Enfoque Estadístico del Aprendizaje y el Descubrimiento 2019

Diego Kozlowski – diegokoz92@gmail.com
Juan Barriola – jmbarriola@gmail.com
María Eugenia Szretter – meszre@dm.uba.a
Andrés Farall – afarall@hotmail.com

Programa de la Materia

Maru

- Regresión lineal simple y múltiple.
  - Estimación por Cuadrados Mínimos.
  - Multicolinealidad.
  - Transformaciones.
  - Variables dummy. Interacción
  - Métodos de ajuste paso a pas
  - Alternativas Robustas
- Modelos Lineales Generalizados (GL
- Regresión logística.
- Regresión de Cuantiles (QR)
- Introducción a los Modelos Lineales Mixtos (LMM)

Andrés

- Enfoques de la inferencia estadística
  - Modelado Estadístico
  - Significatividad Estadística (p-valor)
  - Máxima Verosimilitud e Inferencia Bayesiana.
- El problema de predicción.
  - Aprendizaje Supervisado
  - Medidas de Bondad de Ajuste
  - Trade-off sesgo-varianza
  - Sobreajuste

Andrés

- Regresión Lasso y Ridge
- Regresión No Paramétrica
  - Técnicas de suavizado
  - Modelos Aditivos
  - Projection Pursuit Regression
- Redes Neuronales Artificiales Multicapa (ANN – MLP). Regresión, Clasificación y Reducción de Dimensión (Autoencoders).
- Regresión/Clasificación con SVM y Gradient Boosting (Xgboost).
- Benchmarking, Comparación y selección de modelos: AIC, BIC, Enfoque Multimodel, Model Tuning(Caret).

Objetivos Principales del Curso

- Ofrecer un enfoque Estadístico de las técnicas de Regresión
- Brindar herramientas aplicadas
- Posicionarse en un contexto científico e interdisciplinario
- Enseñar una amplia variedad de técnicas implementadas en R
- Utilizar conjuntos de datos reales
- No profundizar en la matemática sobre la cual se basan los métodos

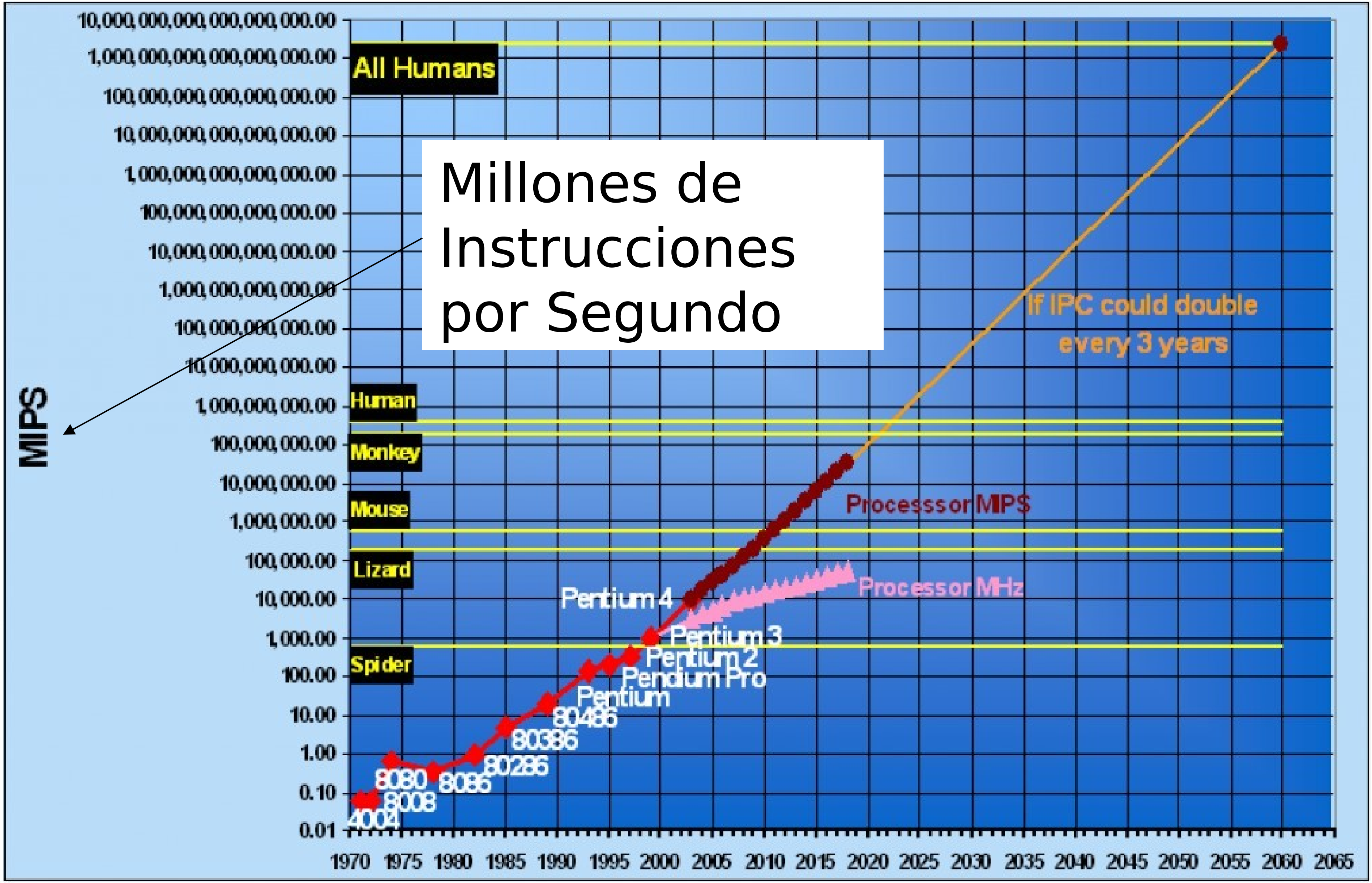
“Un Modelo Lineal No Se Le Niega a Nadie”

- Avanzamos de lo simple a lo complejo.
- Un modelo simple sirve como “benchmark” contra el que comparar el resultado de modelos más complejos.
- Un modelo simple permite interpretar la mecánica de las relaciones entre variables.

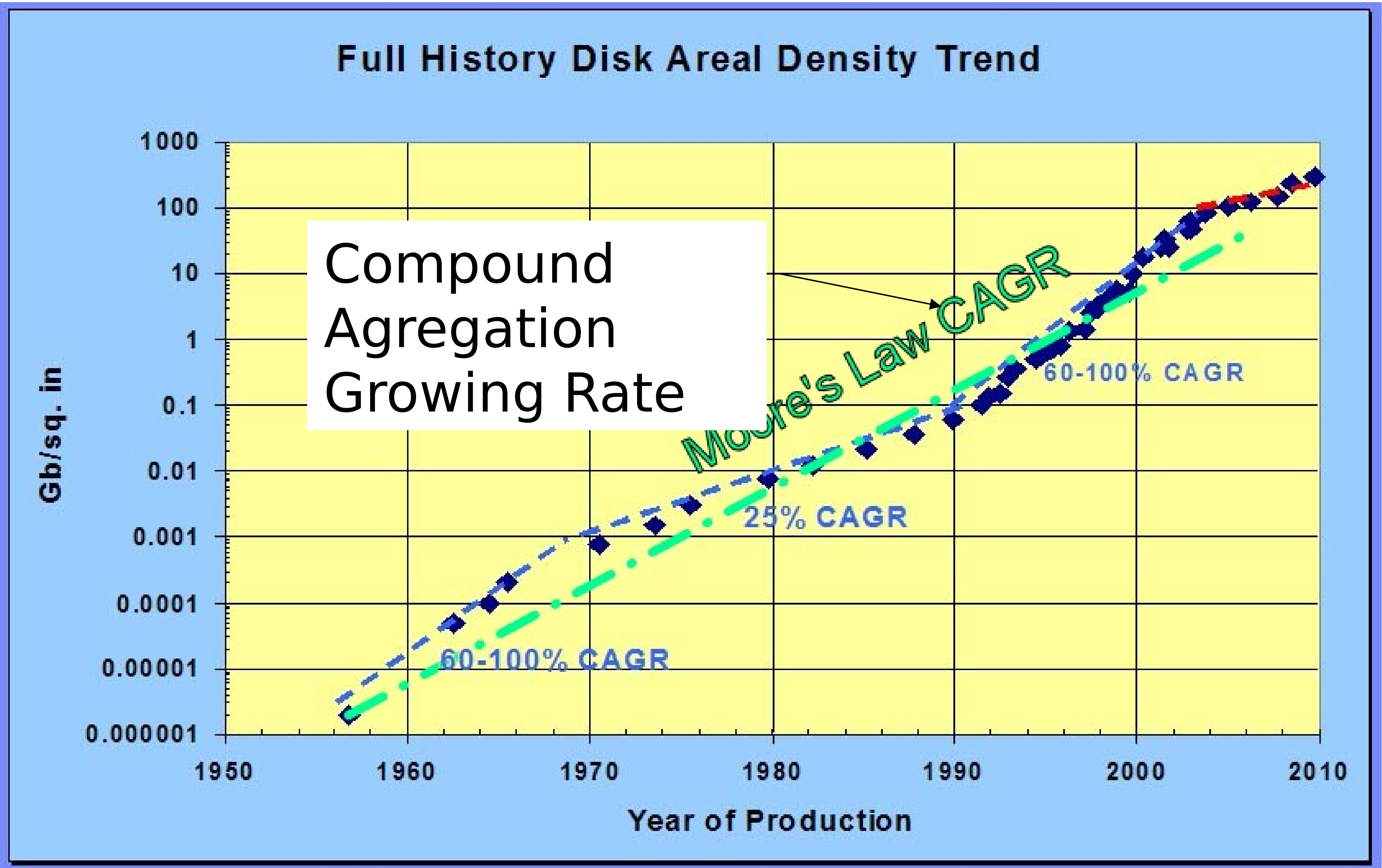
El Contexto Tecnológico

- Capacidad de Cálculo
- Capacidad de Almacenamiento
- Velocidad en la Transmisión de Datos
- Ciencia de Datos
- Machine Learning
- Data Mining
- Big Data
- Optimización

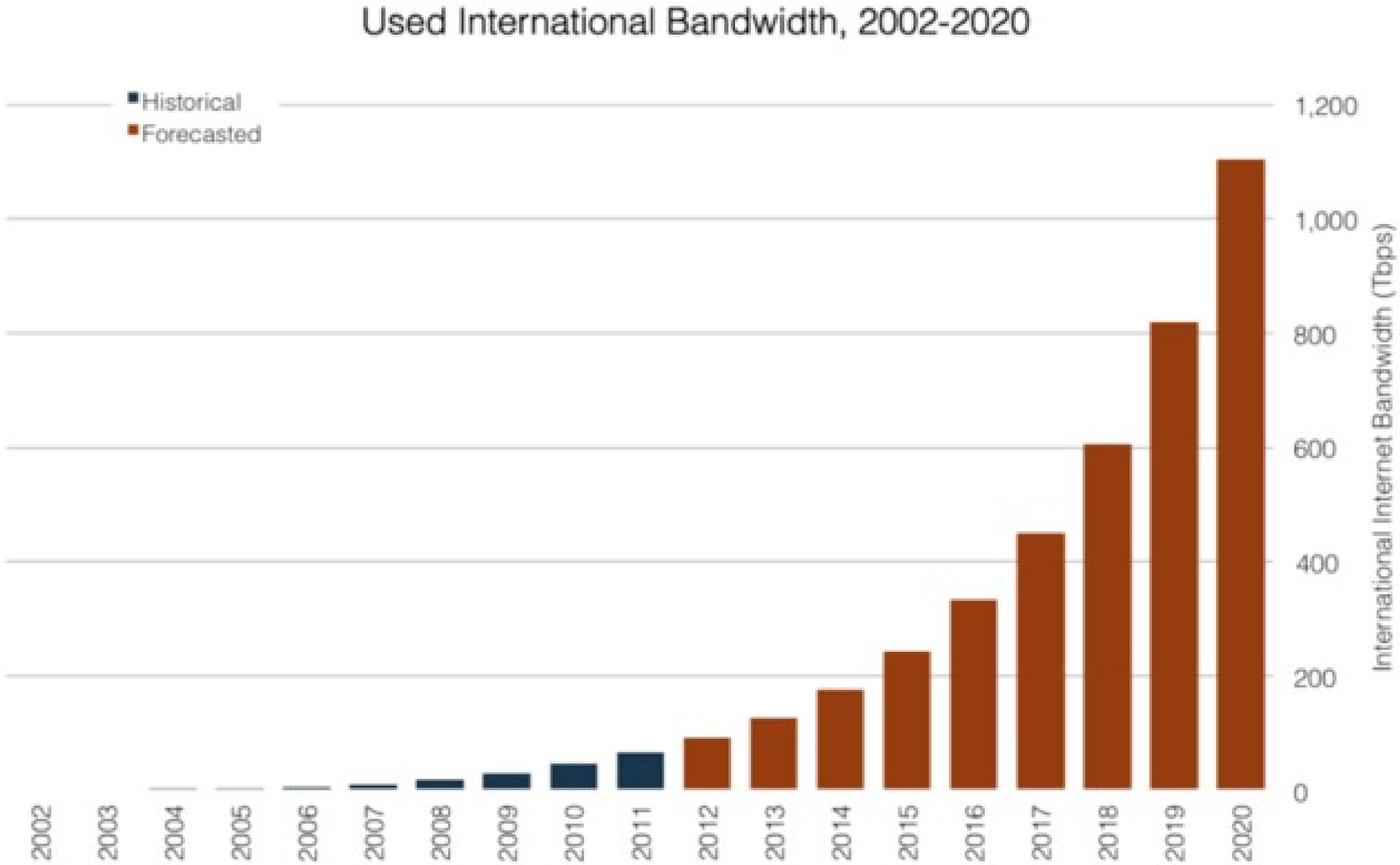
Que pasó con la Capacidad de Cálculo ?

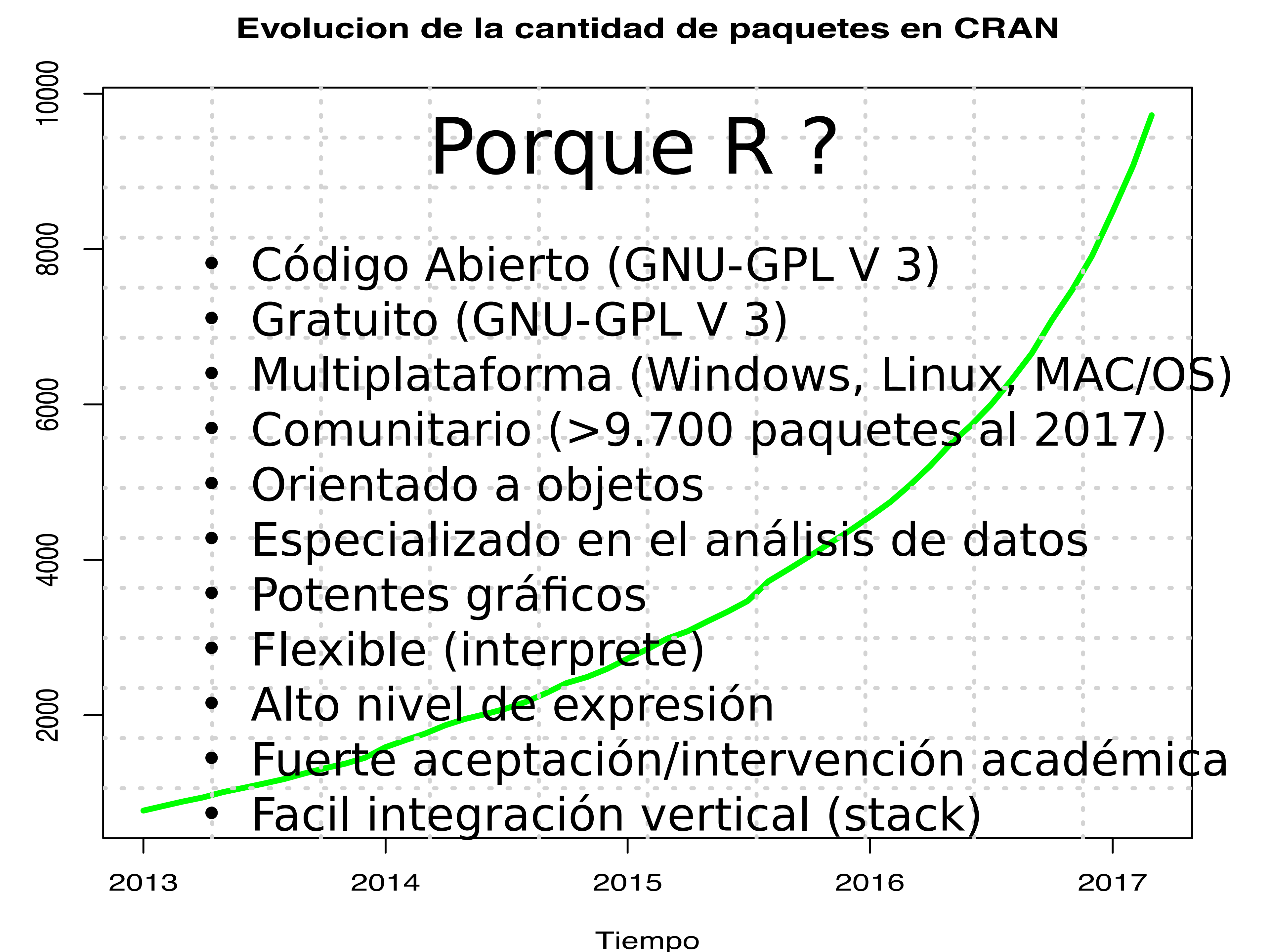


Que pasó con la Capacidad de Almacenamiento ?



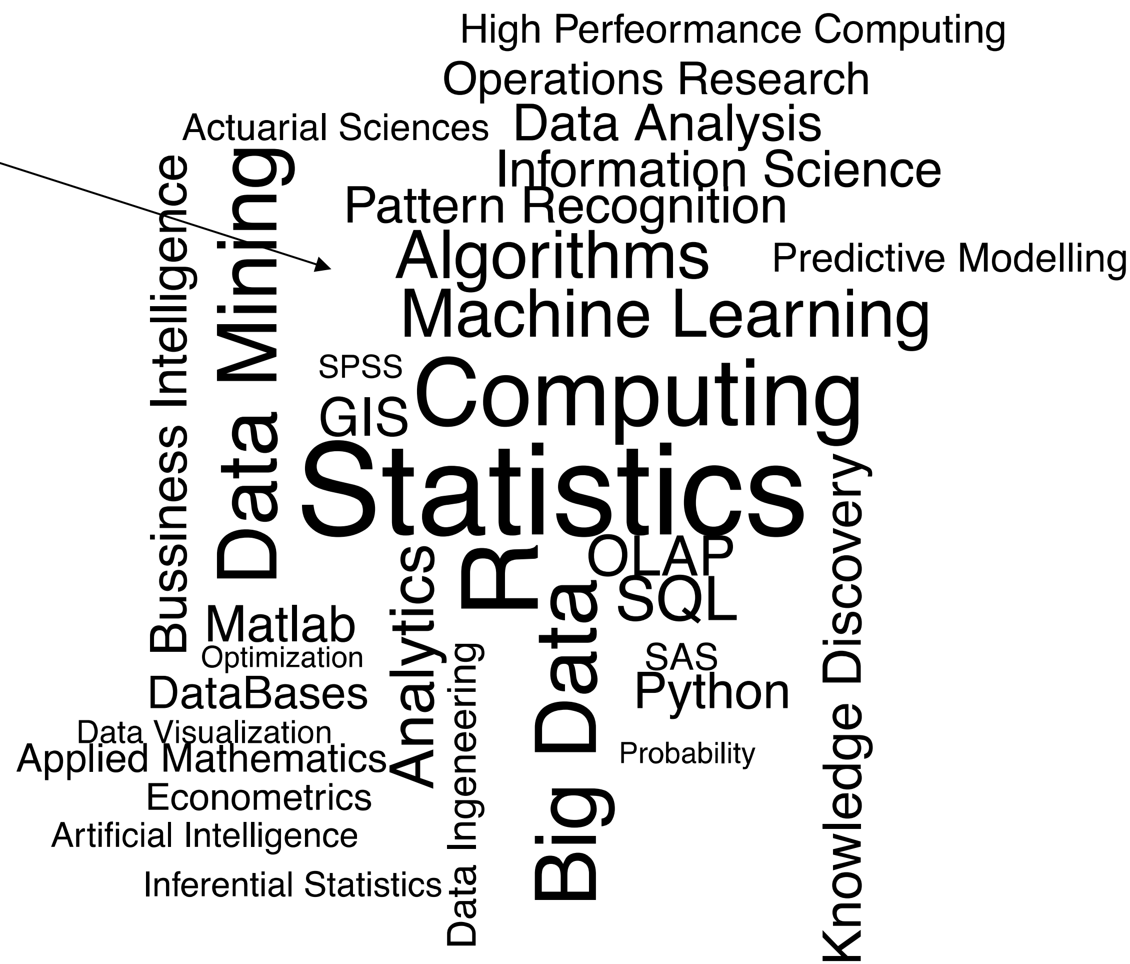
Que pasó con la Capacidad de Transmisión de Datos (Ancho de Banda) ?




Evolucion de la cantidad de paquetes en CRAN
Porque R ?
• Código Abierto (GNU-GPL V 3)
• Gratuito (GNU-GPL V 3)
• Multiplataforma (Windows, Linux, MAC/OS)
• Comunitario (>9.700 paquetes al 2017)
• Orientado a objetos
• Especializado en el análisis de datos
• Potentes gráficos
• Flexible (interprete)
• Alto nivel de expresión
• Fuerte aceptación/intervención académica
• Facil integración vertical (stack)
10000
8000
6000
4000
2000
2013
2014
2015
2016
2017
Tiempo

Que es Ciencia de Datos ?

WordCloud de los Componentes de la Ciencia de Datos



Estadística

- Basada en la Teoría de Probabilidades.
- Formalizó el concepto de incertidumbre en las mediciones/estimaciones.
- Condicionada por la escacez de datos (N > 30 ?)
- Herramientas/conceptos básicos utilizados:
  - Modelo probabilístico
  - Población / Muestra
  - Variable Aleatoria
  - Verosimilitud
  - Inferencia
  - Significancia / P-valor
  - Intervalos de Confianza
  - Test de Hipótesis
  - Interpretabilidad

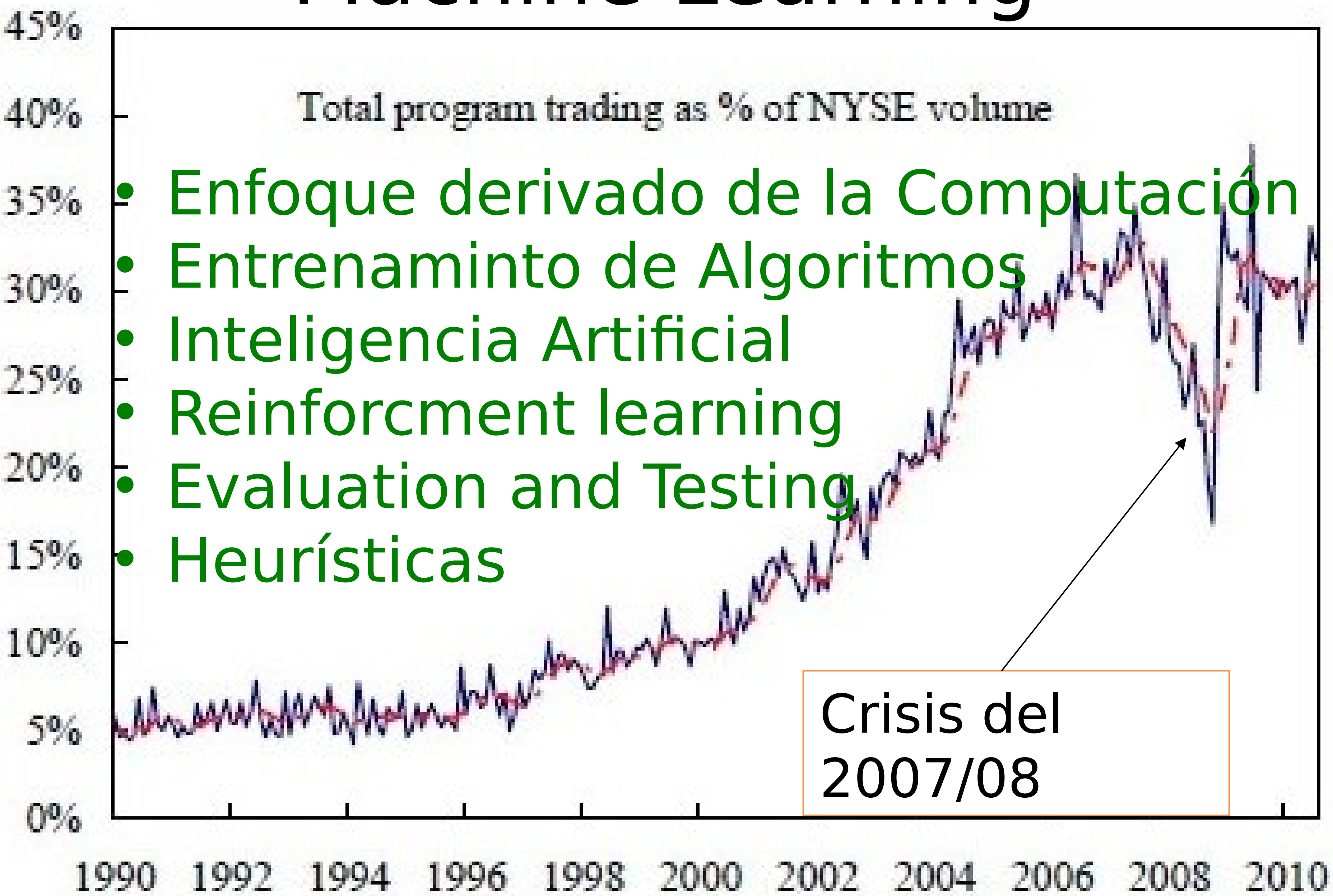
$$\mathbf{X} = \boldsymbol{\mu} + \epsilon$$

