términos de diagramas de dispersión de datos



← cada punto es la coordenada (x_i, y_i) de un individuo



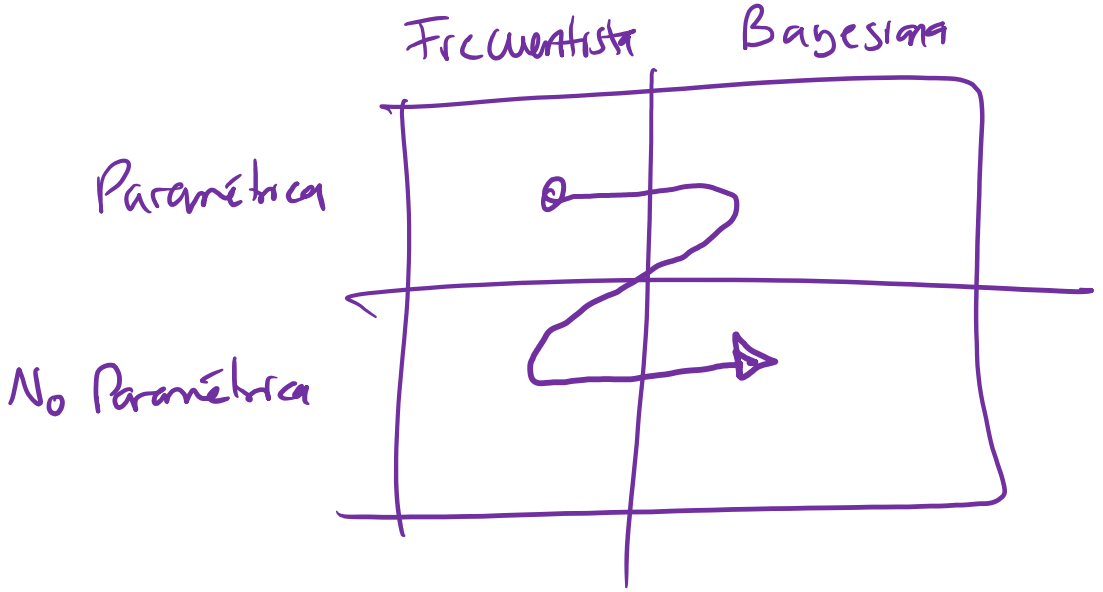
La elongación del diagrama depende de $Cov(X,Y)$

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\langle X-E(X), Y-E(Y)\rangle}{\|X-E(X)\|\|Y-E(Y)\|}$$

$$= \cos\left(\angle\left(X-E(X), Y-E(Y)\right)\right)$$

Continuara....