Componente Determinístico

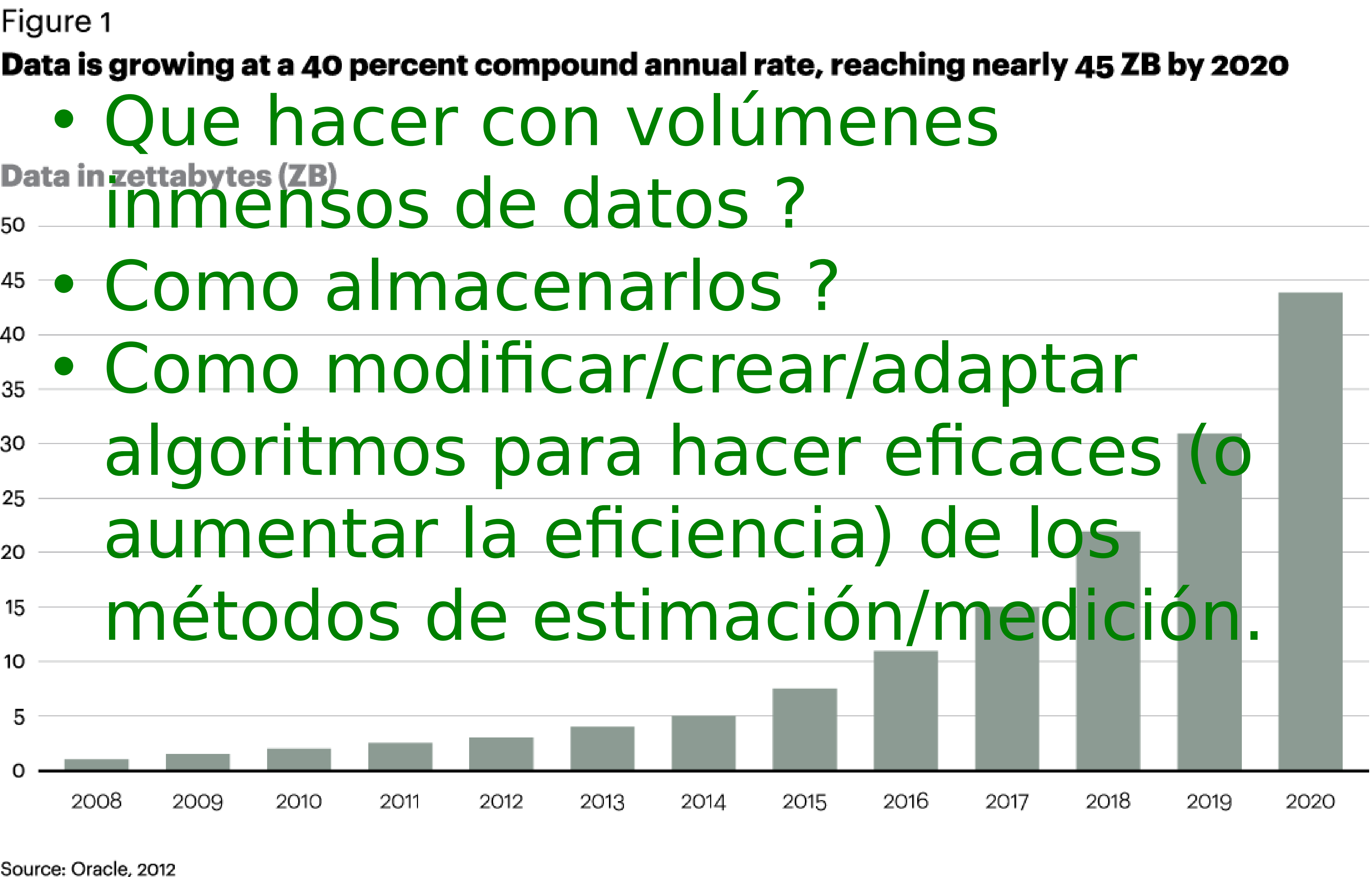
Componente Aleatorio

Machine Learning

- Enfoque derivado de la Computación
- Entrenaminto de Algoritmos
- Inteligencia Artificial
- Reinforcment learning
- Evaluation and Testing
- Heurísticas



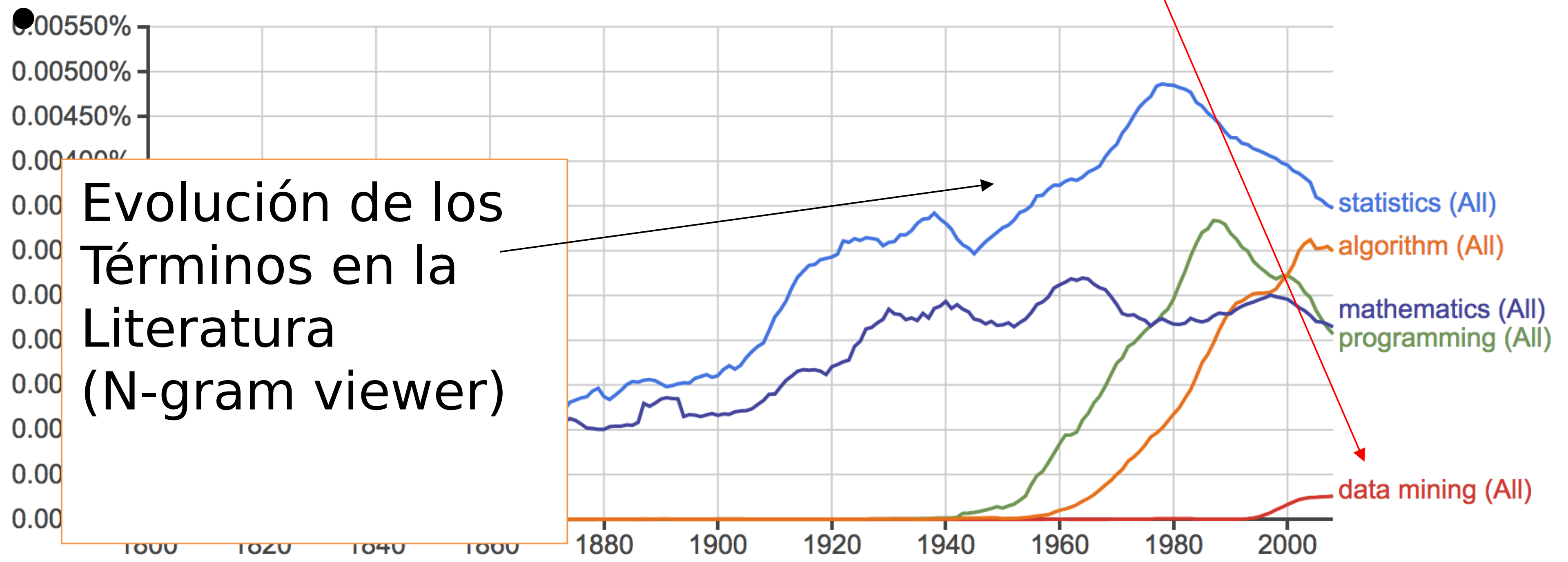
Big Data



- Que hacer con volúmenes inmensos de datos ?
- Como almacenarlos ?
- Como modificar/crear/adaptar algoritmos para hacer eficaces (o aumentar la eficiencia) de los métodos de estimación/medición.

Data Mining

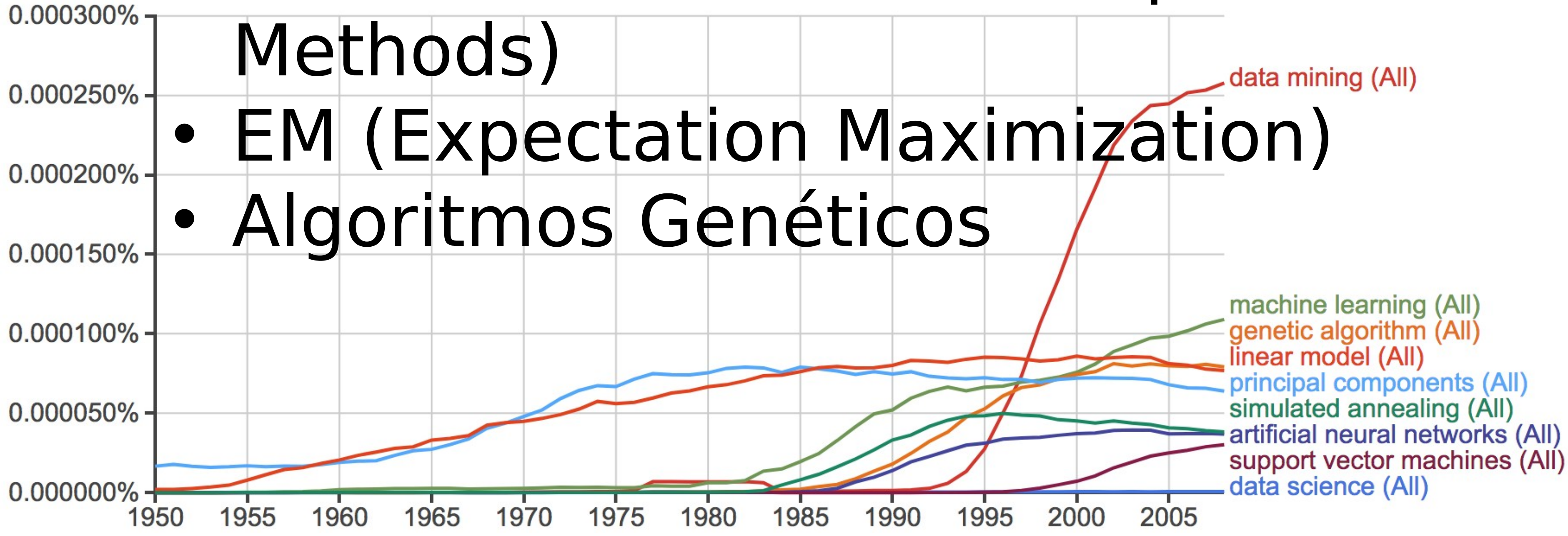
- Que patrones pueden extrarse de los datos ?
- Data Analysis y Analytics en gran escala
- Versión antigua del Data Science
-



Evolución de los Términos en la Literatura (N-gram viewer)

Optimización

- Fuerza Bruta
- Random Optimization (Luus-Jaakola)
- Gradient Descent
- Newton-Rapson (Quasi)
- Simulated Annealing
- Optimización Lineal/Cuadrática con/sin restricciones (Simplex Like Methods)
- EM (Expectation Maximization)
- Algoritmos Genéticos



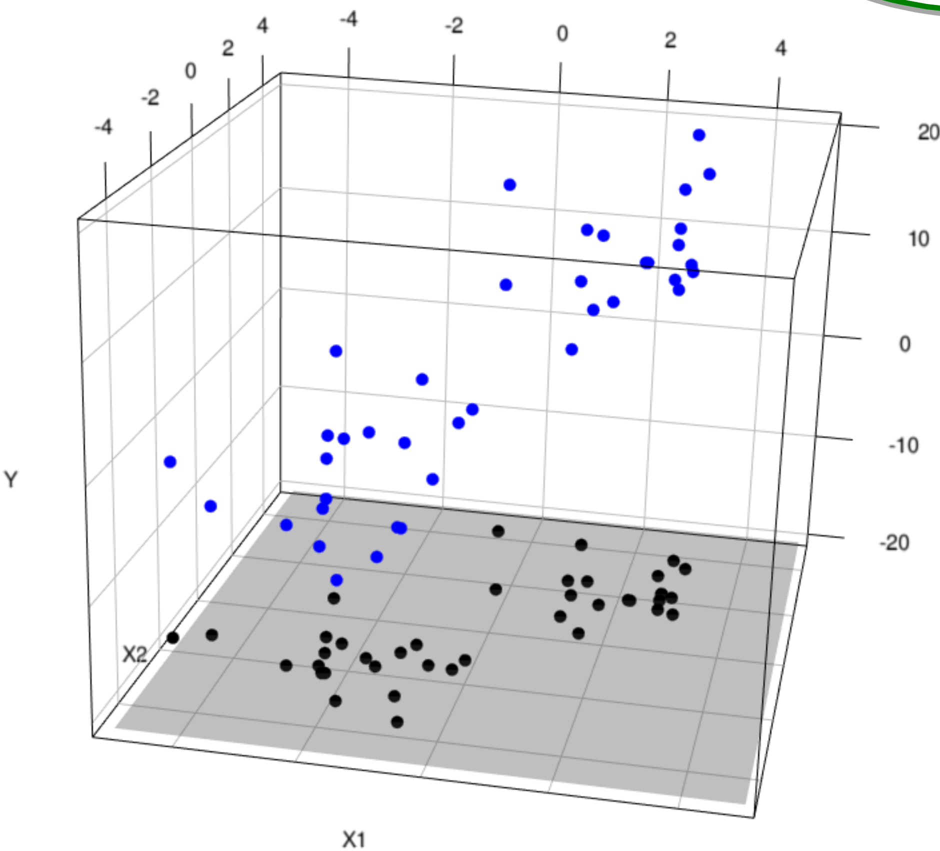
Taxonomía Basica de los Métodos en la Ciencia de Datos

- Métodos Supervisados
  - Clasificación
    - CART
    - Support Vrctor Machines
  - Regresión
    - Modelos Lineales
    - Redes Neuronales
- Métodos No Supervisados
  - Análisis Factorial
    - Componentes Principales
    - Análisis de Correspondencia
  - Segmentación
    - K-medias
    - Clusterización Jerarquica
    - GMM

Supervisado Vs. No Supervisado

Espacio de probabilidad conjunto

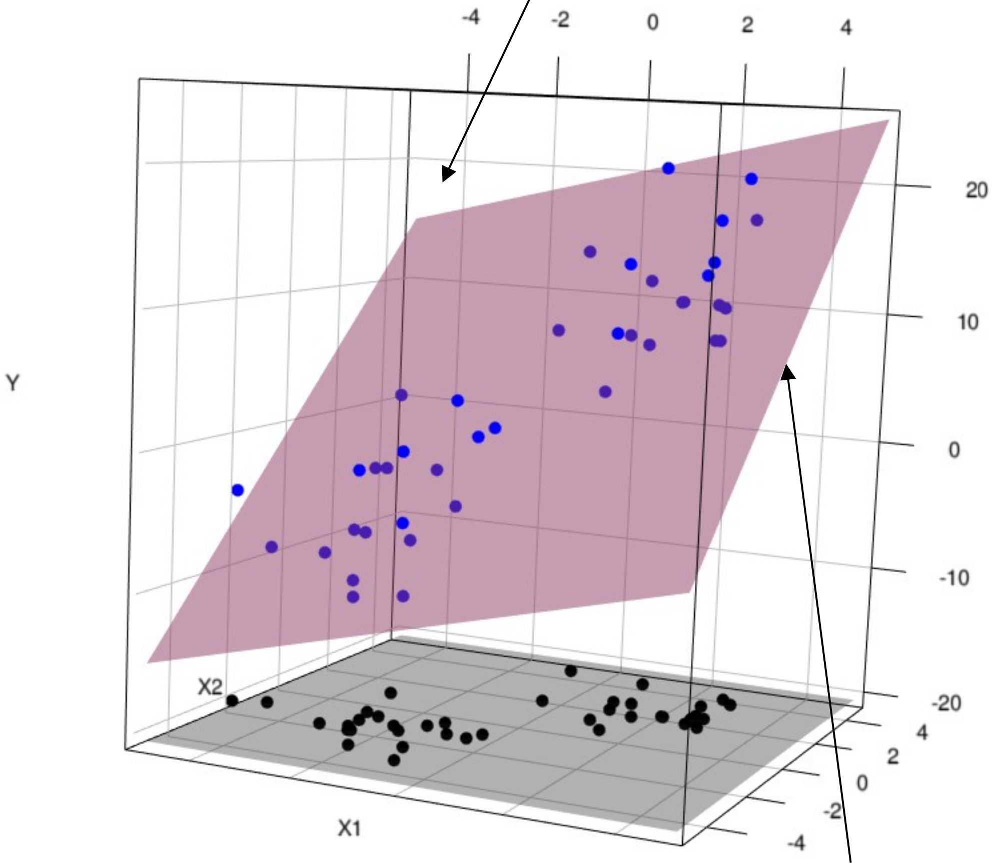
$$\Pr(X,Y) = \Pr(Y|X) \cdot \Pr(X)$$

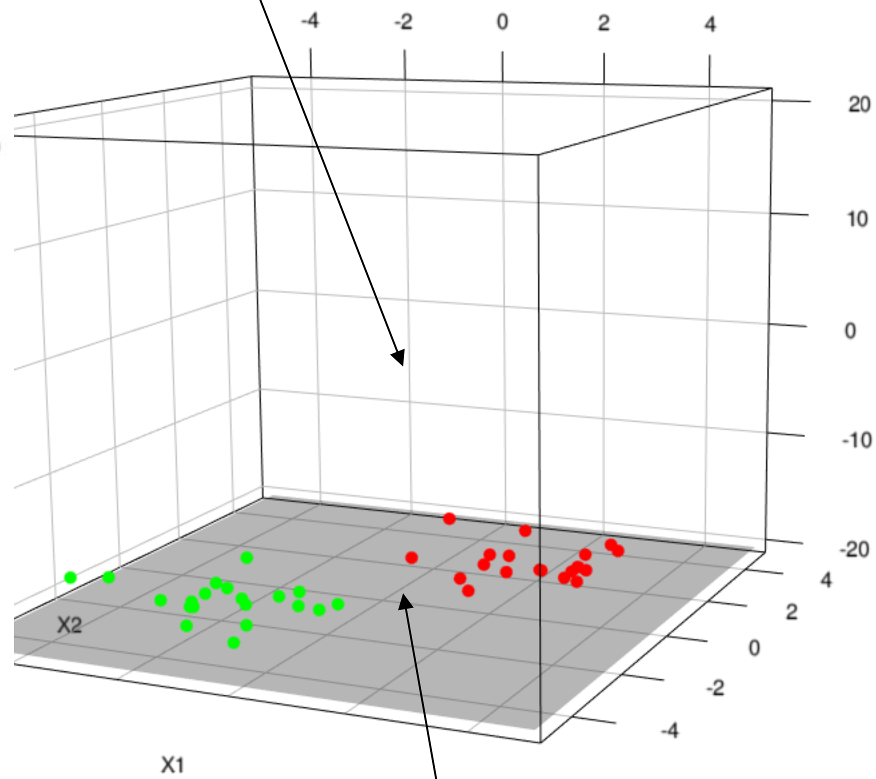


Variables Explicativas o Features

Variables Respuesta o Target

Regresión Vs. Clasificación





Relación entre Target y Features

Sólo Features

Sistema de Recomendaciones

Indiv.	Item 1	Item 2		Item j					Item p
1		3		9			2		
2	7	2			4				10
3			5						
4					8				3
5		1				7			
i			3	9					
N					7		4		

Indiv.	Item 1	Item 2		Item j					Item p
i*	?	6	?	?	?	8	?	3	?

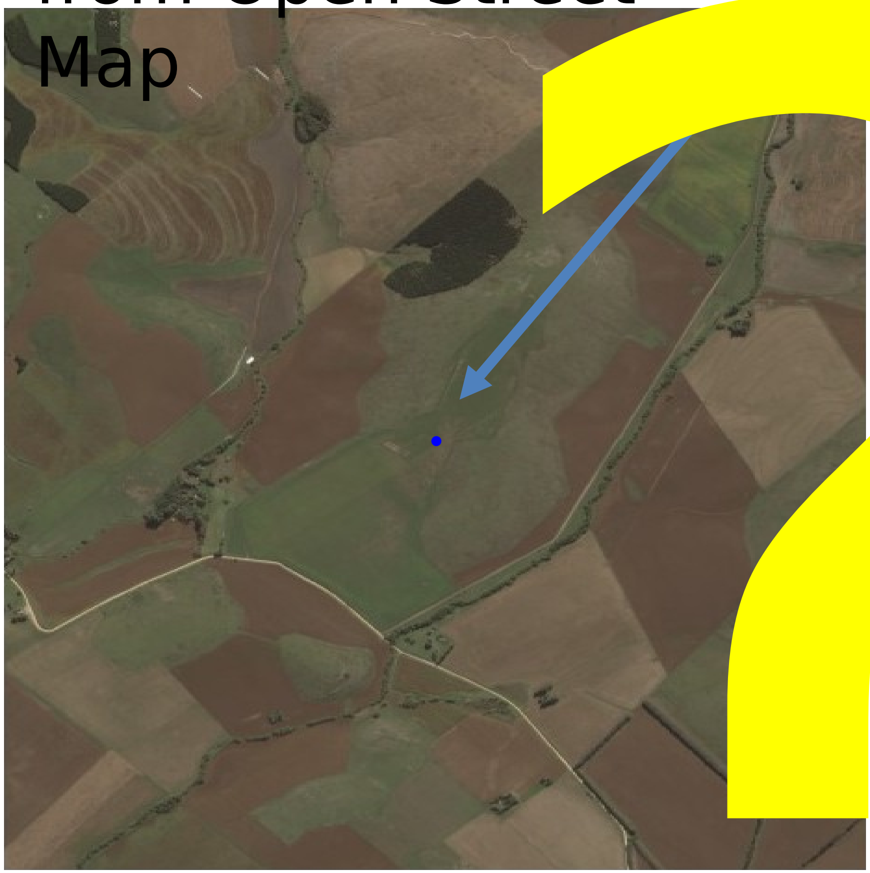
GeoReferenciación Automática

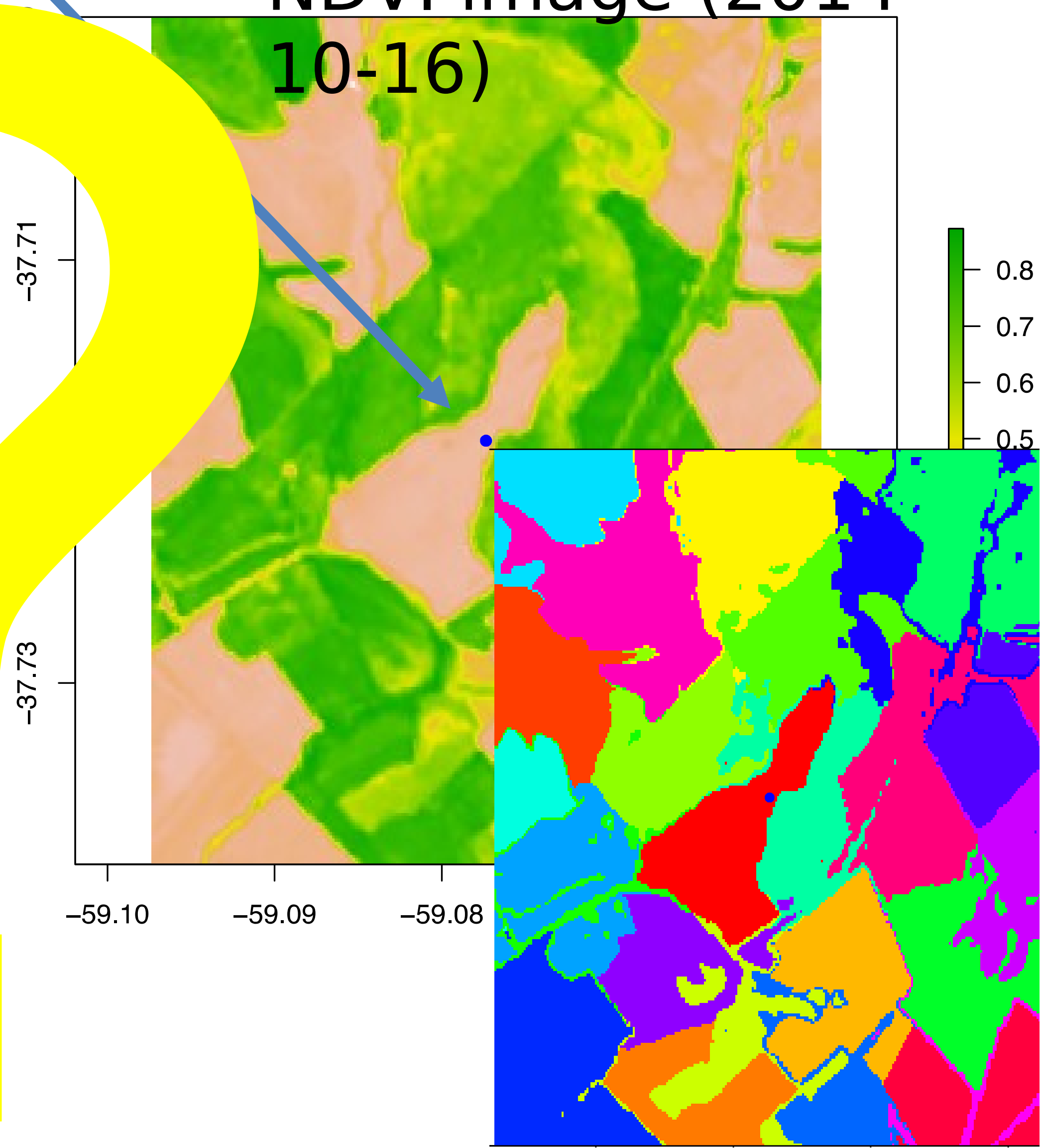
Visible image from Open Street Map

Point of interest

NDVI image (2014-10-16)

?





Symbolic Data Analysis
(Estadística de Objetos?)

Repaso

Basado en el curso de Ciencia de Datos con R Fundamentos Estadísticos

por

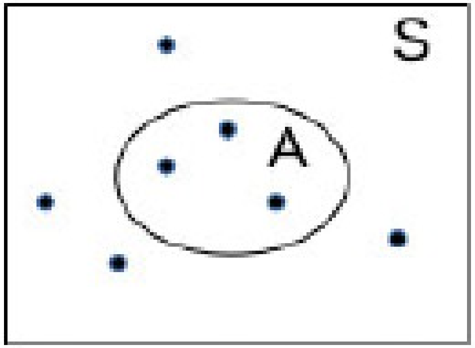
M. Sued, A. Bianco y M. Valdora

Teoría de Probabilidades

1. $\mathcal{S}$ espacio muestral: conjunto con los posibles resultados del experimento.

2. A, B, C: eventos a los cuales vamos a asignarles probabilidad.

3. $\mathbb{P}$: función de probabilidad.