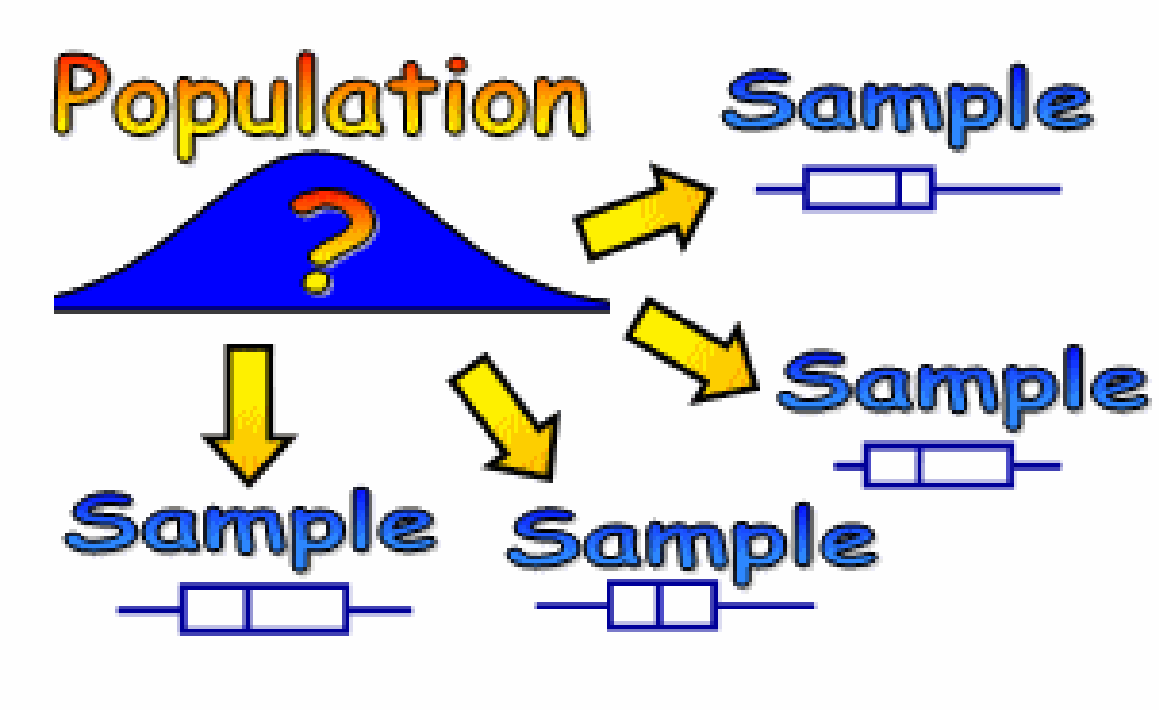
ESTADÍSTICA



Muestras

- Una **muestra** es una colección de variables aleatorias $X_1, \ldots, X_m$ usualmente independientes con una misma distribución F_X
- Usualmente en proba decimos que es una sucesión de v.a.i.i.d.
- Usar la palabra muestra o v.a.i.i.d. es usualmente un problema de *enfoque*.

- Vaiids vs Realizaciones
- X_i vs $X_i(\omega) = x_i$



Estadística Estimadores

- *Una* **estadística** es cualquier cantidad de interés acerca de una distribución F_X *(in fancy words: una funcional)*
- Dada una muestra $X_1, \ldots, X_n$, un **estimador** es una función $h(X_1, \ldots, X_n)$.
- Un estimador es una función que depende de los valores de la muestra y que su propósito es tratar de conocer el valor de alguna estadística.

Ejemplos....

- *Teaser*:¿ Cómo estimarían la esperanza de una distribución data una muestra?

Queremos $\int_{IR} x\, dF_X(x) = \int_{\alpha} x f_X(x) dx = E(X_i) \triangleq \eta(F_X)$

función de la distrib

de una muestra $X_1, \ldots, X_n$

$\hat{\eta}(X_1, \ldots, X_n) = \frac{1}{n} \sum X_i$

La estadística es la media

$\hat{\eta}$ estima η

estimador — estadística

¿ cómo saber si un estimador es bueno/malo?

Estadística Paramétrica

- Tenemos una muestra $\mathrm{X} = \{X_1, \ldots, X_n\}$ con distribución común F_X^{θ} y queremos aprender el "valor" de θ. Queremos "decidir" que valores puede tomar θ.
- Ej:
- $F_X^{\theta} \sim N(\mu, \sigma^2) \quad \theta = (\mu, \sigma^2)$
- $F_X^{\theta} \sim Ber(p) \quad \theta = (p)$

En la vida real nosotros acotamos a una familia paramétrica específica dependiendo del tipo de datos. La elección de familia paramétrica se llama el **modelo.**

Inferencia Bayesiana vs Inferencia Frecuentista

Inferencia Frecuentista	**Inferencia Bayesiana**
Los parámetros θ que queremos conocer son tratados como NÚMEROS	Los parámetros θ que queremos conocer son tratados como VARIABLES ALEATORIAS para modelar nuestra incertidumbre.
NO puedo aportar conocimiento previo.	Puedo aportar conocimiento previo o "a priori" que influye en el resultado final.
Dada una muestra, me devuelve un NÚMERO que un estimador.	Dada una muestra me devuelve una nueva distribución para θ llamada distribución "a posteriori" que combina la "a priori" con los datos.
Máximo verosimilitud es un *ejemplo* de estimación frecuentista.	Hay un método de inferencia: **el método de Bayes** y combina la *versomilitud* con la información a priori.

(*just to be clear*... El método de *máxima verosimilitud* es un método 100% frecuentista, pero la verisimilitud tiene un rol importante en la inferencia bayesiana…)