$m = 7$
$m_A = 3$

$$\boxed{\frac{m_A}{m} \to \mathbb{P}(A)}$$

$\mathbb{P}(A)$ *representa el porcentaje de veces que esperamos que* A *ocurra en* **infinitas** *repeticiones*

Función de Probabilidad

$0 \leq \mathbb{P}(A) \leq 1$

La probabilidad del espacio muestral es uno: $\mathbb{P}(\mathcal{S}) = 1$.

Si A_1 y A_2 son eventos disjuntos($A_1 \cap A_2 = \emptyset$), entonces

$$\mathbb{P}(A_1 \cup A_2) \; = \; \mathbb{P}(A_1) + \mathbb{P}(A_2) \; .$$

$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

Dados dos eventos A y B, tenemos que

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Concepto de Independencia

Dado un evento B con $\mathbb{P}(B) > 0$, definimos la probabilidad del evento A dado que B aconteció mediante la formula:

Probabilidad Condicional

$$\boxed{\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}}$$

A y B se dicen independientes si

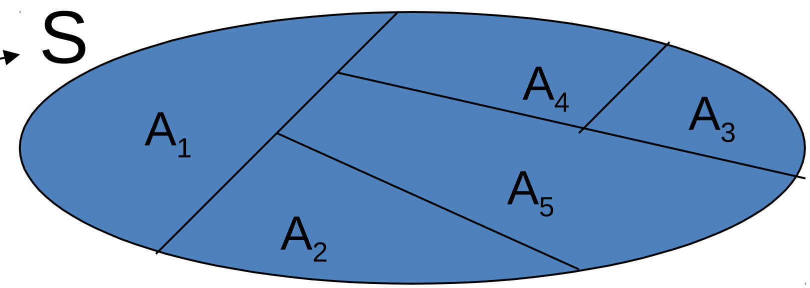
$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)}$$

Lema: Si A y B son independientes y $\mathbb{P}(B) > 0$, entonces

$$\mathbb{P}(A \mid B) = \mathbb{P}(A)$$

Teorema de Bayes

$(A_i)_{i\geq 1}$ partición de $\mathcal{S}$ si



1. Los eventos son disjuntos dos a dos: $A_i \cap A_j = \emptyset$, para $i \neq j$.

2. Los eventos cubren el espacio muestral: $\bigcup_{i\geq 1} A_i = \mathcal{S}$.

Inversión del Condicional

$$\mathbb{P}(C) = \sum_i \mathbb{P}(C \mid A_i)\mathbb{P}(A_i)\ .$$

$$\boxed{\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}{\sum_{j\geq 1}\mathbb{P}(B \mid A_j)\mathbb{P}(A_j)}\ \text{, para } i \geq 1.}$$

Variables Aleatoeias

Una Variable Aleatoria X es una función definida sobre el espacio muestral que toma valores en los reales:

$$X : \mathcal{S} \to \mathbb{R}$$

Conjunto de valores de la Variable Aleatoria

Evento de Interes

Elementos del Espacio Muestral

$$X \in A = X^{-1}(A) = \{s \in \mathcal{S} : X(s) \in A\}$$

$$\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}\left(\{\omega \in \mathcal{S} : X(\omega) \in A\}\right)$$

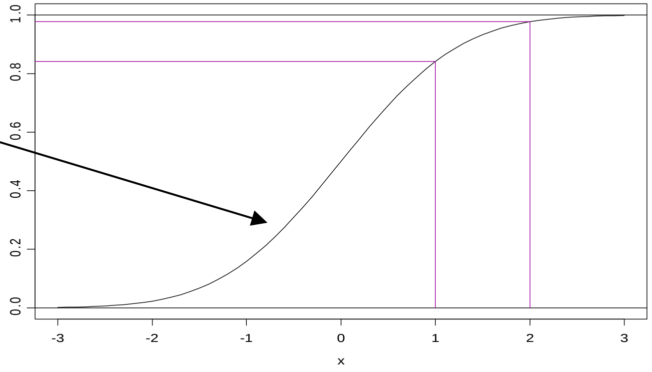
$$\mathbb{P}(X = t) = \mathbb{P}(X^{-1}(t)) = \mathbb{P}\left(\{\omega \in \mathcal{S} : X(\omega) = t\}\right)$$

Valor puntual

La Función de Distribución

La función de distribución acumulada F_X de la variable (aleatoria) X está definida por:

$F_X : \mathbb{R} \to [0,1] \qquad F_X(t) = \mathbb{P}(X \leq t) = F(t)$



1. $0 \leq F_X(t) \leq 1$ para todo $t \in \mathbb{R}$.

2. F_X creciente: si $s \leq t$, entonces $F_X(s) \leq F_X(t)$.

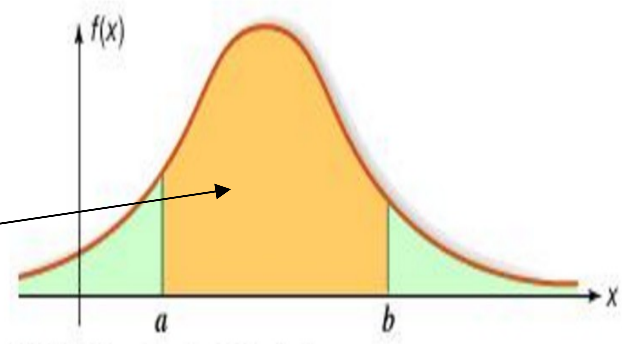
3. $\lim_{t\to-\infty} F_X(t) = 0$ y $\lim_{t\to+\infty} F_X(t) = 1$.
4. F_X es continua a derecha con límite a izquierda:

5. $\mathbb{P}(X = a)$ es el salto en a de F_X

La Función de Densidad

Una variable aleatoria X se dice ($^{(*)}$ absolutamente) continua sii existe una densidad

$$f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$$



tal que

$$\mathbb{P}(X \in A) = \int_A f_X(u)\, du.$$

En particular,

$$F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f_X(u)\, du\ .$$

En tal caso, diremos que f_X es la función de densidad de la variable aleatoria X.

$f : \mathbb{R} \to \mathbb{R}$ se dice densidad si

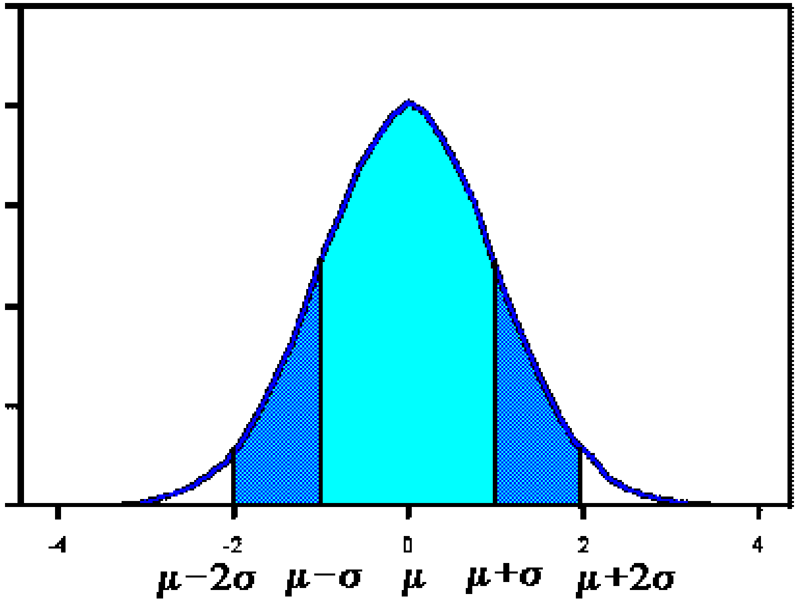
- $f(u) \geq 0$ para todo $u \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f(u)\, du = 1$

La Normal Univariada

- Z normal estandart si

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

- f_Z simétrica en el origen: $f_Z(z) = f_Z(-z)$
- Siendo f_Z simétrica, tenemos que $F_Z(-u) = 1 - F_Z(u)$
- $F_Z(z) = \int_{-\infty}^{z} f_Z(u)du$ no se puede calcular.
- Hay tabla con valores de $F_Z(u)$ para $u > 0$.
- $\phi(z) = F_Z(z)$ se llama función phi.



- Z normal estandar, Sea $X := \sigma Z + \mu$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

- $F_X(x) = \phi\left((x-\mu)/\sigma\right)$
- $X \sim \mathcal{N}(\mu, \sigma^2)$

Esperanza y Varianza

Dada una variable aleatoria discreta X con $\mathbb{R}g(X) = \{x_1, x_2, \cdots\}$ y función de probabilidad puntual $p_X(x_i)$, definimos la esperanza de X mediante la fórmula,

$$\mathbb{E}[X] \;=\; \sum_{i=i}^{\infty} x_i \, p_X(x_i) \;,$$

Dada una variable aleatoria continua X con función de densidad f_X, definimos la esperanza de X como

$$\mathbb{E}[X] \;=\; \int_{-\infty}^{\infty} u \; f_X(u) \, du \;.$$

X v.a.discreta con $\mathbb{E}[X] = \mu$. La varianza de X, está definida mediante la fórmula

$$\mathbb{V}(X) = \mathbb{E}\Big[(X - \mu)^2\Big] \;.$$

Covarianza y Correlación

- Medidas de asociación lineal entre variables (x e y)

Covarianza Empírica

$$cov\left(x,y\right)=\sum_{i=1}^{n}\frac{1}{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)$$

Correlación

Medias

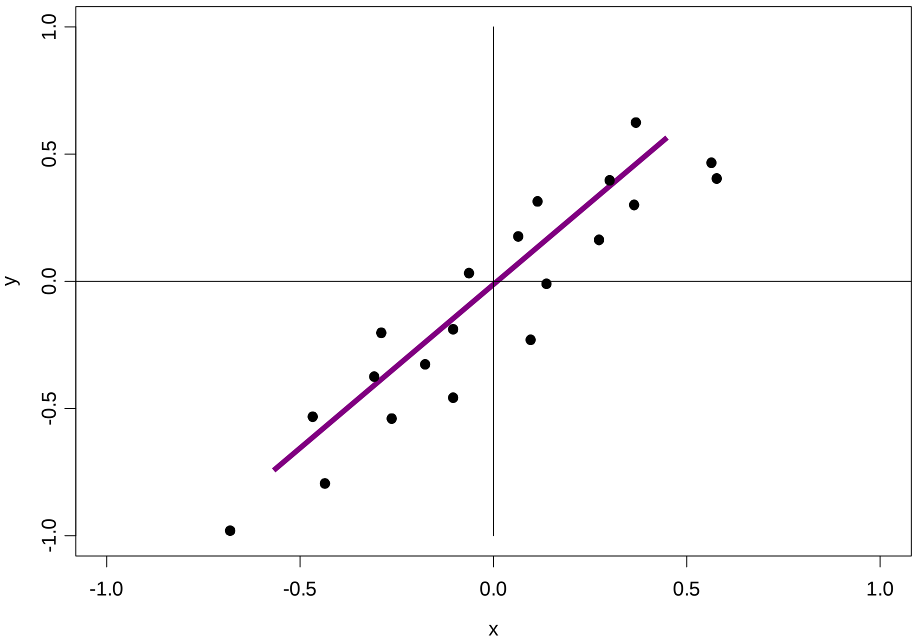
$$\rho\left(x,y\right)=cor\left(x,y\right)=\frac{cov(x,y)}{\sigma(x)\sigma(y)}$$

$$-1\leq\rho(x,y)\leq1$$

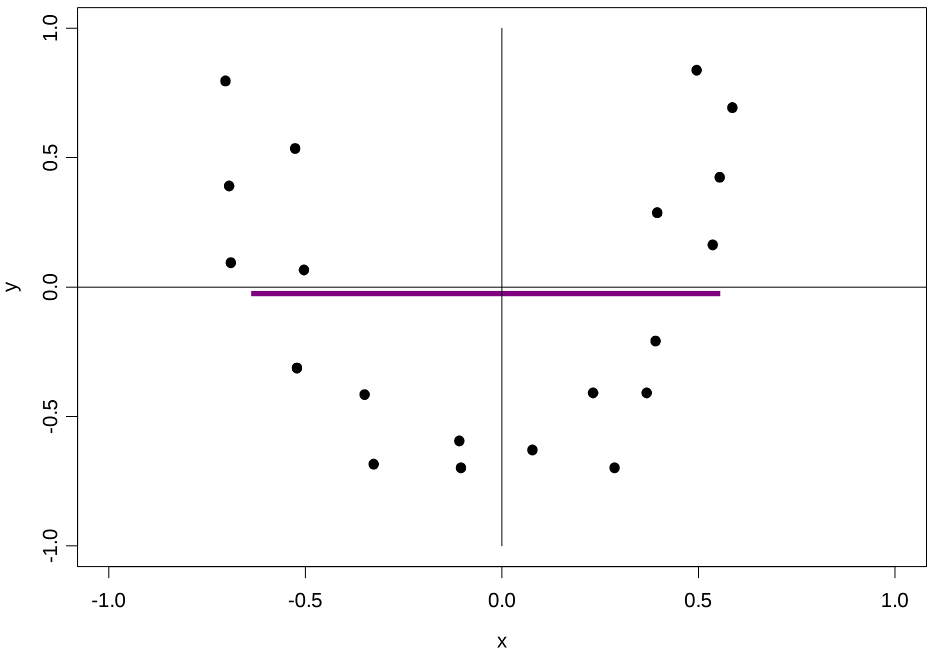
Desvios

Asociación Vs. Correlación

Cor(x,y) = 0.9



Cor(x,y) = 0.05



Tchevichev

Sea W una v.a. con media μ_W y $V[W] = \sigma_W^2$.Luego, $\epsilon > 0$

$$\boxed{\mathbb{P}\left(|W - \mu_W| \geq \varepsilon\right) \leq \frac{\sigma_W^2}{\varepsilon^2}}$$

Ley de los Grandes Números

Sean $(X_i)_{i\geq 1}$ i.i.d., con $E[X_i] = \mu$ y $V[X_i] = \sigma^2$, para todo i.
Entonces, el promedio converge a μ en probabilidad:

$$\lim_{n\to\infty} \mathbb{P}((|\bar{X}_n - \mu| > \varepsilon) = 0$$

Promedio

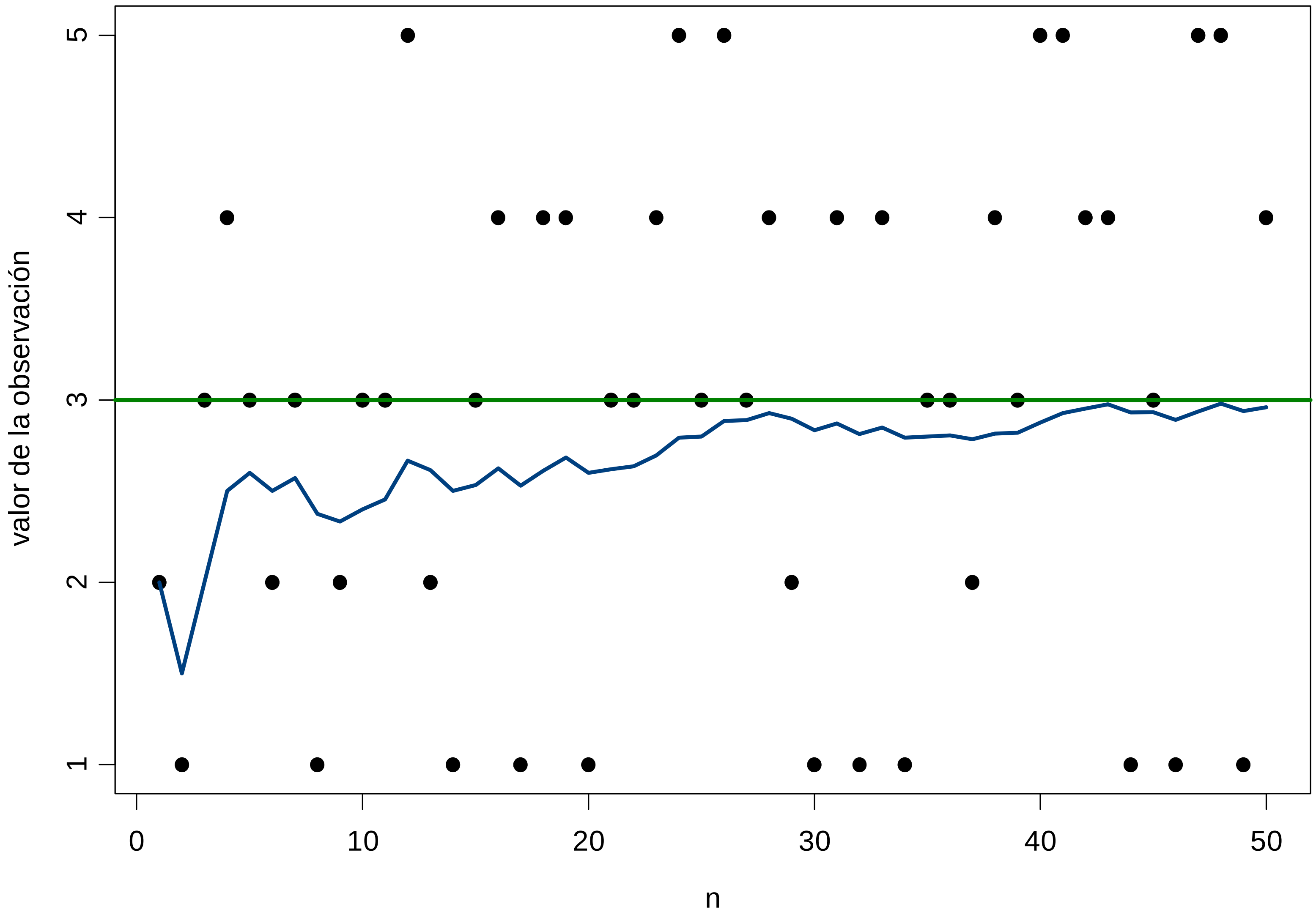
Es decir, para todo $\varepsilon > 0$, vale que

$$\boxed{\bar{X}_n \to \mu \text{ en probabilidad}}$$

Ejemplo de Ley de los Grandes Números

Valores de la Variable

Val	Prob
1	0.2
2	0.2
3	0.2
4	0.2
5	0.2



Teorema Central del Límite

Variables Independiantes e Identicamente Distribuidas

Sean $(X_i)_{i\geq 1}$ v.a. i.i.d. con $E[X]=\mu$ y $V[X]=\sigma^2$, entonces tenemos que

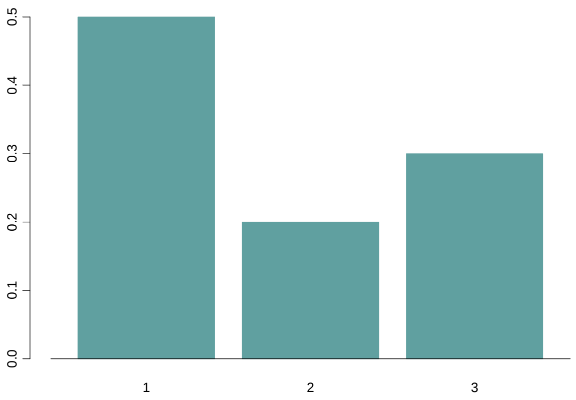
$$\mathbb{P}(\left(\frac{\bar{X}_n - E(X_1)}{\sqrt{Var(\bar{X}_n)}} \leq x\right) \underset{n\to\infty}{\to} \phi(x)\ \forall x \in \mathbb{R}\ ,$$

Promedio de Variables Aleatorias

Función de Distribución Normal

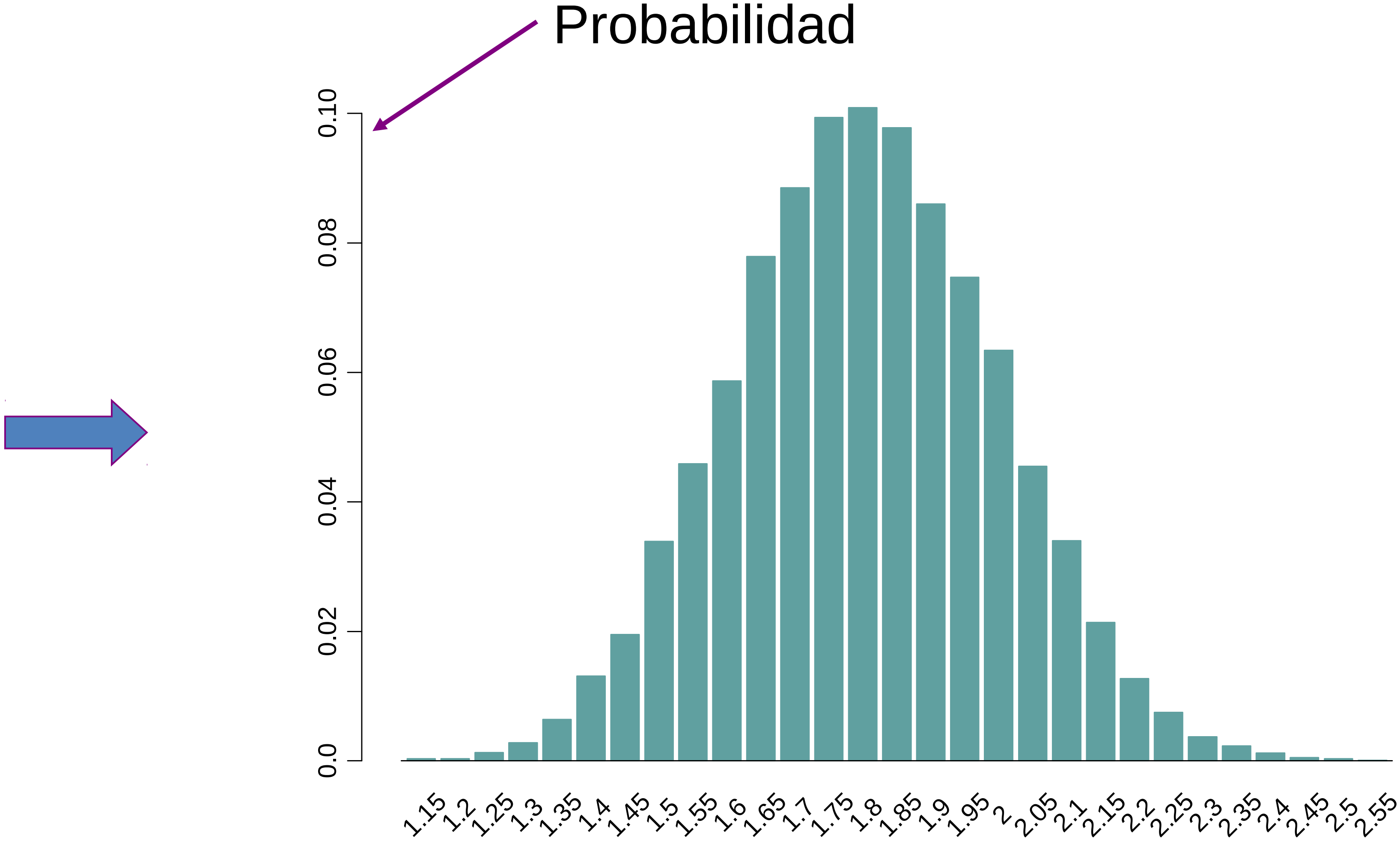
Ejemplo del Teorema Central

X_i	$P(x_i)$
1	0.5
2	0.2
3	0.3



$(X_1+x_2)/2$	$P((X_1+x_2)/2)$
2/2	0.5*0.5
3/2	0.5*0.2 + 0.2*0.5
4/2	0.5*0.3 + 0.2*0.2 + 0.3*0.5
5/2	0.2*0.3 + 0.3*0.2
6/2	0.3*0.3

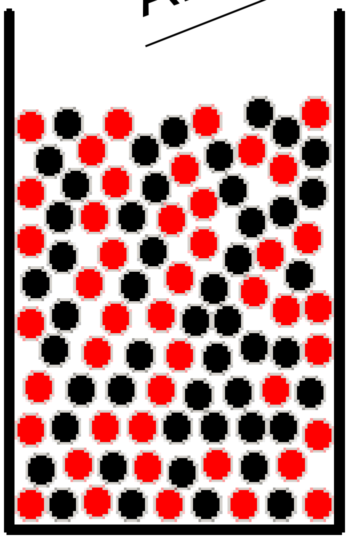
Distribución de $(x_1+x_2+\ldots+x_{20})/20$



Inferencia estadística

Trata de estimar o inferir mediante una muestra (aleatoria) el valor (desconocido) de un parámetro poblacional

Probabilidad e Inferencia, el problema de la “Inversión”