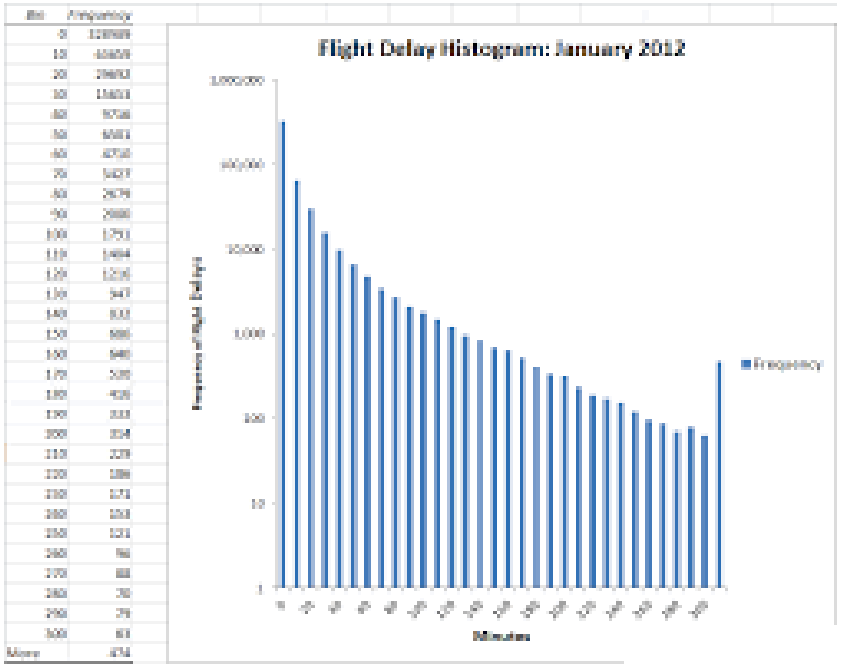
Máxima verosimilitud

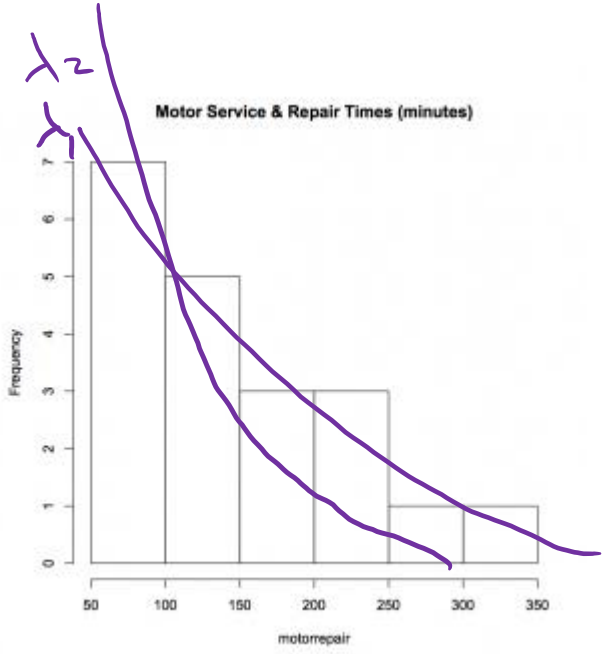
- Si $\mathrm{F_X} \sim Ber(p)$ y $X_1, \ldots, X_n$ es una muestra cuya realización fue:

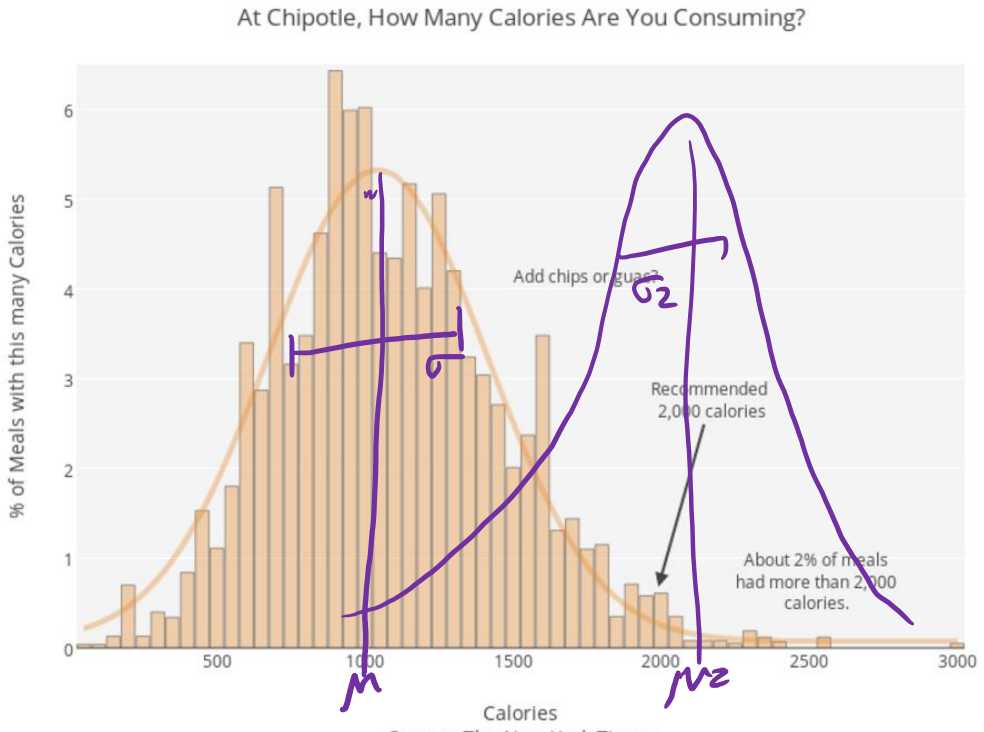
  1,0,1,1,1,1,0,1,0,1,0,1,1,1,0,1,1,1,1,1