Muestra Aleatoria

$$X_1, \ldots, X_n$$

Estimador

$$\widehat{\theta}_n = h_n(X_1, \ldots, X_n)$$

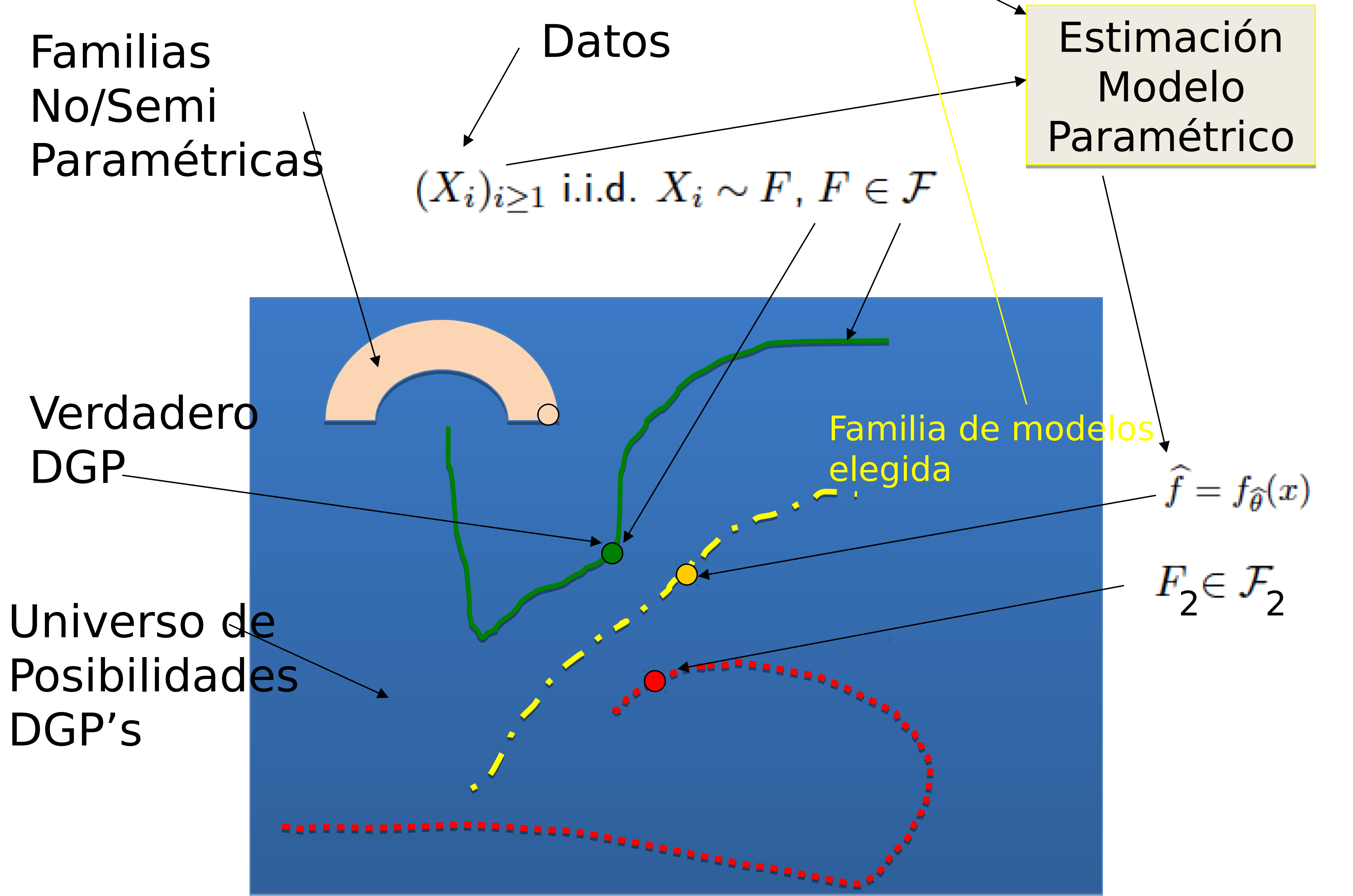
Estimación

$\theta \;=$ Proporción poblacional de bolitas rojas

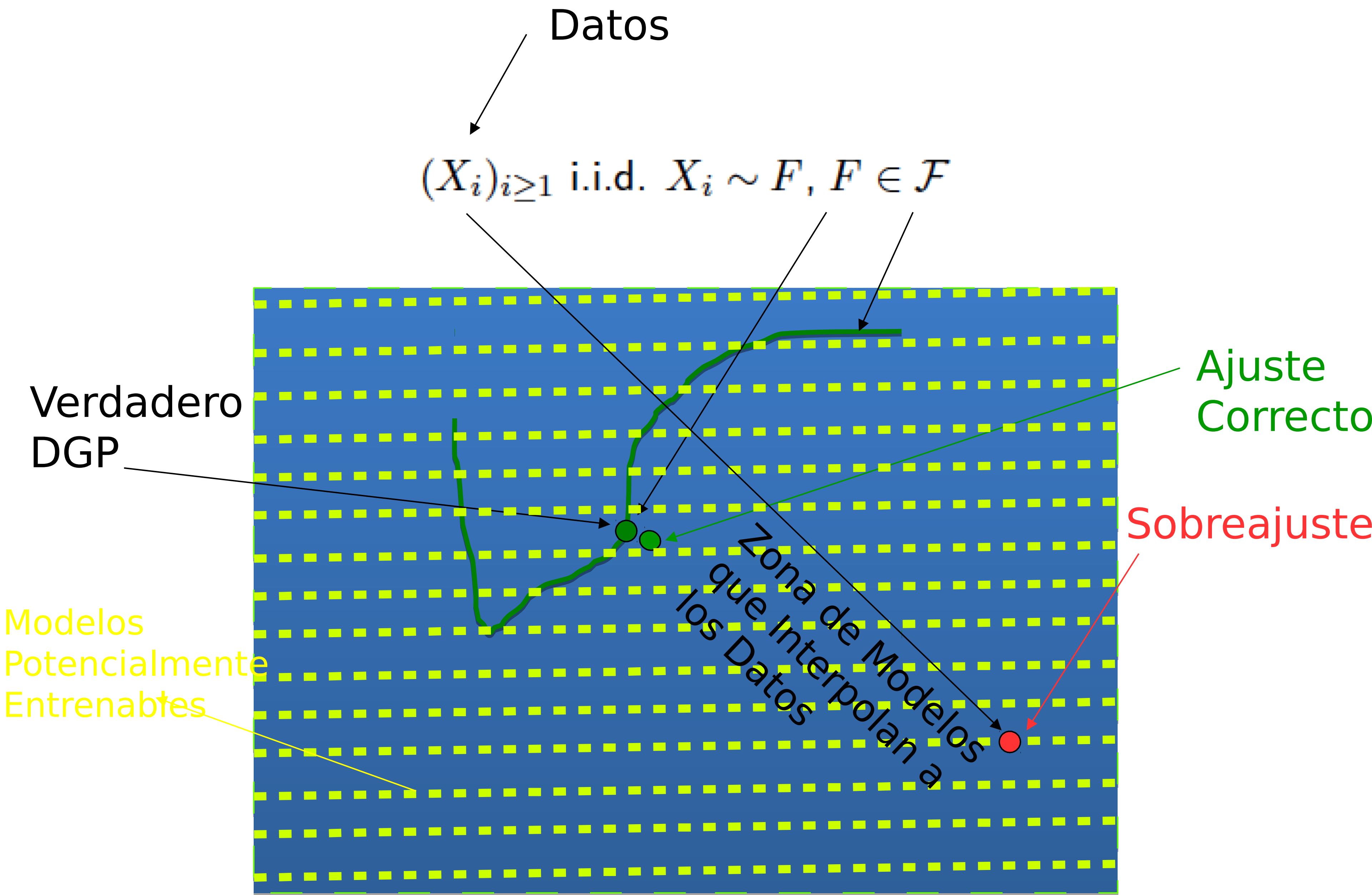
$h_n\left(x_1, x_2, \ldots, x_n\right) =$ Proporción muestral de bolitas rojas

Inferencia Estadística

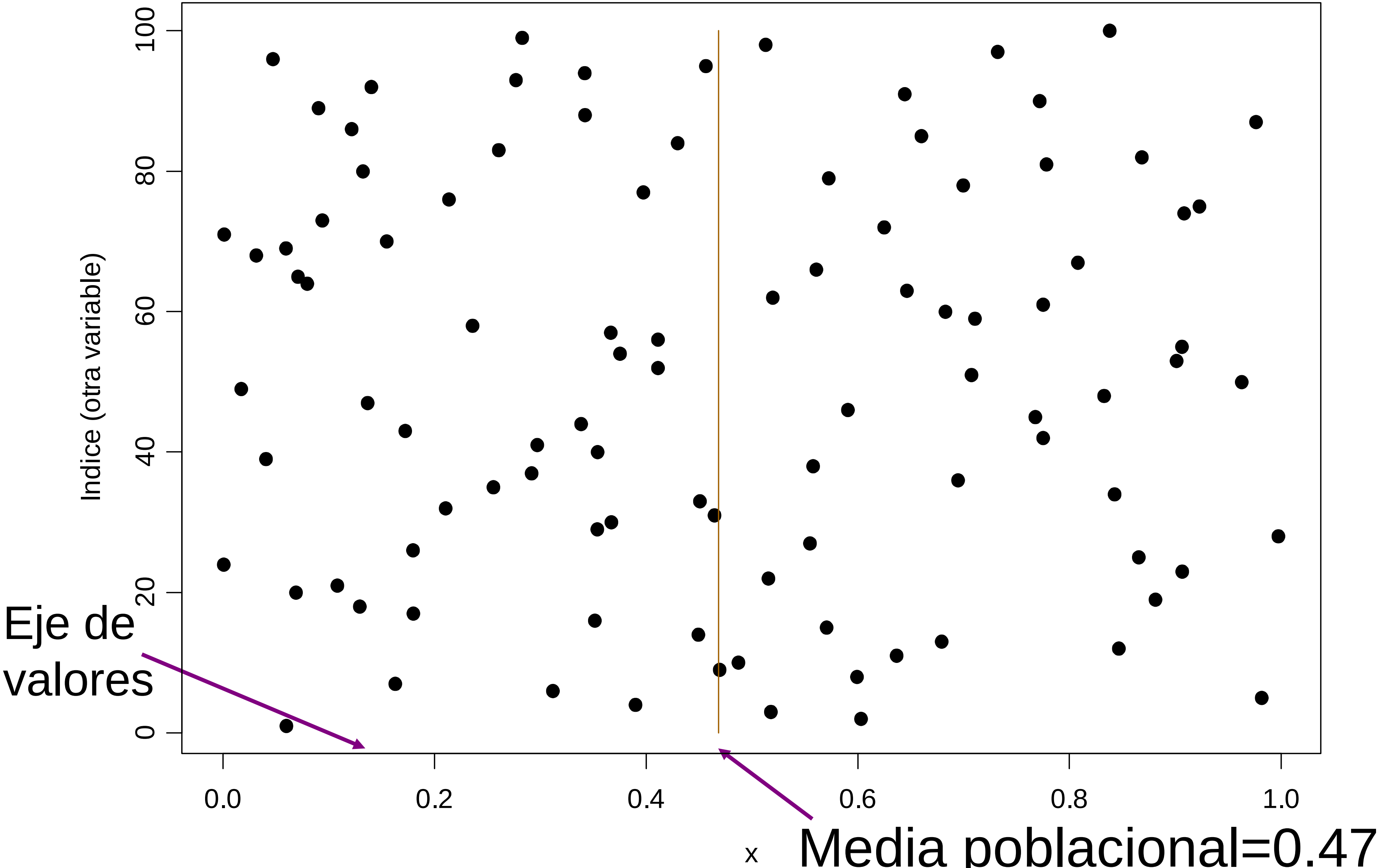
$f \in \mathcal{M} = \{f(\cdot,\theta), \theta \in \Theta\}.$



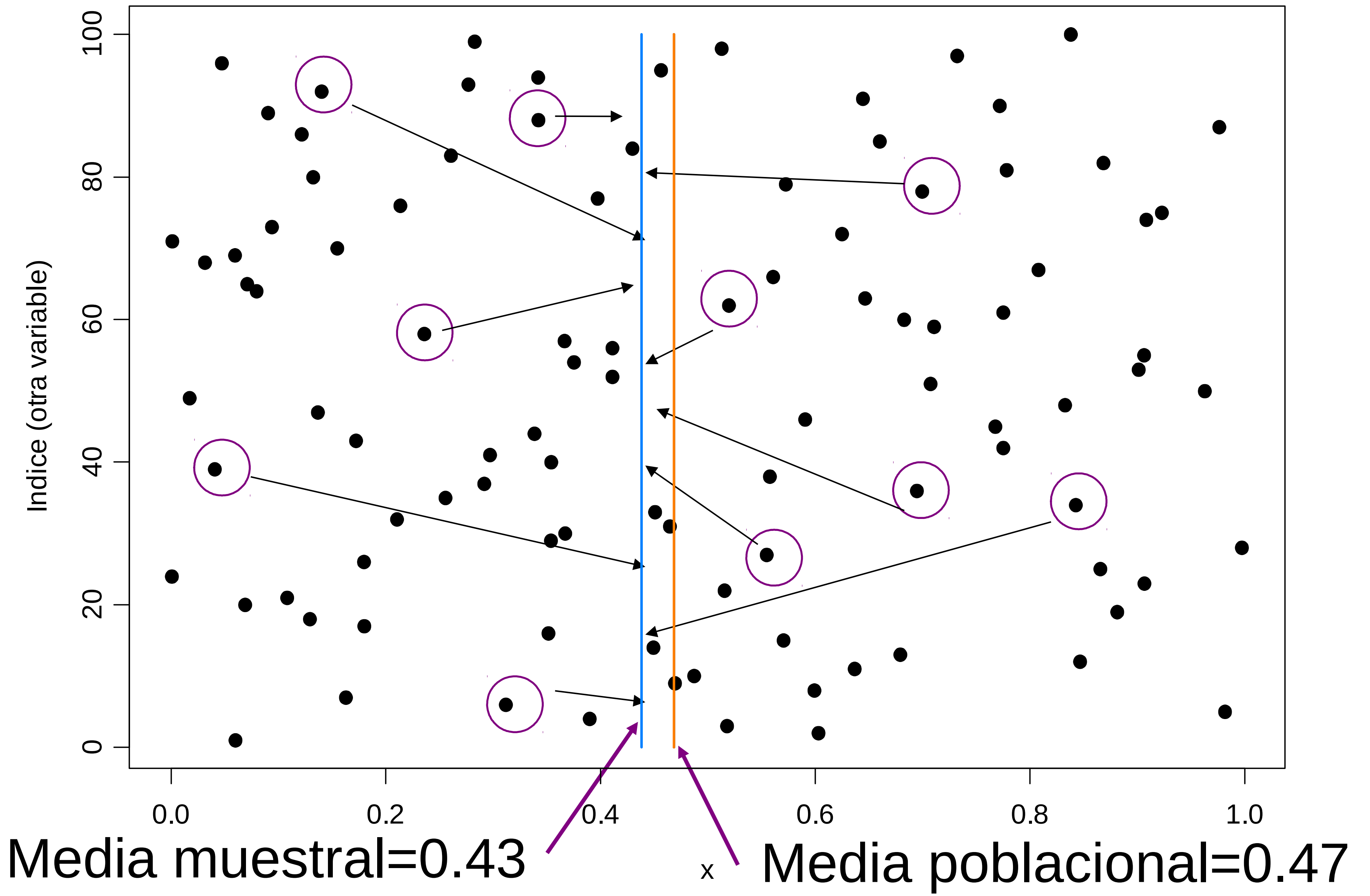
Aproximadores Universales (ANN - MLP)



La población (N=100)



La muestra (n=10)



El estimador y la estimación

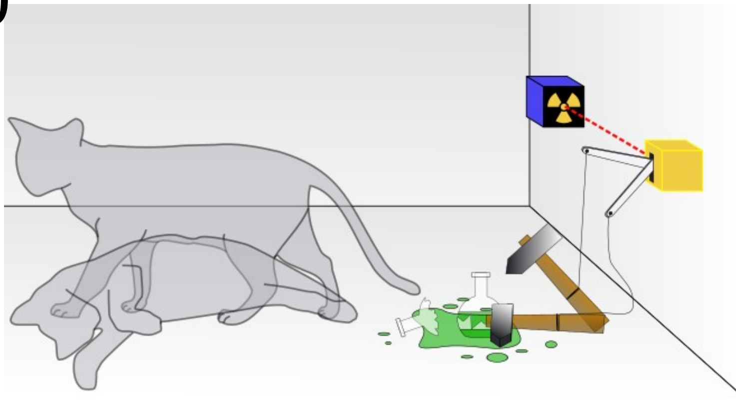
Dados los datos: x_1, x_2 … x_{10}

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{x_1}{10} + \frac{x_2}{10} + \dots + \frac{x_{10}}{10} = 0.43$$

Dados las variables aleatorias: x_1, x_2 … x_{10}

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{x_1}{10} + \frac{x_2}{10} + \dots + \frac{x_{10}}{10}$$

Schrödinger's cat

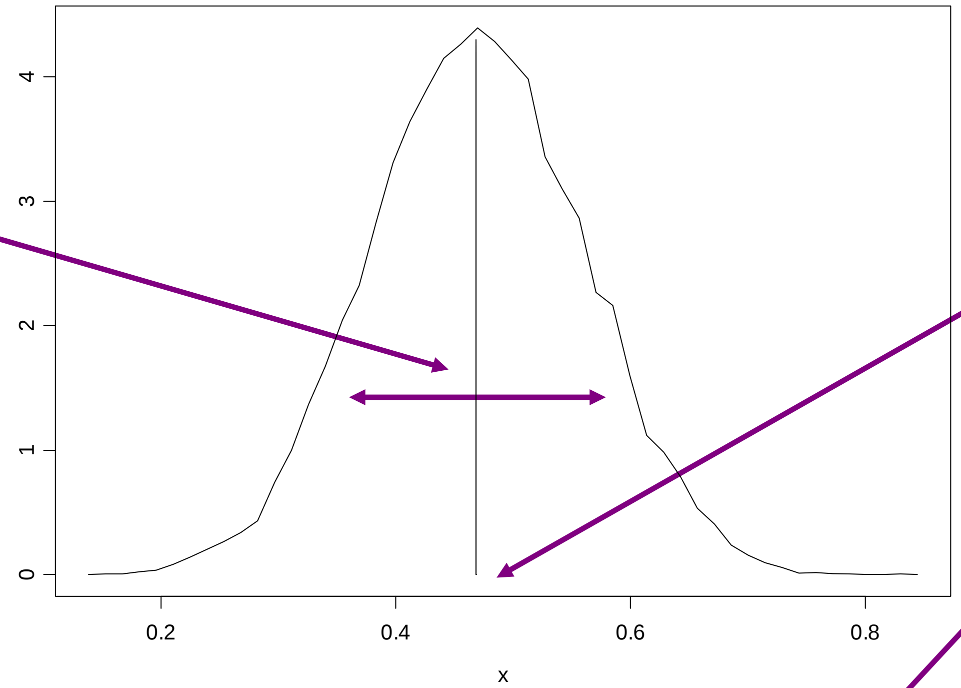


¿ Que tan bueno es ?

¿ Cuales son sus propiedades?

Repito el experimento 10000 veces

Desvío
= 0.09



Media
poblacional
= 0.468

Insesgado

```
  Min.   st Qu.  Median    Mean   3rd Qu.   Max.
0.15705 0.40697 0.46900 0.46879 0.53011 0.82568
```

Propiedades de los Estimadores

- Consistencia

$$\boxed{\widehat{\theta}_n \to \theta}$$

Parámetro

Estimador

- Insesgadez

$$\boxed{\mathbb{E}(\widehat{\theta}_n) = \theta}$$

- Error Cuadrático Medio

$$\boxed{\text{ECM}(\widehat{\theta}_n) = \underbrace{\mathbb{V}(\widehat{\theta}_n)}_{\text{Varianza}} + \underbrace{\left\{\mathbb{E}(\widehat{\theta}_n) - \theta\right\}^2}_{\text{Sesgo\^{}2}}}$$

$$\boxed{\mathbb{E}(\widehat{\theta}_n - \theta)^2}$$

Verosimilitud

Parámetros

Variables Aleatorias

- Modelo: $\mathcal{M} = \{p(\cdot, \theta) \, , \theta \in \Theta\}$.
- $\mathbf{x} = x_1, \cdots, x_n$ realización de $X_1, \cdots, X_n$ i.i.d.
- Función de verosimilitud asociada a $\mathbf{x} = x_1, \cdots, x_n$:

$$L(\cdot \, ; \, \mathbf{x}) \, : \Theta \to \mathbb{R}$$

Los Parámetros varían !!!

$$L(\theta \, ; \, \mathbf{x}) \, = \, \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) \, , X_i \sim p(\cdot, \theta) \, .$$

$$\boxed{L(\theta \, ; \, \mathbf{x}) \, = \, \prod_{i=1}^{n} p(x_i, \theta) \, ,}$$

Observaciones están fijas !!!

Estimador de Máxima-Verosimilitud

- Función de verosimilitud: $L(\cdot\ ;\ \mathbf{x}\ )\ : \Theta \to \mathbb{R}$

$$L(\theta\ ;\ \mathbf{x}\ )\ =\ f_{X_1,\cdots,X_n}(x_1,\ldots,x_n)\ , X_i \sim f(\cdot,\theta)\ .$$

$$L(\theta\ ;\ \mathbf{x}\ )\ =\ \prod_{i=1}^{n} f(x_i,\theta)\ ,$$

- Propuesta de Máxima Verosimilitud:

$$\boxed{h_n(\mathbf{x}) = \underset{\theta\in\Theta}{\operatorname{argmax}}\ L(\theta\ ;\ \mathbf{x}\ )\ .}$$

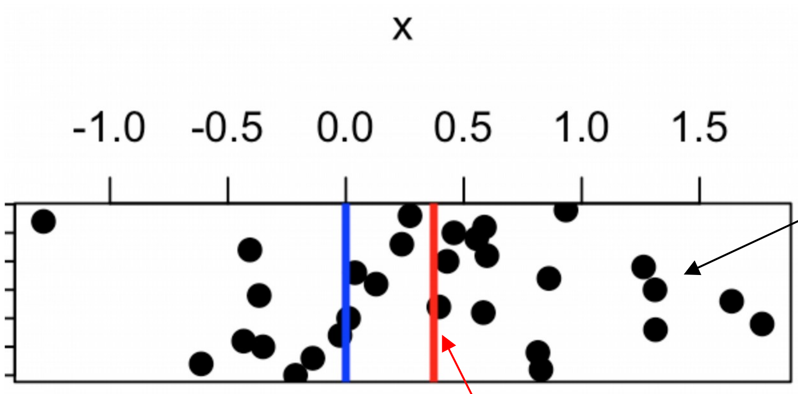
o sea

$$L(h_n(\mathbf{x}),\mathbf{x}) \geq L(\theta,\mathbf{x})$$

- Definimos el EMV siendo $\widehat{\theta}_n = h_n(X_1,\cdots,X_n)$.

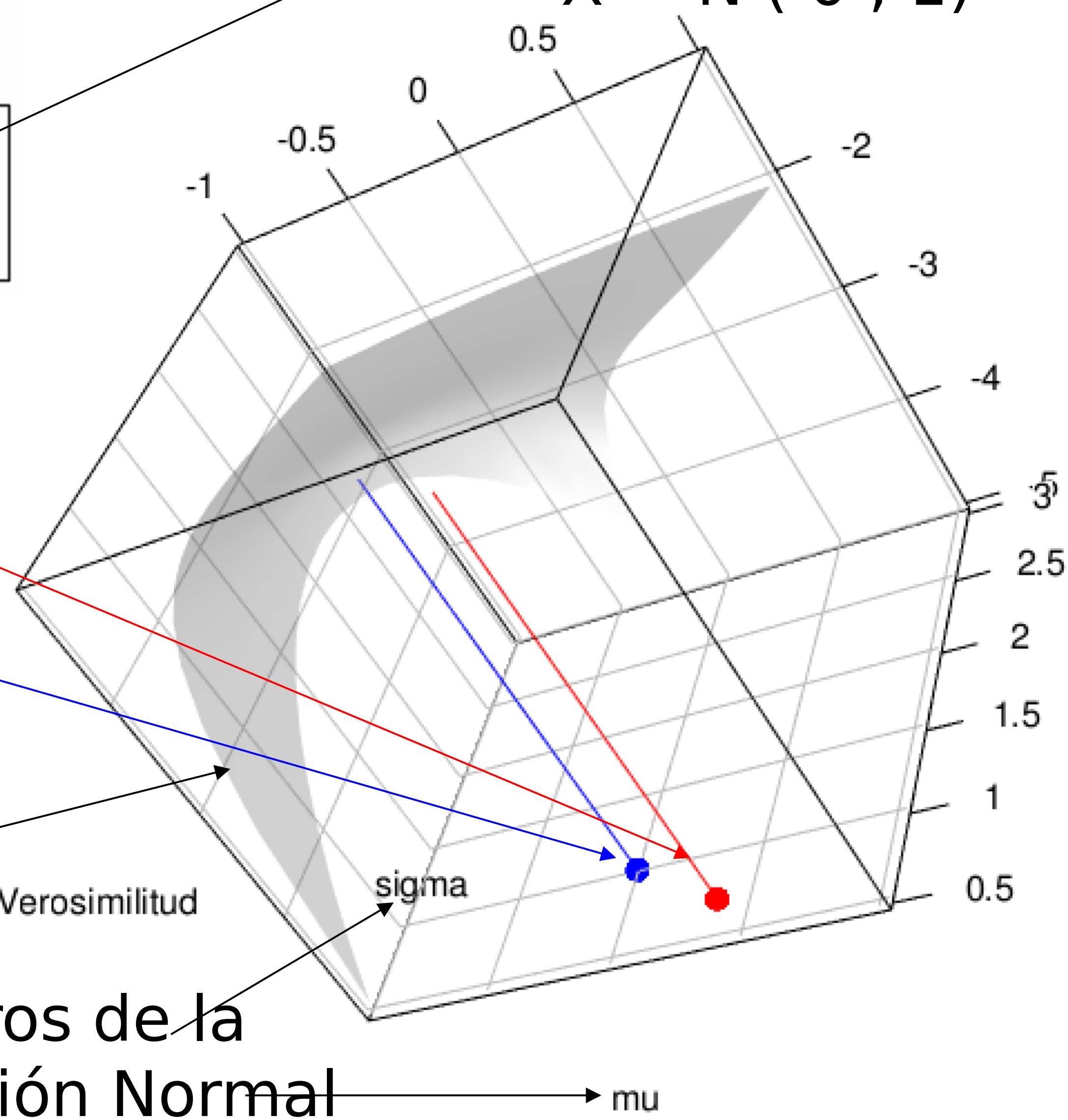
Ejemplo de Verosimilitud

30 realizaciones X ~ N (0 , 1)



Media verdadera = 0

Media estimada MV = 0.37



Función de Verosimilitud (Normal) basada en los 30 x’s fijos

Parámetros de la Distribución Normal

Selección de Modelos

- Modelo: $\mathcal{M} = \{f(\cdot, \theta), \theta \in \Theta\}$

- Verosimilitud (likelihood): $L(\theta\ ,\ \mathbf{x}\ )\ =\ \prod_{i=1}^{n} f(x_i, \theta)$

$$\widehat{\theta}(\mathbf{x}) = \underset{\theta\in\Theta}{\operatorname{argmax}}\ L(\theta\ ;\ \mathbf{x}\ )\ , \quad \text{o sea} \quad L(\widehat{\theta}(\mathbf{x}), \mathbf{x}) \geq L(\theta, \mathbf{x}).$$

- log-vero (log- likelihood): $\ell(\theta\ ,\ \mathbf{x}\ )\ =\ \sum_{i=1}^{n} log(f(x_i, \theta))$
- Bondad del Modelo:

$$\text{Bondad}(\mathcal{M}, \mathbf{x}) := \ell(\widehat{\theta}(\mathbf{x})\ ,\ \mathbf{x}\ )\ =\ \sum_{i=1}^{n} log(f(x_i, \widehat{\theta}(\mathbf{x})))$$

$$\boxed{\text{AIC} = \text{AIC}(\mathcal{M}, \mathbf{x}) := -2\left(\ell(\widehat{\theta}(\mathbf{x}),\ \mathbf{x}\ ) - \#\text{parámetros}\right)}$$

Intervalos de confianza

Dado un parámetro poblacional desconocido, buscamos un intervalo (dependiente de la muestra) que con alta probabilidad contenga al verdadero valor del parámetro.

Intervalos de confianza

Dados $x_1 \ldots x_n$ muestra aleatoria proveniente de una población con parámetro θ

Límite inferior

Límite superior

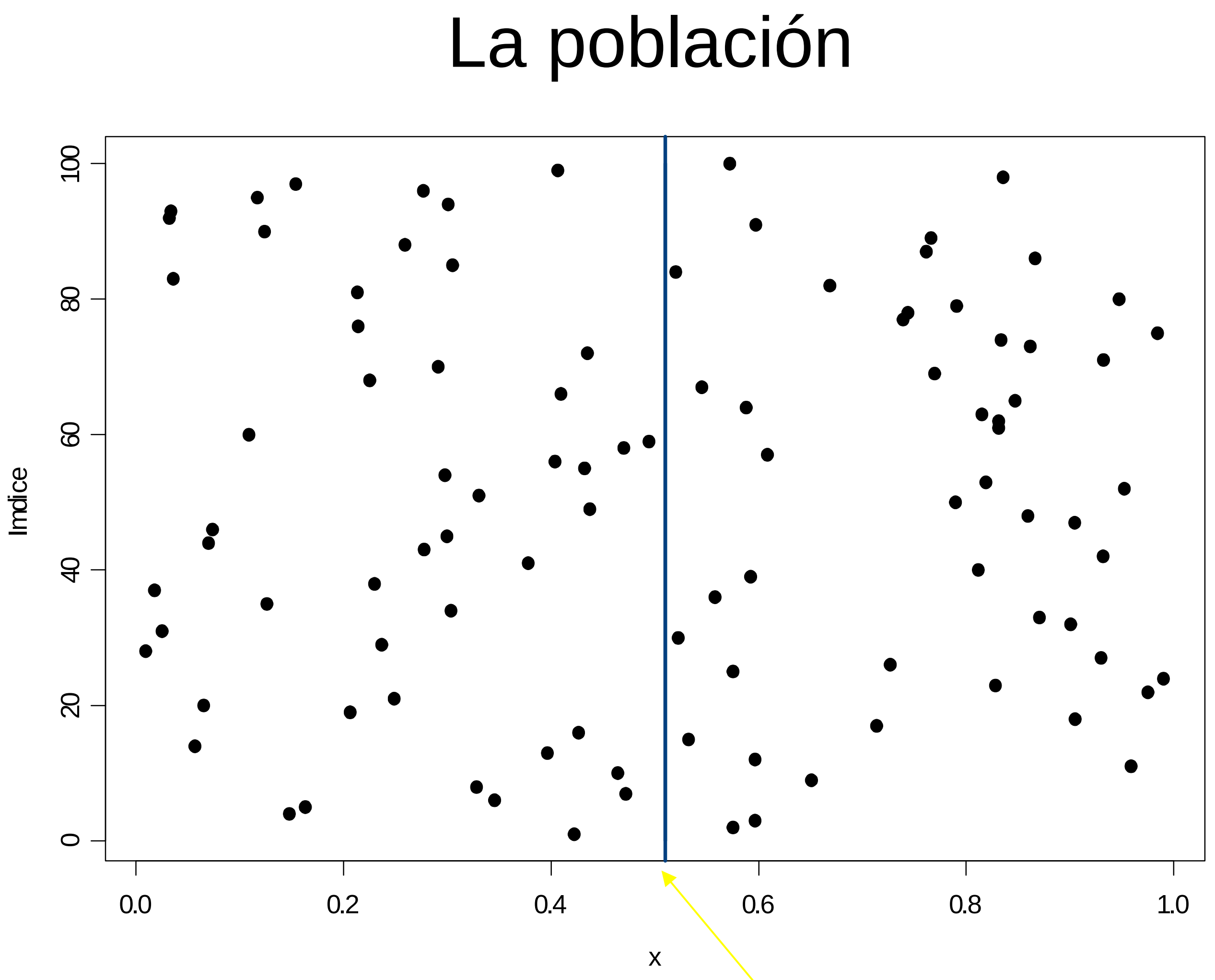
$\left[I(x_1 \ldots x_n),\, D(x_1 \ldots x_n) \right]$ es un intervalo de confianza 0.95 si

Fijo

$$P\left(I(x_1 \ldots x_n) \le \theta \le D(x_1 \ldots x_n)\right) = 0.95$$

Aleatorio

Intervalos de confianza: Ejemplo



Media poblacional = 0.47

Propongo un intervalo de confianza (cualquiera)

Dada una muestra aleatoria de dos elementos x_1 y x_2

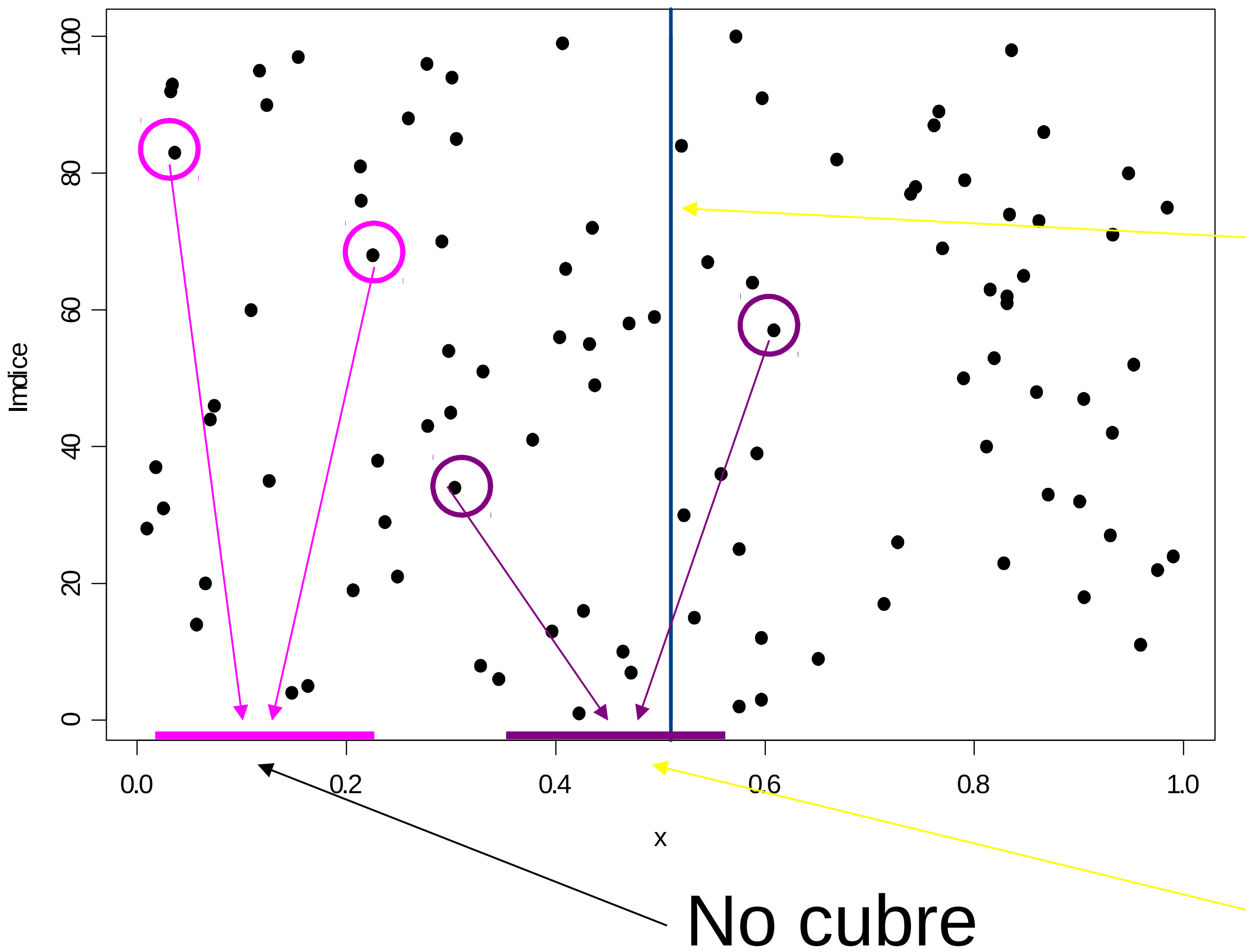
$$\left\lfloor I\left(x_1 \ldots x_n\right), D\left(x_1 \ldots x_n\right) \right\rfloor$$

$$=$$

$$\left[\left(\frac{x_1 + x_2}{2} \right) - 0.1, \left(\frac{x_1 + x_2}{2} \right) + 0.1 \right]$$

¿ Con que probabilidad cubre a la verdadera media poblacional (0.47)?

Dos realizaciones del intervalo de confianza (cualquiera)



Media poblacional = 0.47

No cubre

Si cubre

Repito el experimento 10000 veces

Repetición	Cubre ?
1	NO
2	NO
3	SI
4	NO
…	…
10000	NO

Proporción de intervalos que cubren = 0.3503

¿ Que pasa si tomo tamaño de muestra = 4 ?

Proporción de intervalos que cubren = 0.4871

¿ Que pasa si tomo tamaño de muestra = 8 ?

Proporción de intervalos que cubren = 0.6543

¿ Que pasa si tomo tamaño de muestra = 8 y longitud de intervalo =0.4 ?

Proporción de intervalos que cubren = 0.8558

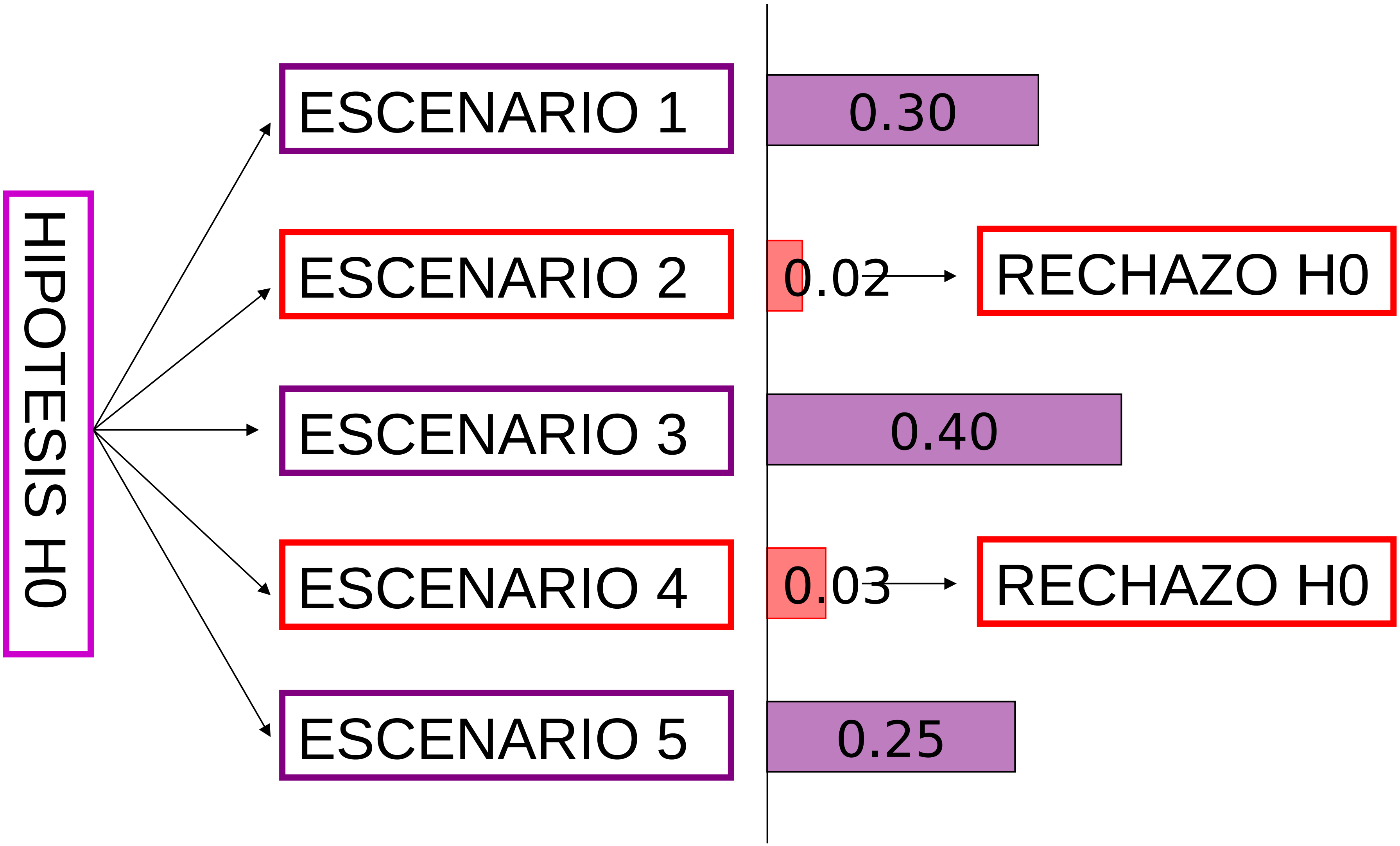
Test de Hipótesis

Es un mecanismo para decidir acerca de la validez de una hipótesis, controlando la probabilidad de rechazar la misma siendo que esta es verdadera.

Intuitivamente

- Nos paramos en la hipótesis que queremos validar y pensamos los diferentes escenarios posibles con sus probabilidades (según la hipótesis)
- Si la realidad se corresponde con un escenario que bajo la hipótesis es poco probable, rechazamos la hipótesis

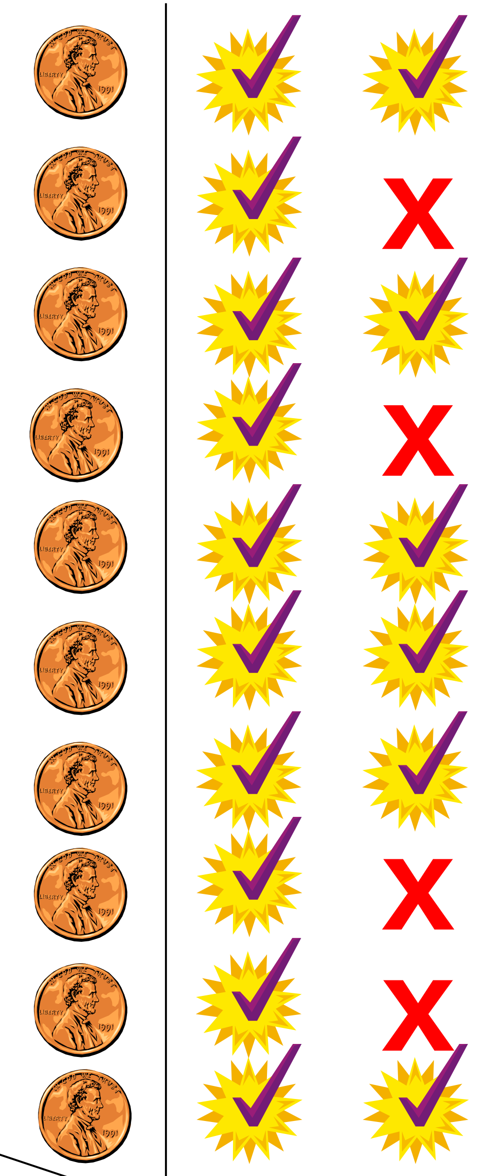
Gráficamente



Receta para armar un Test de Hipótesis

- Definir la Hipótesis nula (H0)
- Elegir un estadístico que mida o refleje el alejamiento de la evidencia de H0
- Definir un valor de probabilidad (α) por debajo del cual creamos que los eventos son suficientemente “raros”
- Evaluar el estadístico en los datos y comparar la probabilidad de un resultado como ese o “mas extremo” con α

H0: Juan NO es adivino



$P(Z >= 10) =$
$= 1/1024 = 0.000977$

$P(Z >= 6) =$
$= 0.377$

Z = Cantidad de aciertos 10 6

Potencia, Específicidad y Errores

Modelo

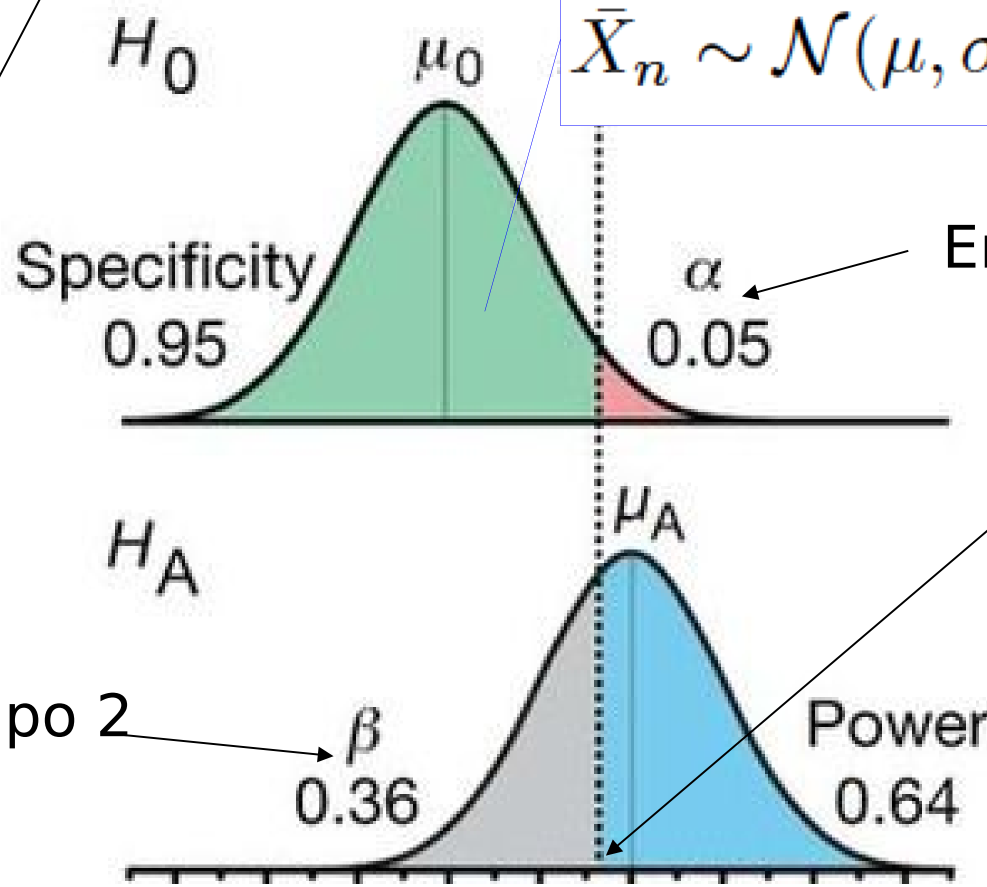
$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

Estimador

$$\bar{X} = \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$N(0, \sigma)$$

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$



Error Tipo 1

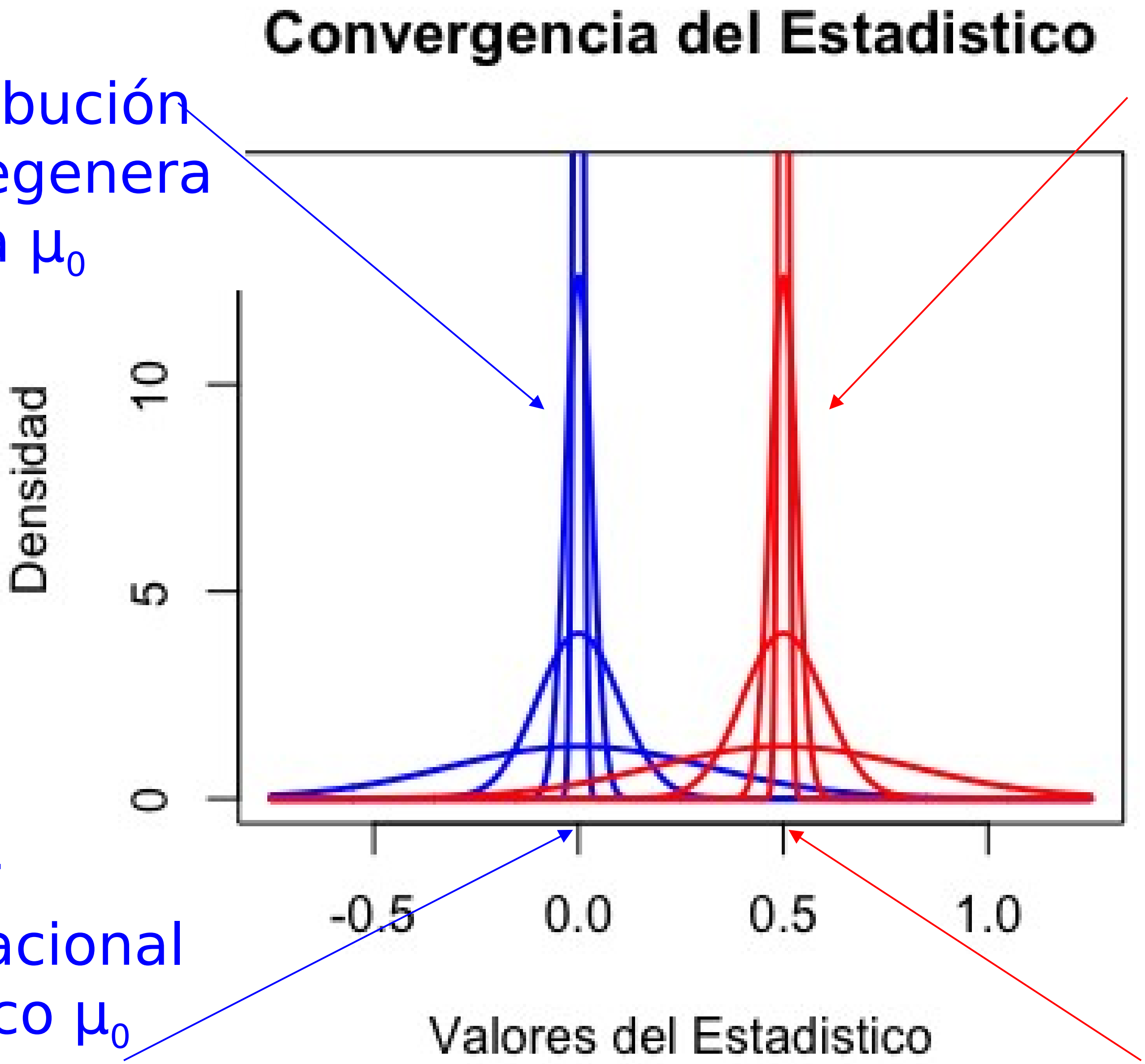
Contra mayor, rechazo si supera el umbral

Error Tipo 2

La Virtud y El Problema de la Consistencia

La distribución se degenera hacia μ_0

La distribución se degenera hacia μ_A



Valor poblacional teórico μ_0

Valor poblacional real μ_A

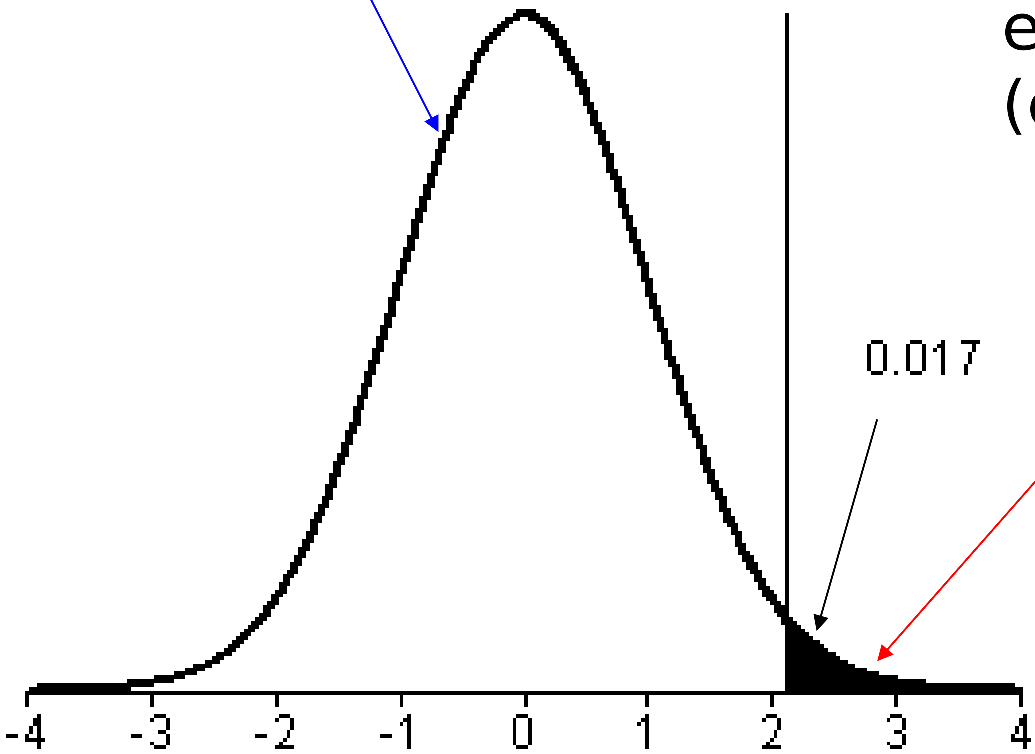
El p-valor

- Concepto fundamental de la Estadística que cuantifica objetivamente la evidencia acerca de la validez de una hipótesis.
- Específicamente, mide en base a las observaciones el graodo de "compatibilidad" de una hipótesis en términos del comportamiento distribucional de un estimador /estadístico.

Gráficamente

Distribución del Estadístico bajo la Hipótesis

Valor observado del estadístico (estimación)



Probabilidad de observar “algo tan o mas extremo” que lo observado basado en la muestra

Evidencia a favor H0 Evidencia contra H0

Para que sirve el Enfoque Estadístico ?