- ¿Qué p es más **creíble**? $p = .3$ o $p = .75$? Y maximizarla…

Nuestro objetivo es buscar una "medida" de qué tan creíble es cada p y buscar la p que maximice esta credibilidad: en esencia eso es máxima verosimilitud.

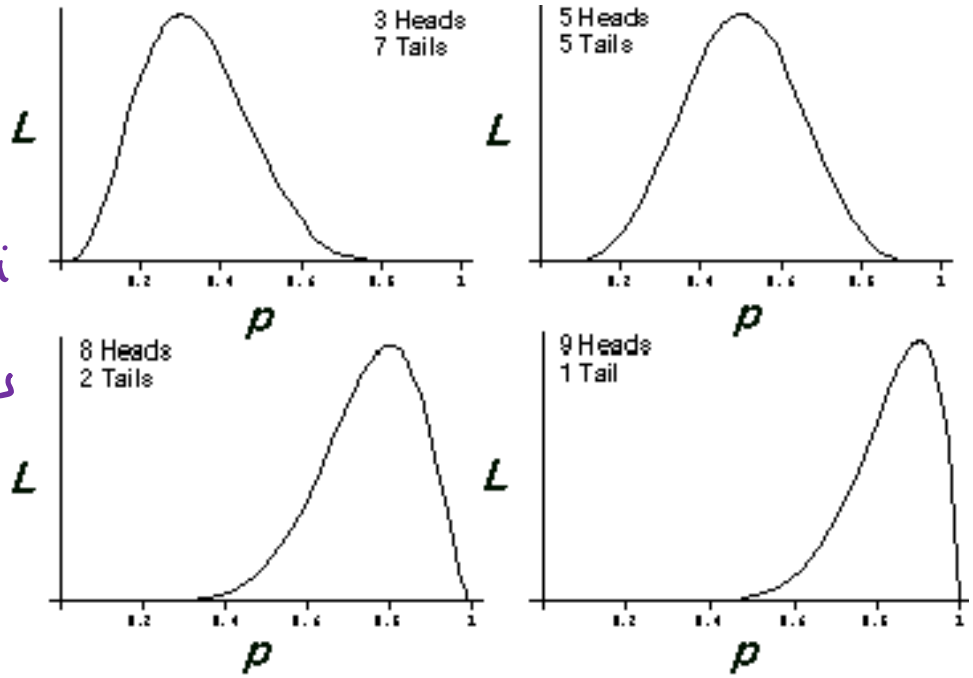

Flight Delay Histogram: January 2012
Frequency
Minutes


Motor Service & Repair Times (minutes)
Frequency
motorrepair


At Chipotle, How Many Calories Are You Consuming?
% of Meals with this many Calories
Add chips or guac?
Recommended 2,000 calories
About 2% of meals had more than 2,000 calories.
Calories

- Existe una elección de función de "credibilidad" que es muy natural:
- Si en el ejemplo anterior graficamos
- $P(X_1 = x_1, \ldots, X_n = x_n; p)$ para distintos valores de p tenemos obtenemos algo que se ve así:

$$L(p) \equiv \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\#1's}(1-p)^{\#0's}$$



- La función de **verosimilitud se define como:**

$$L(\theta|\mathrm{X}) = P(X_1 = x_1, \ldots, X_n = x_n; \theta)$$
$$= \prod_i P(X_i = x_i; \theta)$$

Si la distribución es discreta solo hay que reemplazar la probabilidad por la función de densidad:

$$L(\theta|\mathrm{X}) = f_X(X_1 = x_1, \ldots, X_n = x_n; \theta)$$
$$= \prod_i f_X(X_i = x_i; \theta)$$

Método de Máxima verosimilitud
Inferencia clásica: Dada una muestra $X_1, \ldots, X_n$ - $\hat{\theta}_{MLE}(X_1, \ldots, X_n) = \arg\max_\theta L(\theta|X)$

Inferencia Bayesiana

Teorema de Bayes:

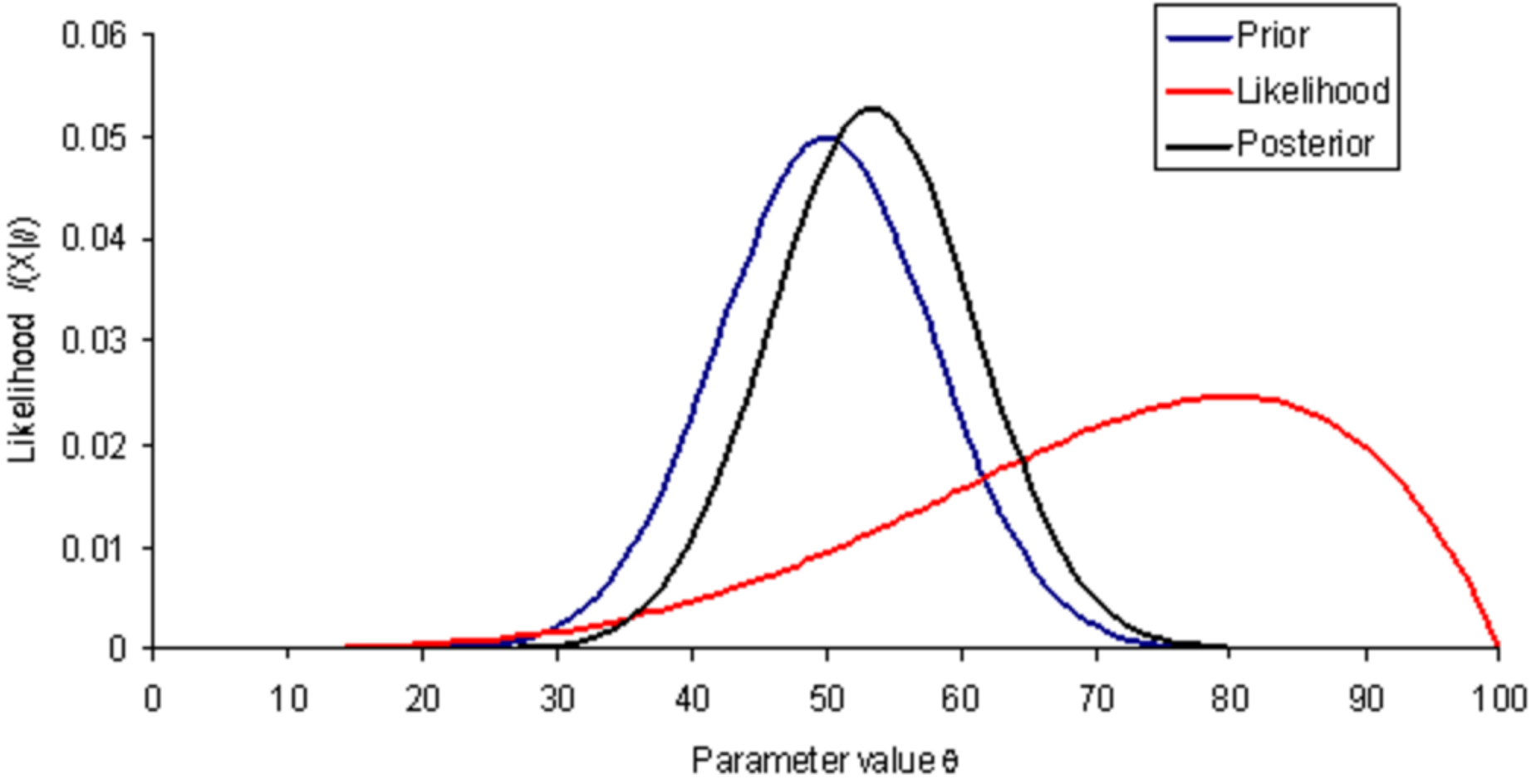
$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)} \propto P(B|A)P(A)$$

Llevémoslo a variables aleatorias y pongan θ en vez de A y su muestra $\mathrm{X} = \{X_1, \ldots, X_n\}$ en vez de B y obtienen

$$f_{\theta|X}(\theta|\mathrm{X}) = \frac{L(\theta \mid \mathrm{X}) f(\theta)}{\int L(\tau \mid \mathrm{X}) f(\tau) d\tau}$$
$$\propto L(\theta \mid \mathrm{X}) f(\theta)$$

en el caso continuo… pongan sumas y probabilidades en vez de integrales y densidades en el caso discreto y ya…

"están mezclando la a priori con la verosimilitud"


Prior
Likelihood
Posterior
0.06
0.05
0.04
0.03
0.02
0.01
0
Likelihood f(X|θ)
0
10
20
30
40
50
60
70
80
90
100
Parameter value θ