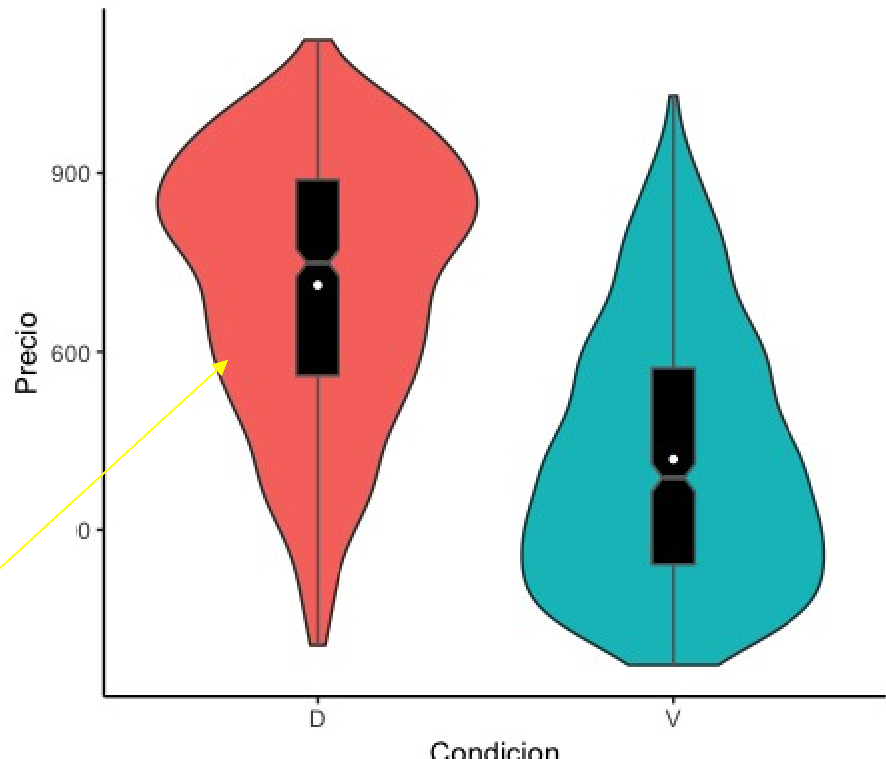
- OVBIO: Para cuantificar la incertidumbre de las estimaciones.
-
- OVBIO: Para completar la falta de información con “relaciones” matemáticas razonables/justificadas.

- NO TAN OVBIO: Para Modelar correctamente los fenómenos de interes, discriminando las relaciones “concomitantes” de aquellas que son “esenciales”

Los Agentes Inmobiliarios venden sus casas mas caras que las de sus clientes ?

	pre	due
	639.4115	V
	218.7228	V
	498.5153	D
	307.9519	V
	604.6274	D
	452.0448	V

Precios en miles de u$s



Claramente, venden SUS casas mas caras

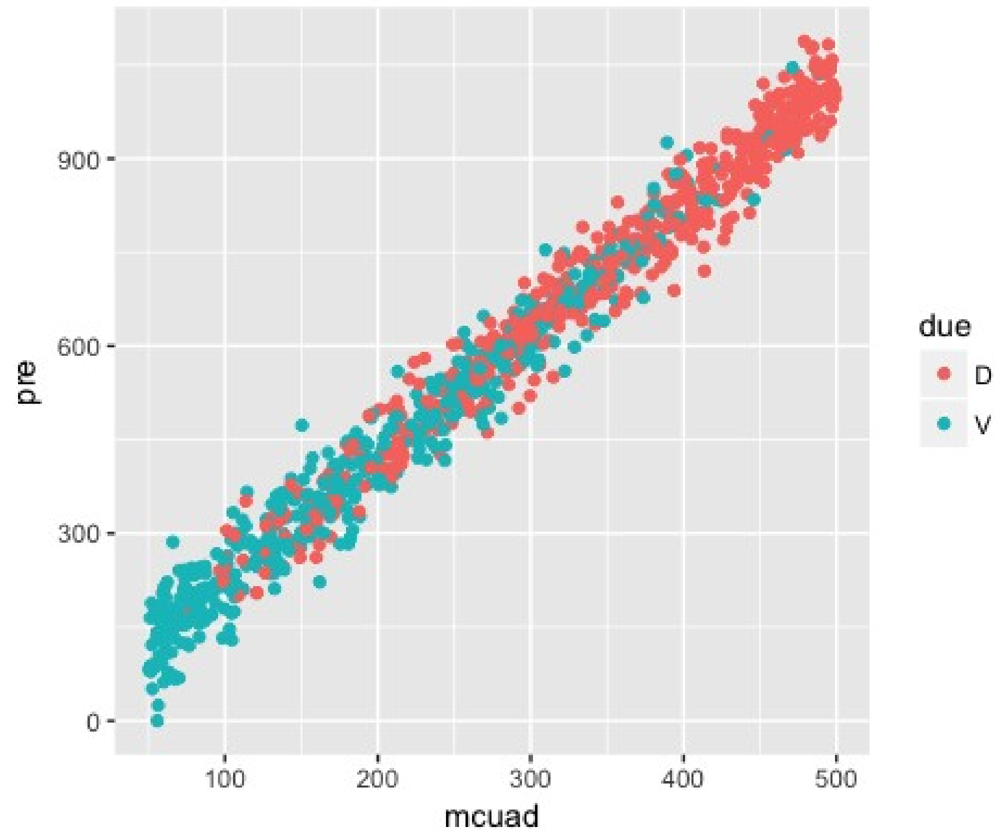
```
> table(due)
due
  D   V
503 497
```

Co

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	710.391	9.633	73.75	<2e-16	***
dueV	-284.165	13.555	-20.96	<2e-16	***

Problema de Especificación

	pre	due	mcuad
1	421.2050	V	199.8858
2	335.3559	V	158.8452
3	852.9682	D	403.0014
4	601.3251	V	280.2027
5	674.4204	D	327.0117
6	710.6430	V	357.1478



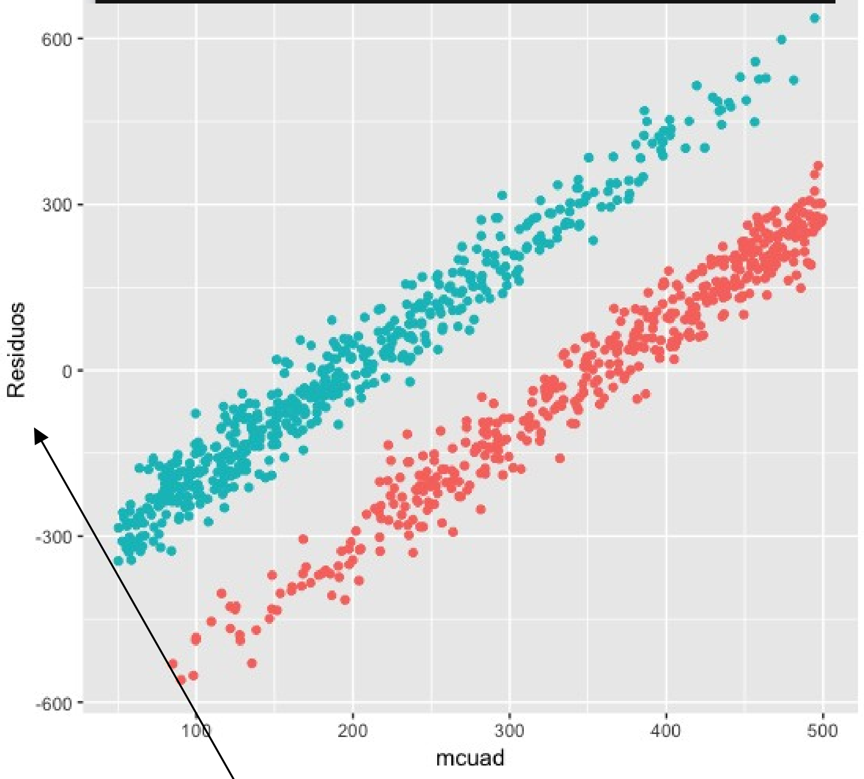
El factor (D/V) NO es significativo

Coefficients:

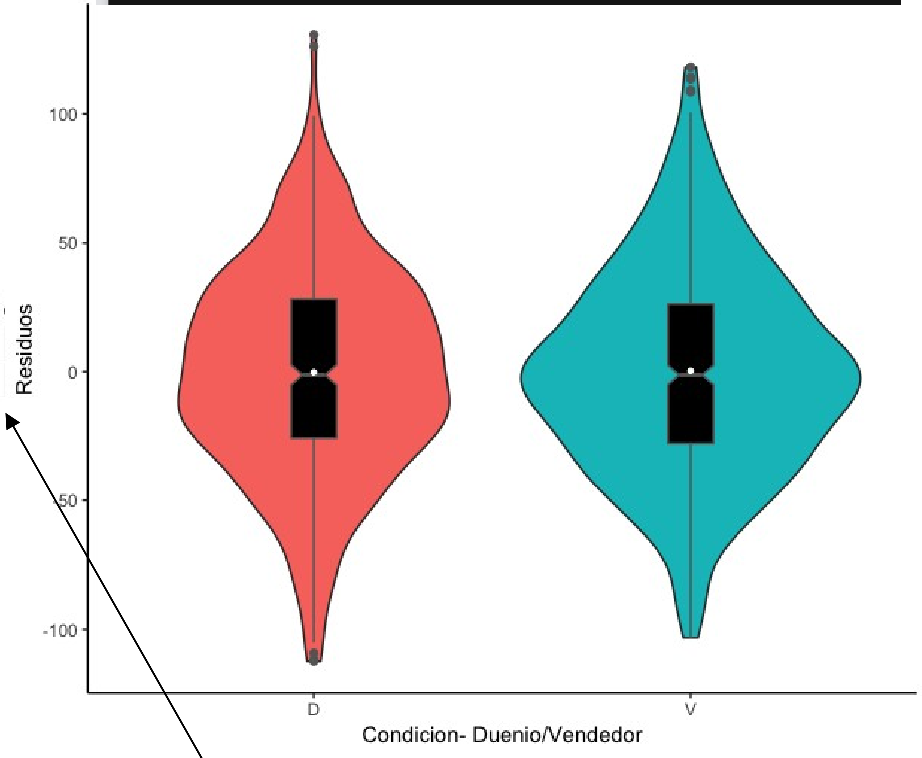
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	23.49817	4.58556	5.124	3.58e-07	***
mcuad	1.99409	0.01221	163.352	< 2e-16	***
dueV	-1.84336	3.10025	-0.595	0.552	

Que Está Pasando ?

```
lm(formula = pre ~ due)
```




```
lm(formula = pre ~ mcuad)
```



Despues del ajuste queda mucha estructura en los residuos

Nada por explicar en los residuos

Inferencia Bayesiana

Distribución Posterior de los Parámetros considerando: Prior + Datos

Likelihood de los Datos bajo el Modelo

Distribución Prior

$$p(\Theta|\mathbf{y}) = \frac{p(\mathbf{y}|\Theta)p(\Theta)}{p(\mathbf{y})}$$

Parámetros aleatorios

Observaciones

Distribución Marginal de las Observaciones

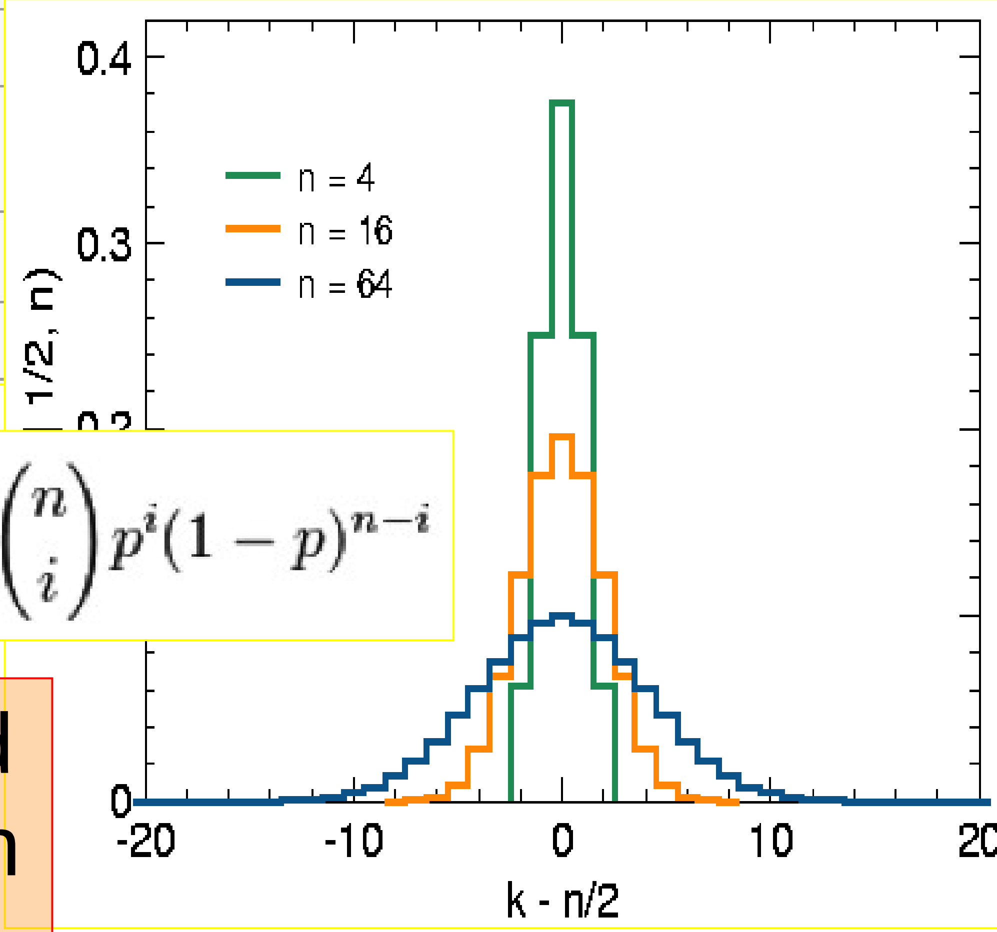
$$p(\Theta|\mathbf{y}) \propto p(\mathbf{y}|\Theta)p(\Theta)$$

$$p(\mathbf{y}) = \int p(\mathbf{y}|\Theta)p(\Theta)d\Theta$$

Mecanismo de Actualización de la Distr. de los Parámetros

Distribución Binomial

notation:	$B(n, p)$
parameters:	$n \in \mathbf{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
support:	$k \in \{0, \ldots, n\}$
pmf:	$\binom{n}{k} p^k (1-p)^{n-k}$
cdf:	$I_{1-p}(n-k, 1+k)$
mean:	np

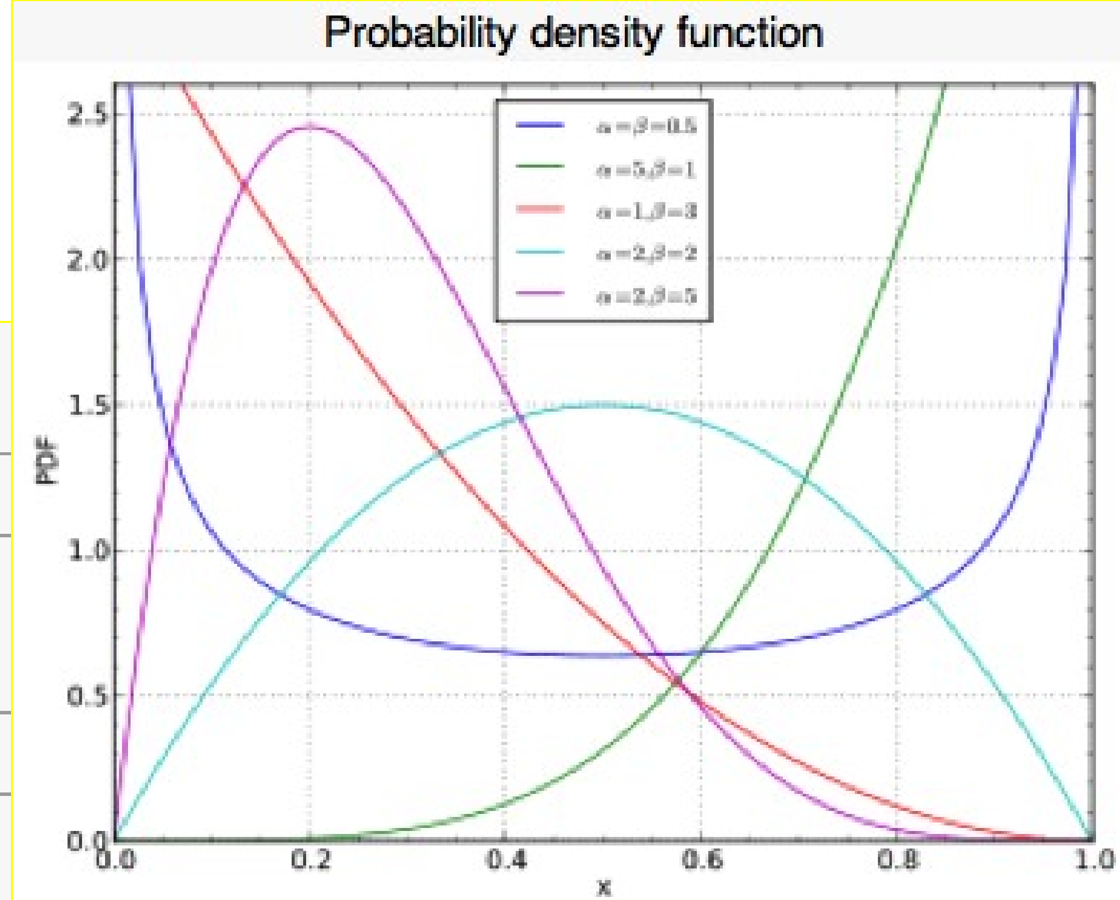


$$F(x; n, p) = \Pr(X \leq x) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

Cual es la probabilidad de observar k éxitos en n intentos ?

Distribución Beta

parameters:	$\alpha > 0$ shape (real) $\beta > 0$ shape (real)
support:	$x \in (0;1)$
pdf:	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$
cdf:	$I_x(\alpha,\beta)$
mean:	$\frac{\alpha}{\alpha+\beta}$
median:	$I_{0.5}^{-1}(\alpha,\beta)$ no closed form
mode:	$\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha > 1, \beta > 1$
variance:	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



Cual es la probabilidad de una probabilidad ?

Ejemplo Sencillo: Binomial y Beta

$$\mathrm{Beta}(\alpha, \beta)$$

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \pi(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

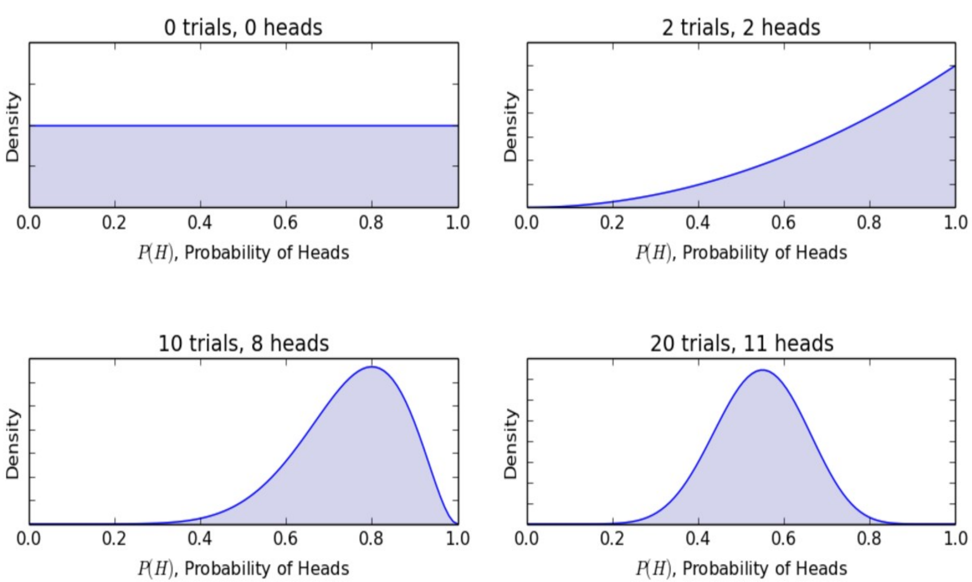
Evento dicotómico $\begin{cases} 1 & p \\ 0 & 1\text{-}p \end{cases}$

$$f(p|x) = \frac{f(x|p)}{f_X(x)}\,\pi(p)$$

$$\propto p^x(1-p)^{n-x}\; p^{\alpha-1}(1-p)^{\beta-1}$$

$$\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1}$$

$$\mathrm{Beta}(x+\alpha, n-x+\beta)$$



de exitos

de casos

La información

Medición

P > 1 => multivariado

	Var 1	…	Var j	…	Var p
Ind 1	$X_{1,1}$		$X_{1,j}$		$X_{1,p}$
…					
Ind i	$X_{i,1}$		$X_{i,j}$		$X_{i,p}$
…					
Ind n	$X_{n,1}$		$X_{n,j}$		$X_{n,p}$

Las variables (columnas)

- Características o atributos cambiantes de los individuos que interesa analizar.

Los individuos (filas)

- Elementos sobre los cuales se miden los atributos.

Estadísticos básicos

	X_1	…	X_j	…	X_p
X_1	$\mathrm{Var}(X_1)$	…	$\mathrm{Cov}(X_1, X_j)$	…	$\mathrm{Cov}(X_1, X_p)$
⋮	⋮	⋮	⋮	⋮	⋮
X_j	$\mathrm{Cov}(X_j, X_1)$	…	$\mathrm{Var}(X_j)$	…	…
⋮	⋮	⋮	⋮	⋮	⋮
X_p	$\mathrm{Cov}(X_p, X_1)$	…	…	…	$\mathrm{Var}(X_p)$

Matriz de varianzas y covarianza

	Var 1	…	Var j	…	Var p
Ind 1	$X_{1,1}$		$X_{1,j}$		$X_{1,p}$
…					
Ind i	$X_{i,1}$		$X_{i,j}$		$X_{i,p}$
…					
Ind n	$X_{n,1}$		$X_{n,j}$		$X_{n,p}$

Varianza

Covarianza

Vector de medias

	Var 1	…	Var j	…	Var p
Medias	$\overline{X}_1$		$\overline{X}_j$		$\overline{X}_p$

Resumen de Información

Técnicas Factoriales

Componentes Principales – Análisis Factorial – Análisis de Correspondencia

Técnicas de Segmentación

Clusterización jerárquica – Métodos de Partición (K-medias)

	Var 1	…	Var j	…	Var p
Ind 1	$X_{1,1}$		$X_{1,j}$		$X_{1,p}$
…					
Ind i	$X_{i,1}$		$X_{i,j}$		$X_{i,p}$
…					
Ind n	$X_{n,j}$		$X_{n,j}$		$X_{n,p}$

Analisis Factorial

Inferencia

No observable

Factores o variables latentes

Observable

Variables o atributos

Medicion

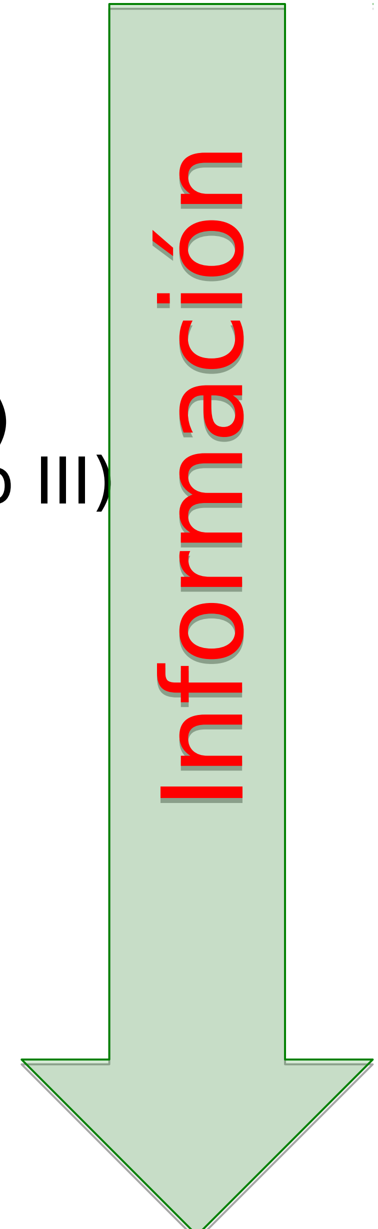
Estadística Descriptiva y Análisis Exploratorio de Datos

Objetivos:

- Conocer los datos
- Descubrir patrones
- Verificar la existencia de patrones
- Entender los patrones
- Resumir información
- Hallar asociaciones de variables
- Detectar anomalías

Las variables

- Categóricas o cualitativas
  - Color de pelo
  - Tipo de auto
  - Sexo
- Ordinales
  - Calificación de examen (A, B, C, D y E)
  - Etapa de una enfermedad (etapa I, II o III)
- Discretas
  - Cantidad de hijos
- Continuas
  - Salario
  - Peso
  - Edad
  - Tiempo



Descripciones multivariadas

- Tablas cruzadas
- Gráficos de dispersión (scatterplot)
- Hexbin
- Estimación de densidad por nucleo
- Gráficos de mosaico (mossaic plot)
- Gráficos de estrella (star plots)
- Caras

Tablas cruzadas (cross tabulation) o de Contingencia

Obsrvaciones conjuntas

Tuvo en cuenta el CONSUMO

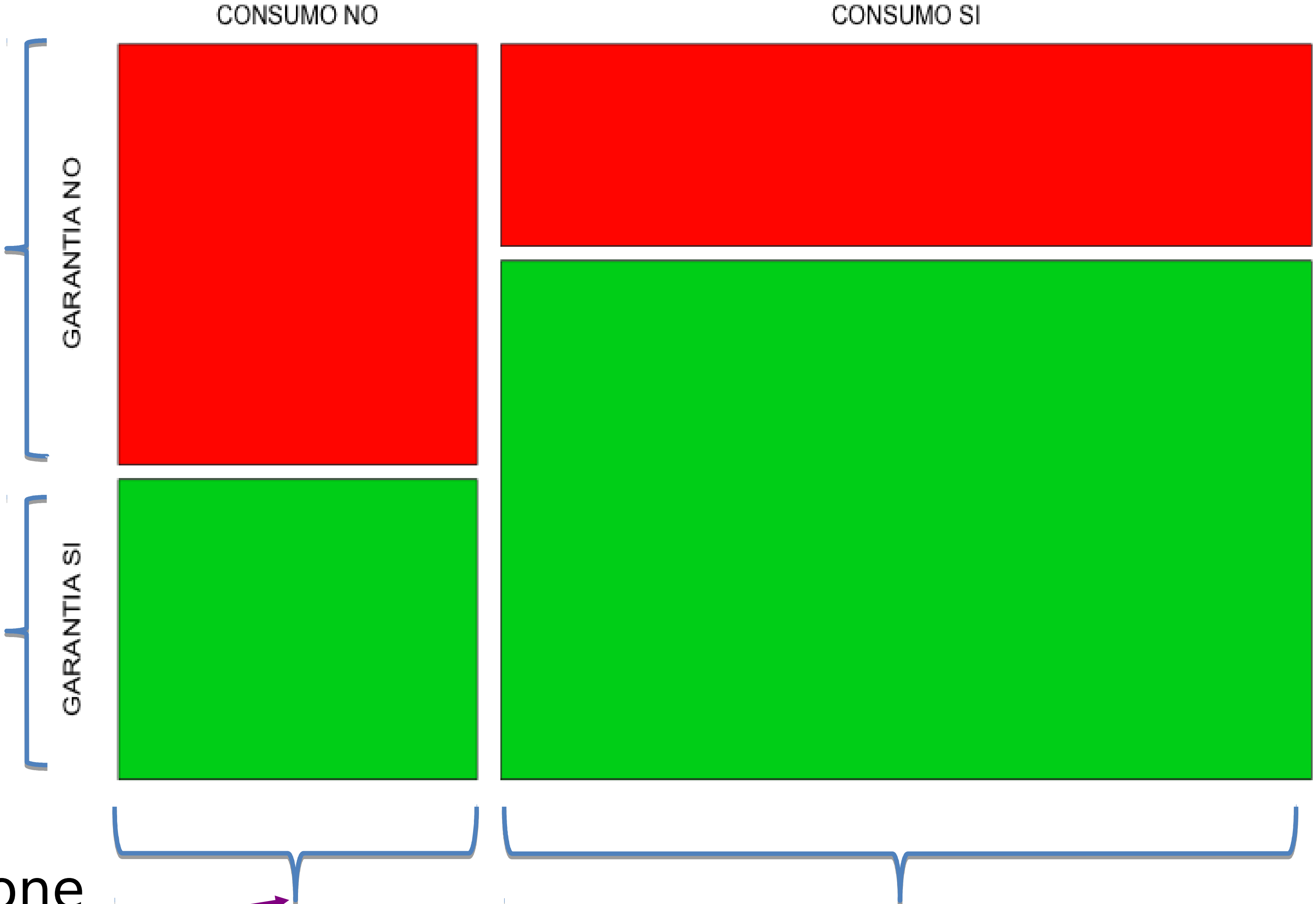
Tuvo en cuenta la GARANTIA

	NO	SI	Total
NO	258	280	538
SI	184	719	903
Total	442	999	1441

Totales marginales

Total de casos

Gráficos de mosaicos



Proporciones condicionales

Proporciones marginales

Sobrevivientes del Titanic

, , Age = Child, Survived = No

Class	Sex: Male	Female
1st	0	0
2nd	0	0
3rd	35	17
Crew	0	0

, , Age = Child, Survived = Yes

Class	Sex: Male	Female
1st	5	1
2nd	11	13
3rd	13	14
Crew	0	0

, , Age = Adult, Survived = No

Class	Sex: Male	Female
1st	118	4
2nd	154	13
3rd	387	89
Crew	670	3

Total de casos = 2201

, , Age = Adult, Survived = Yes

Class	Sex: Male	Female
1st	57	140
2nd	14	80
3rd	75	76
Crew	192	20

Mosaico del Titanic

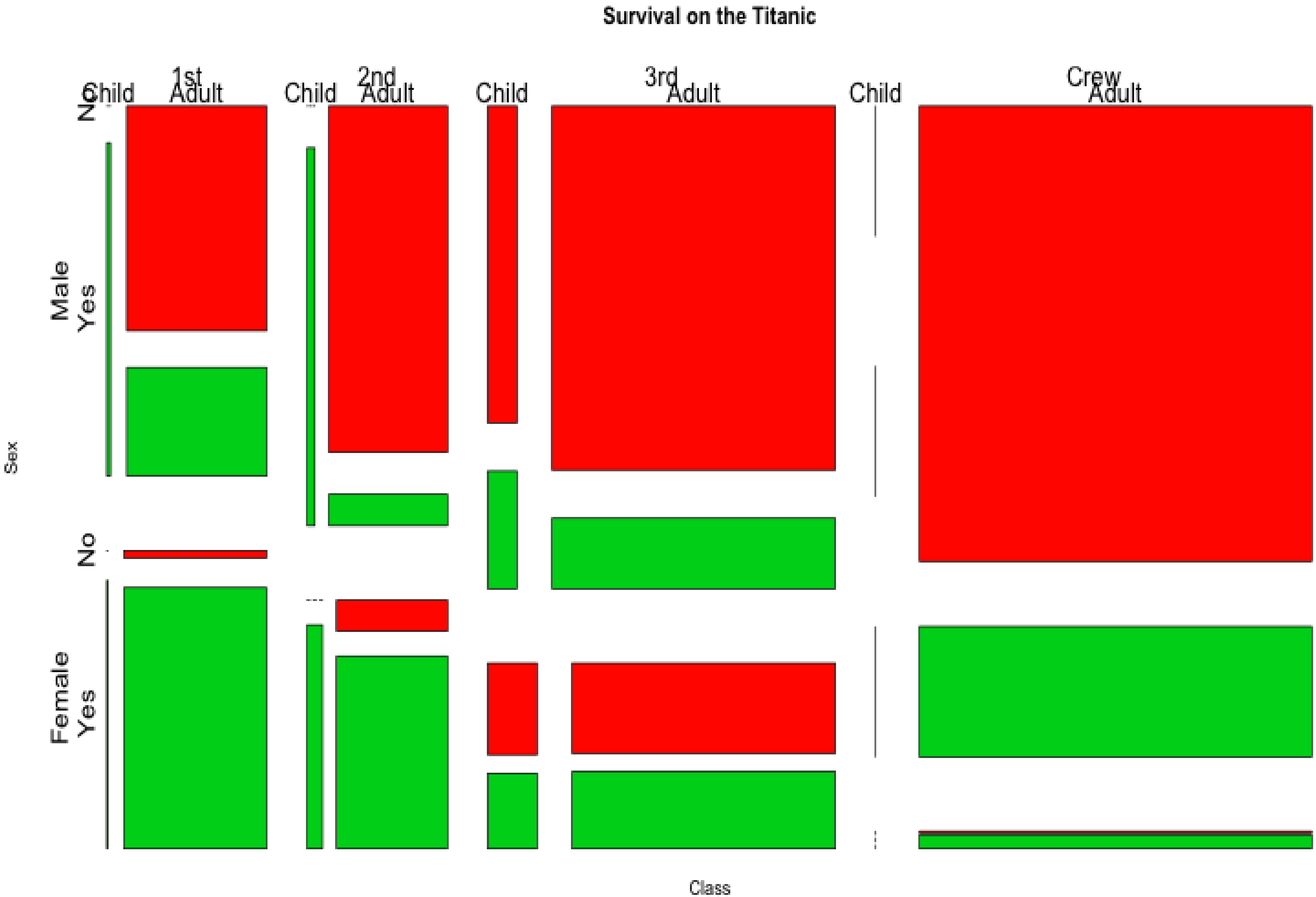


Gráfico de dispersión (X,Y)

Dos variables

Mediciónes

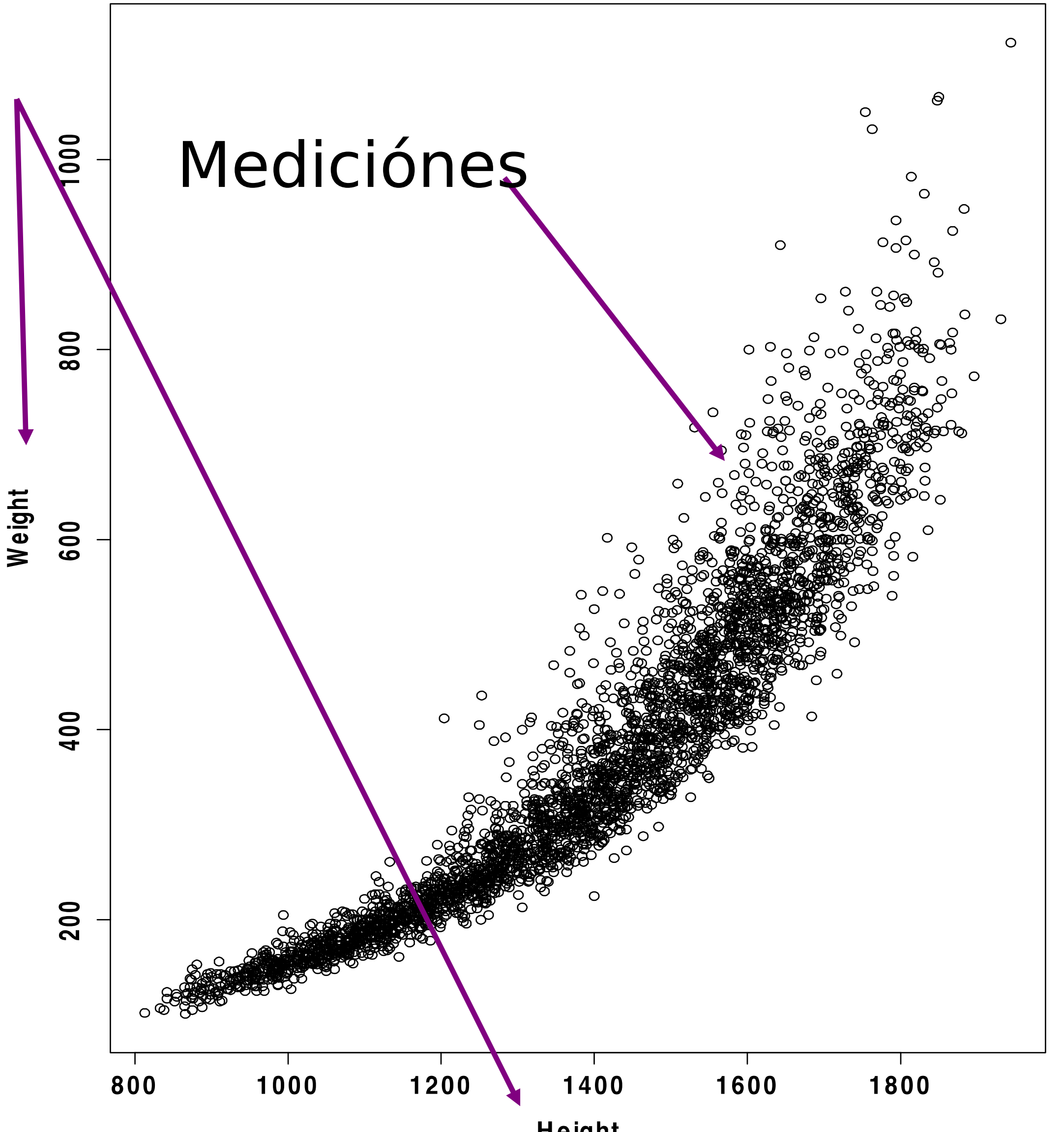
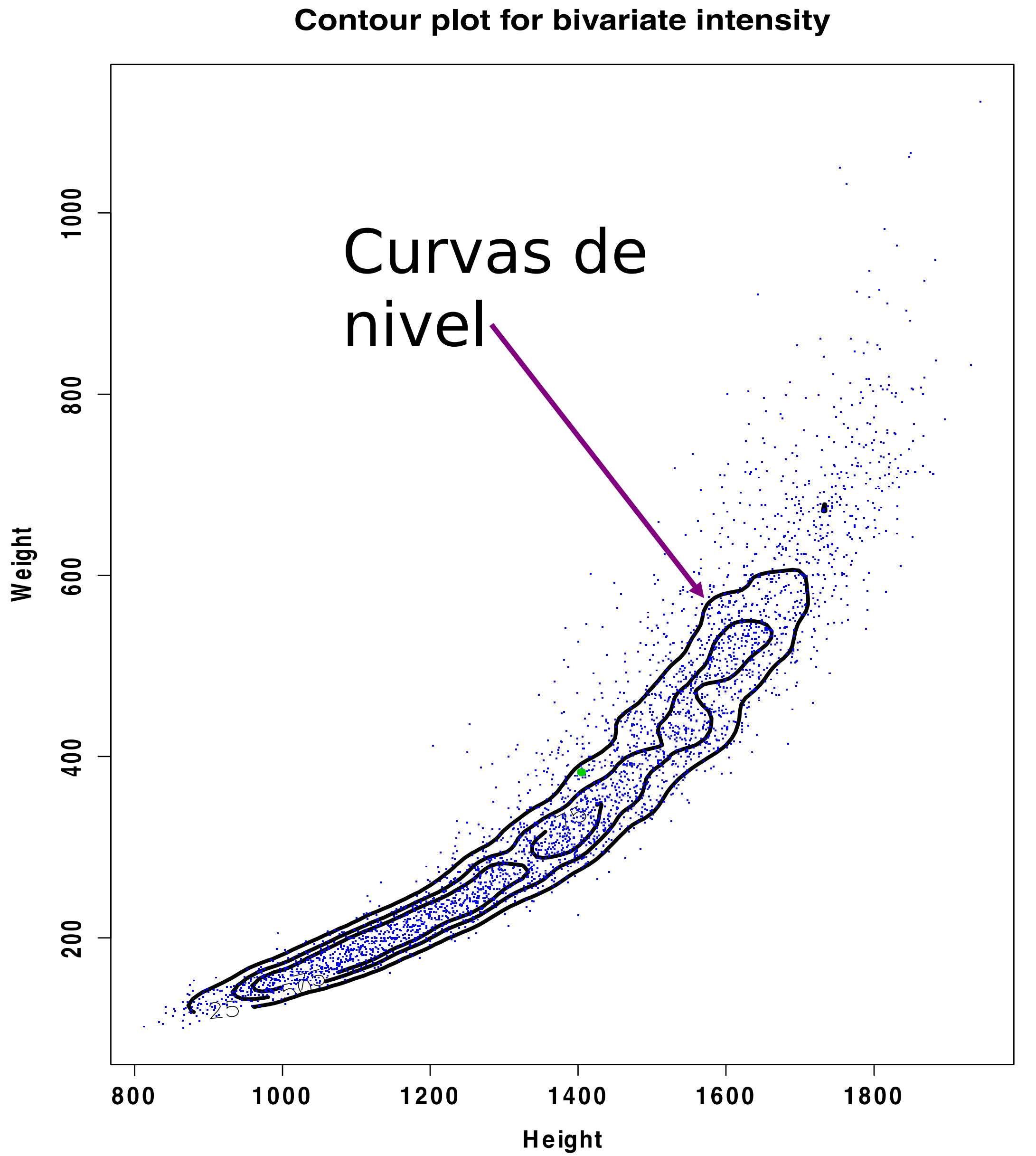
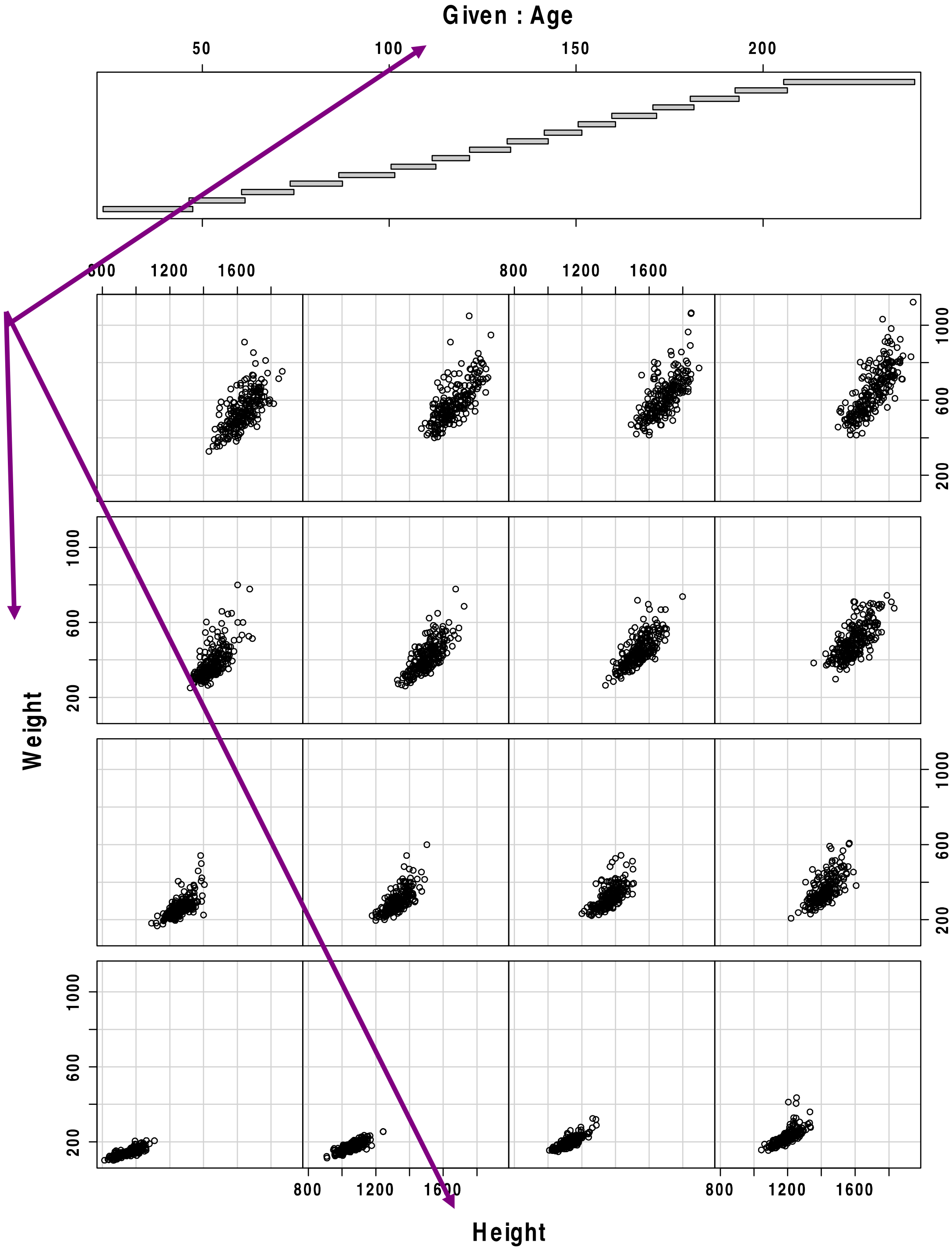


Grafico Contour



Tres variables

Gráfico "CoPlot"



CoPlot De Temperaturas

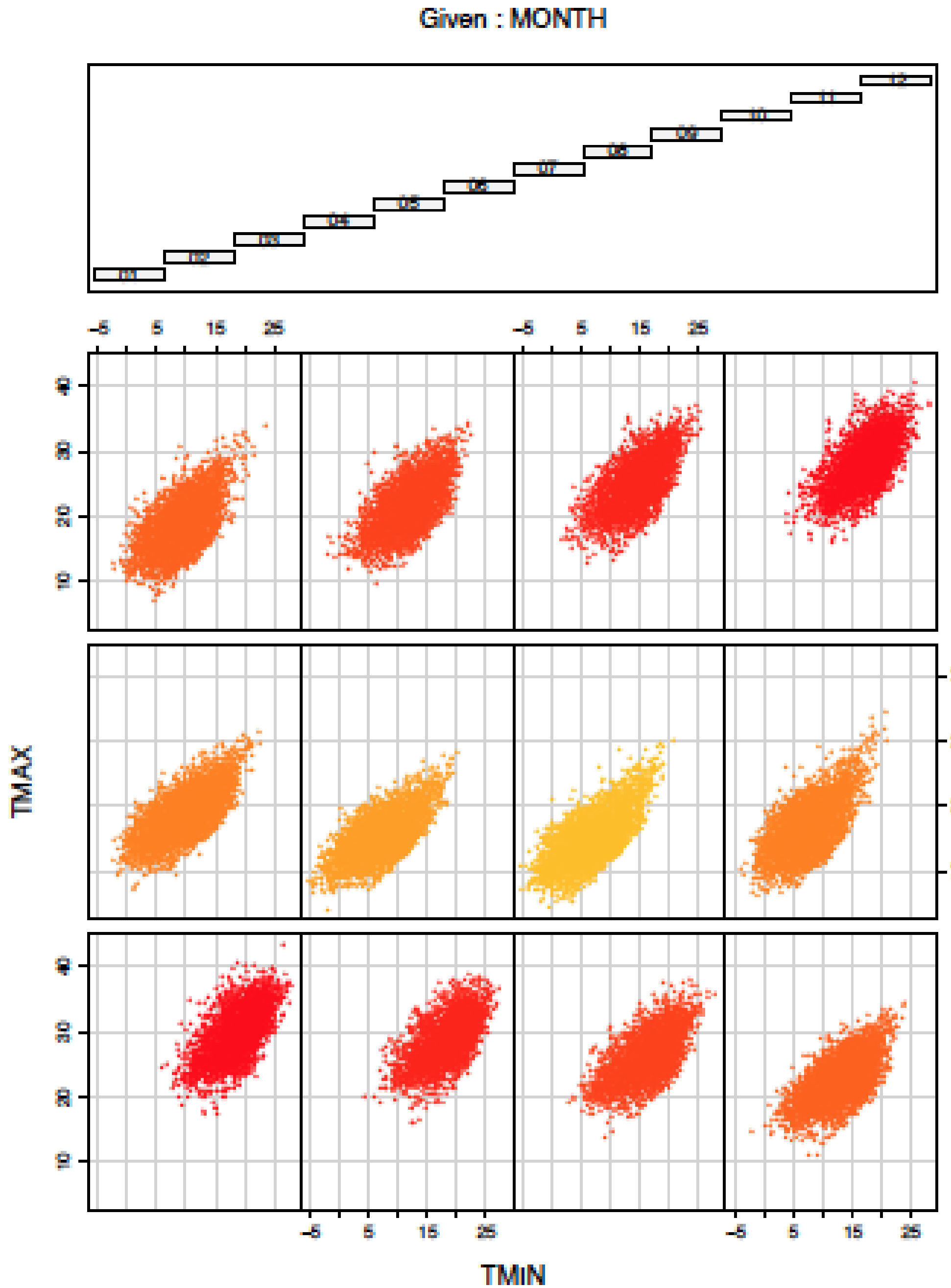


Gráfico de dispersión

Muchos
puntos
superpuestos

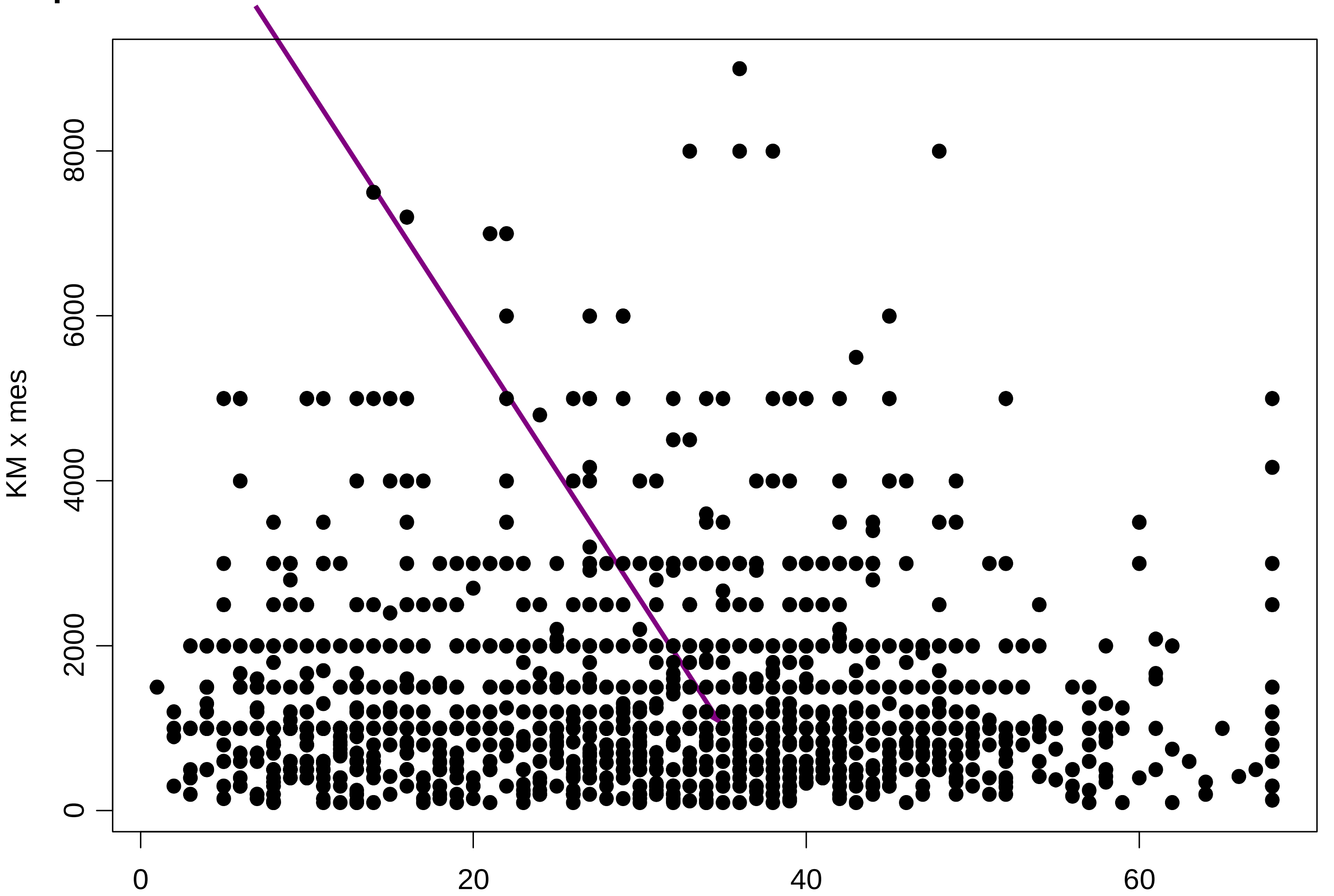


Gráfico hexbin

Mayor
concentración

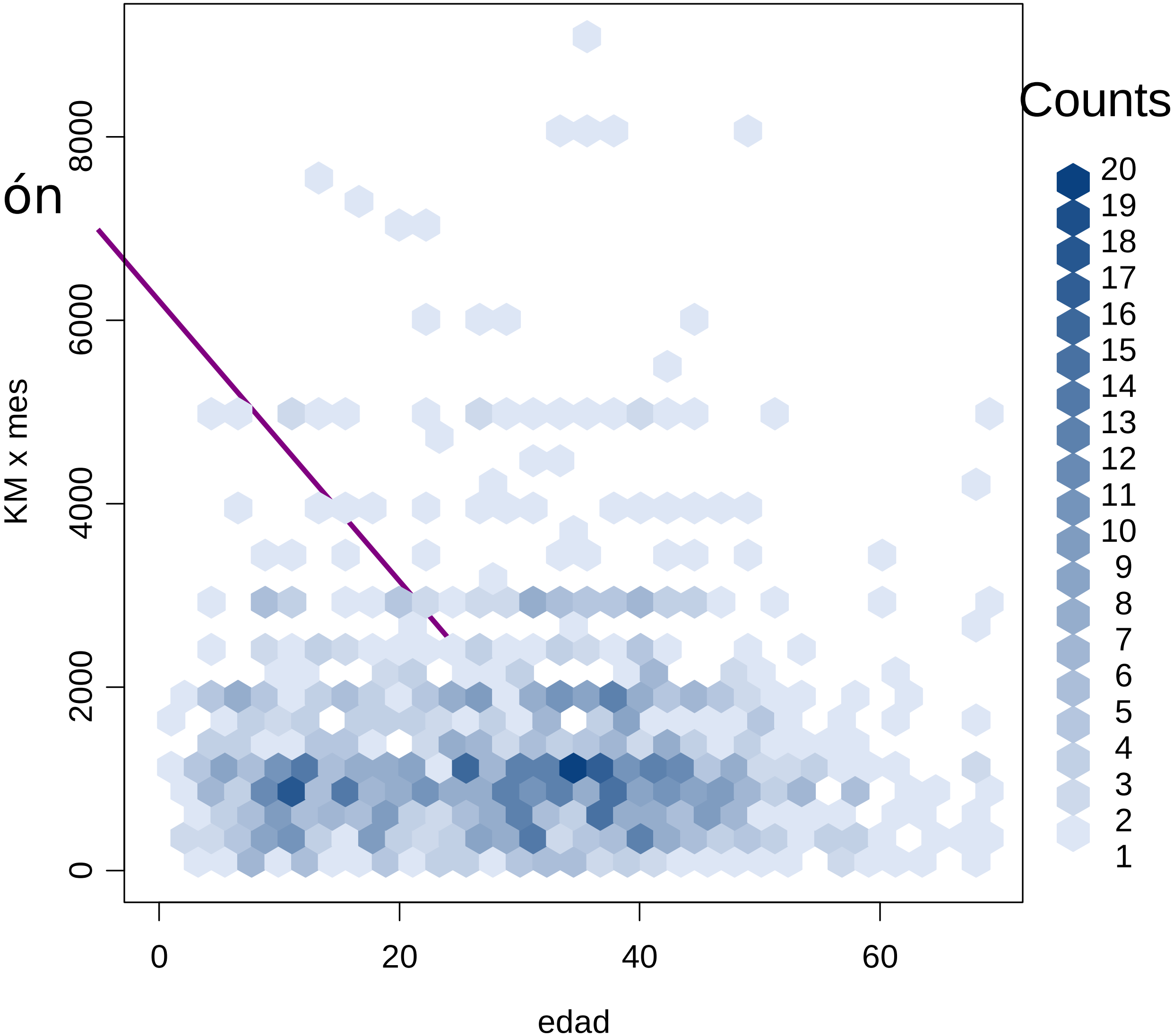


Gráfico hexbin

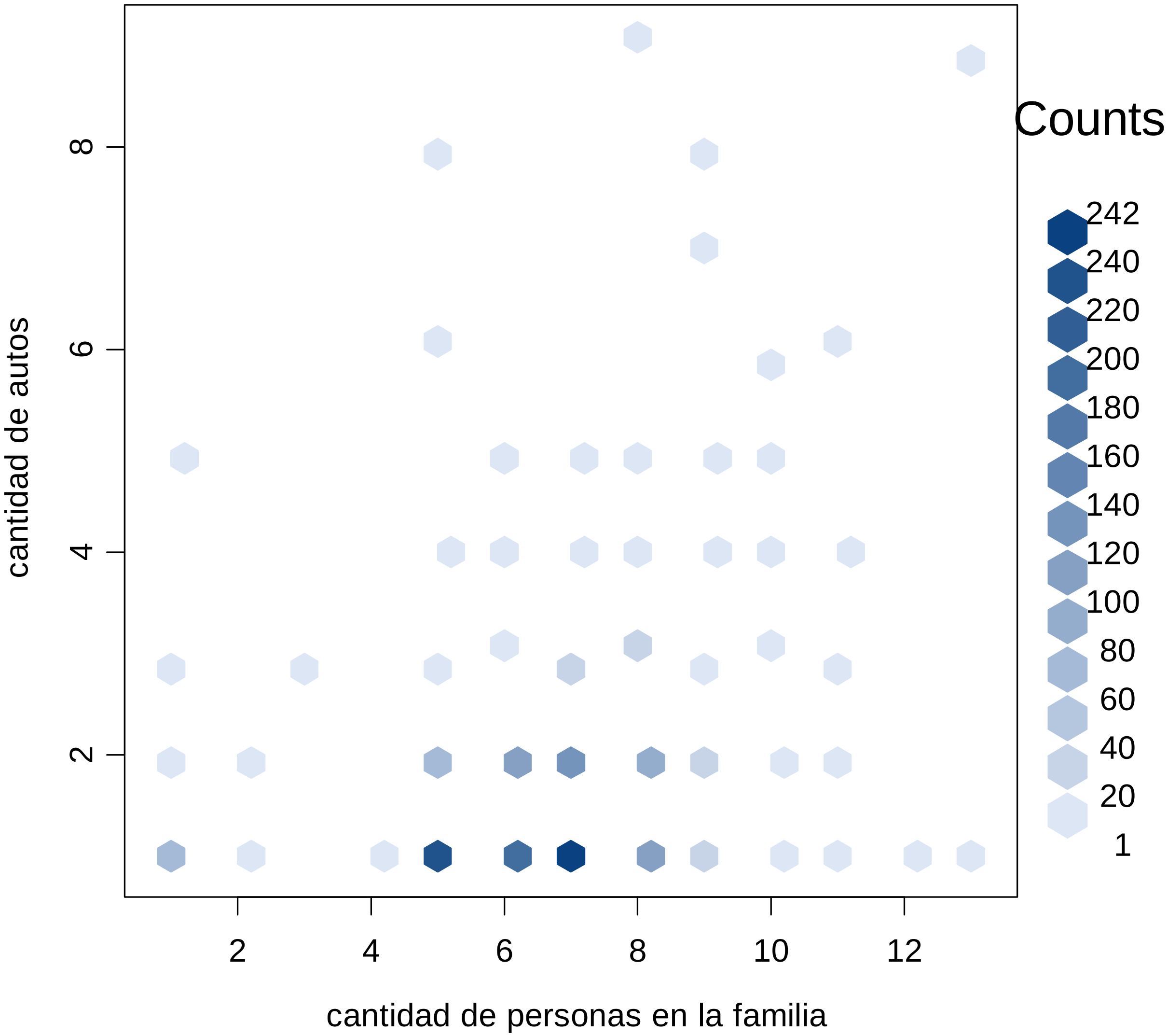
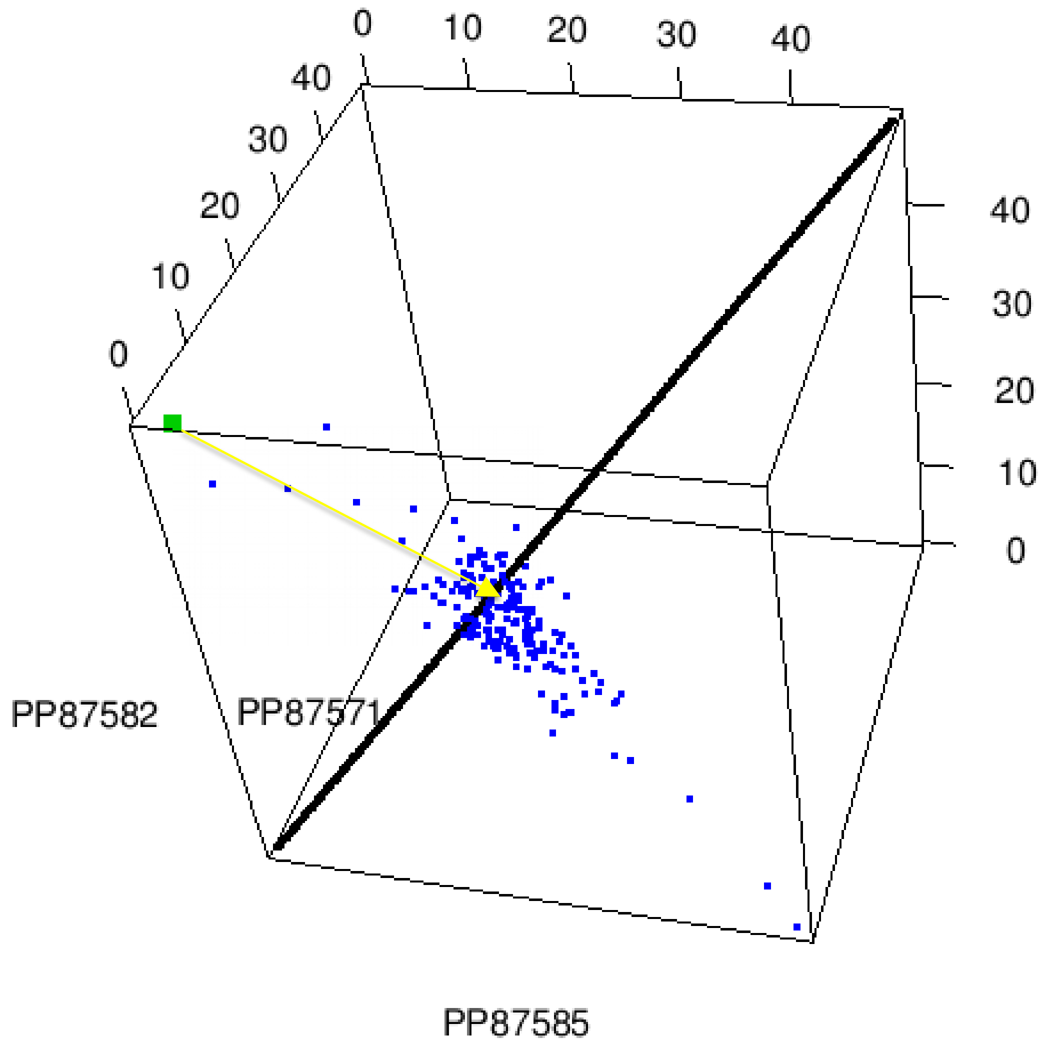
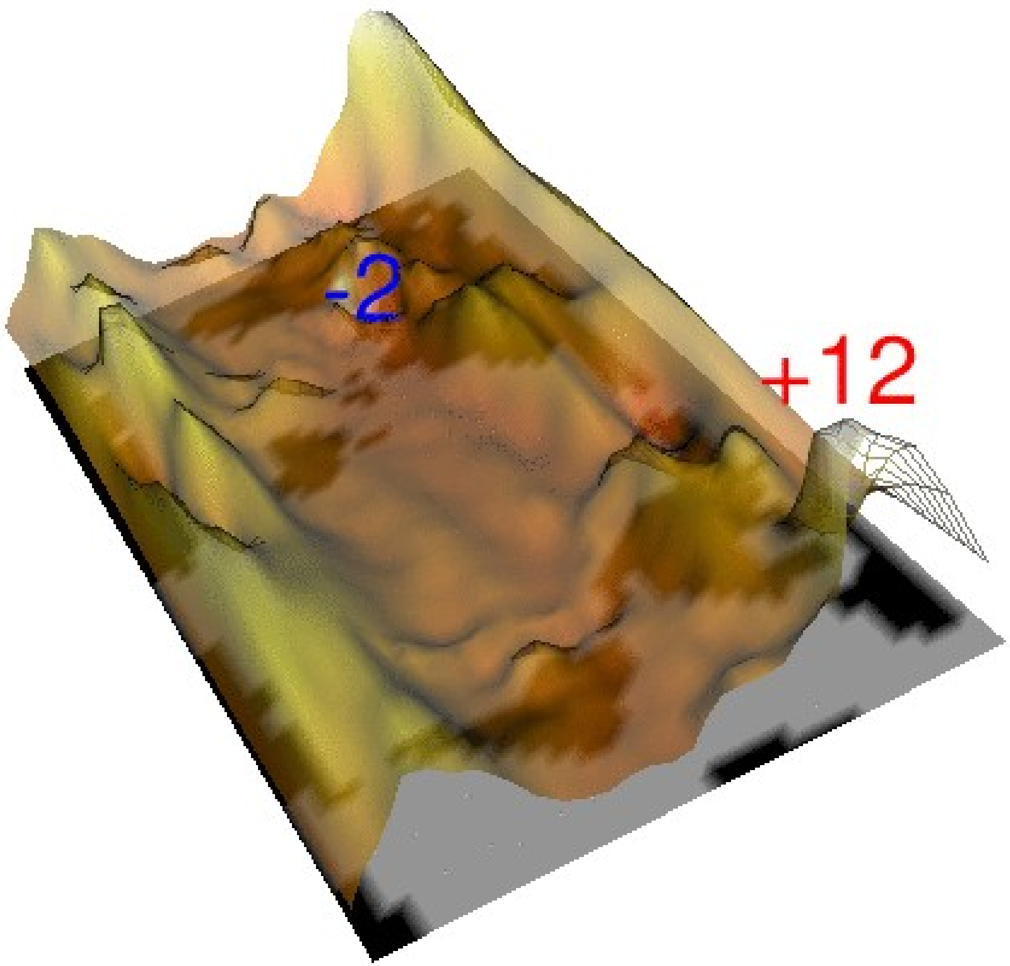


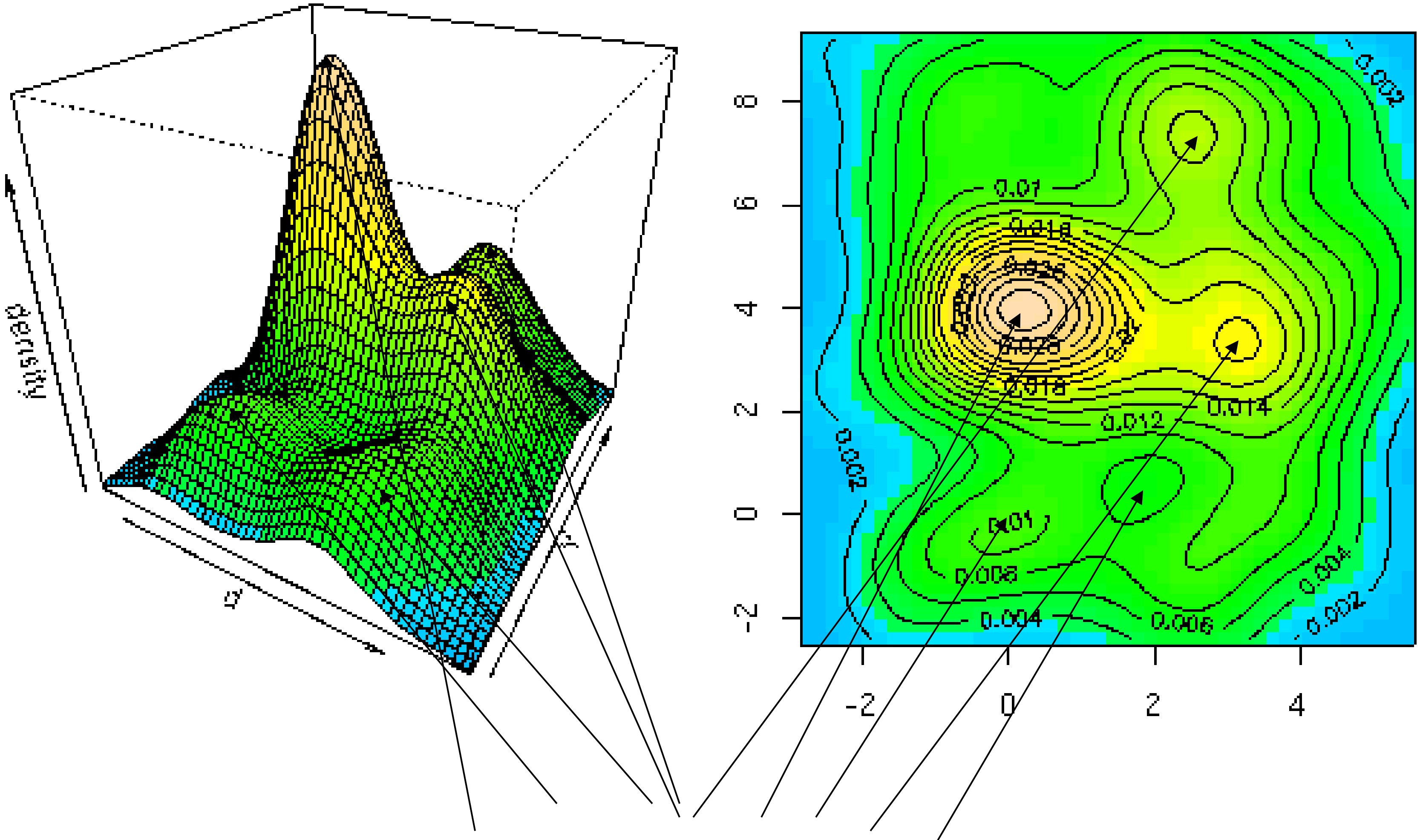
Gráfico de dispersión (X,Y,Z)



Campo medio de diferencias entre R4 y control para el periodo 2046-2050



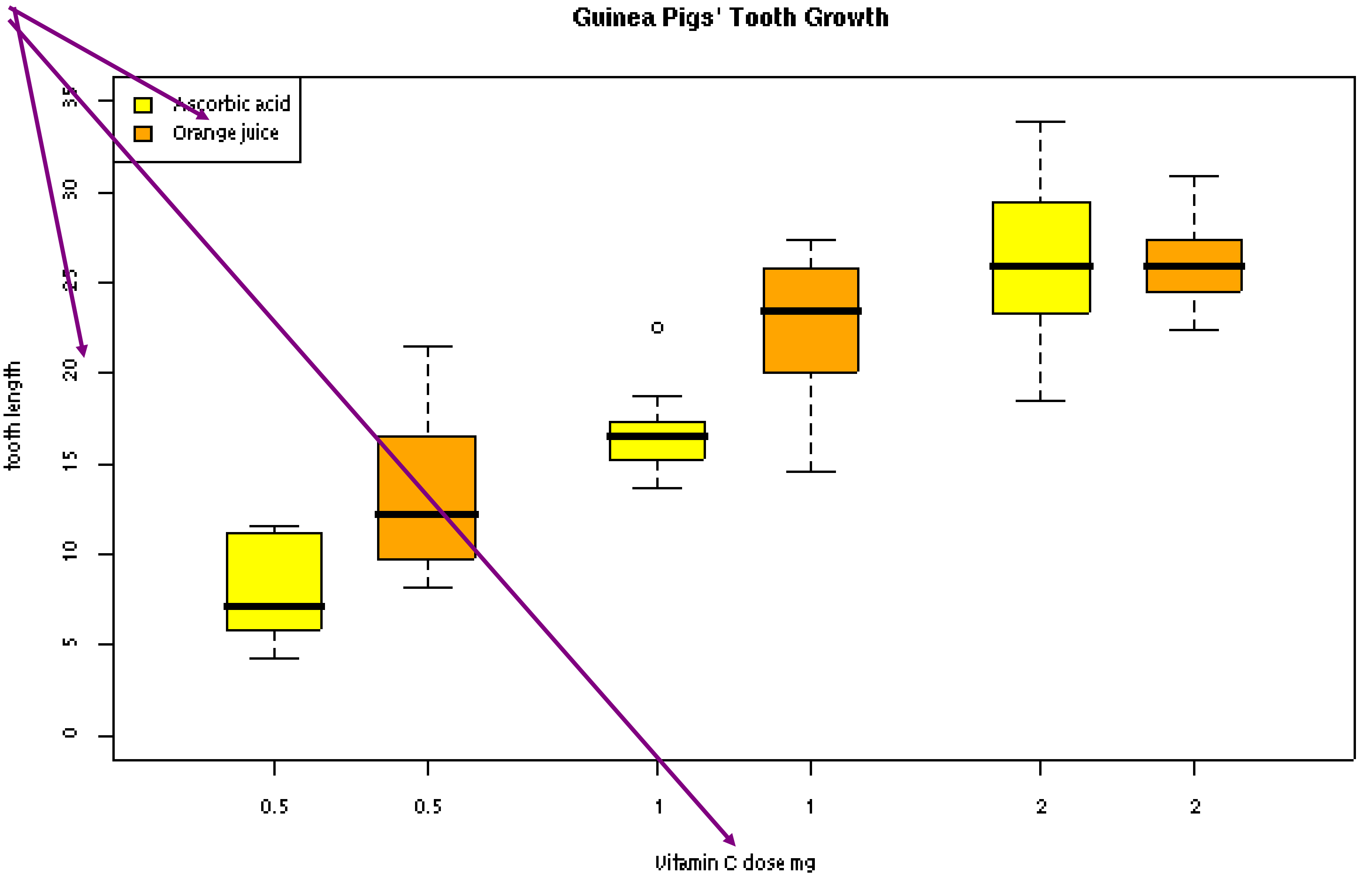
Estimación de densidad por núcleos (Bivariado)



Posibles clusters

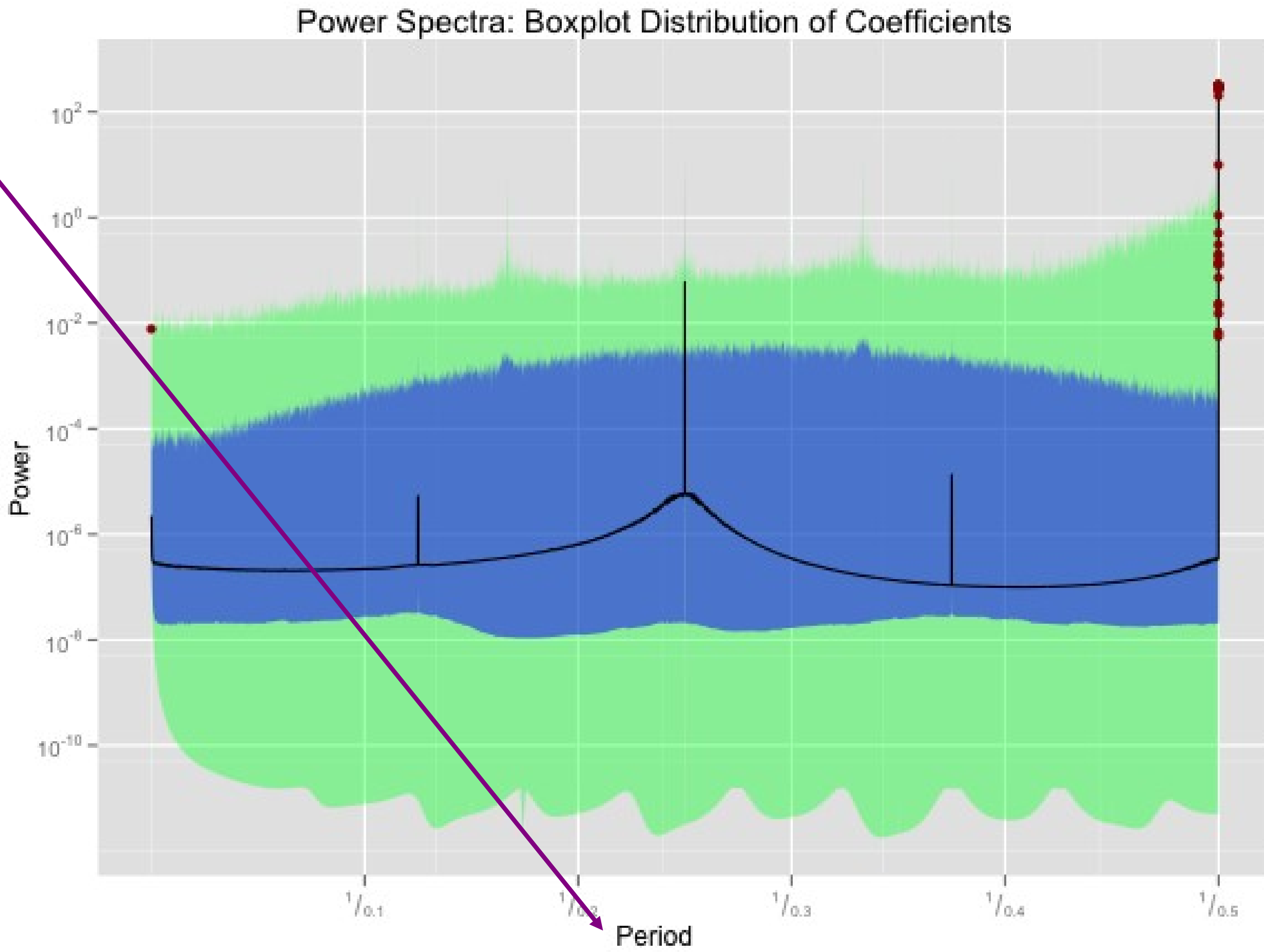
Múltiples Boxplots

Tres variables



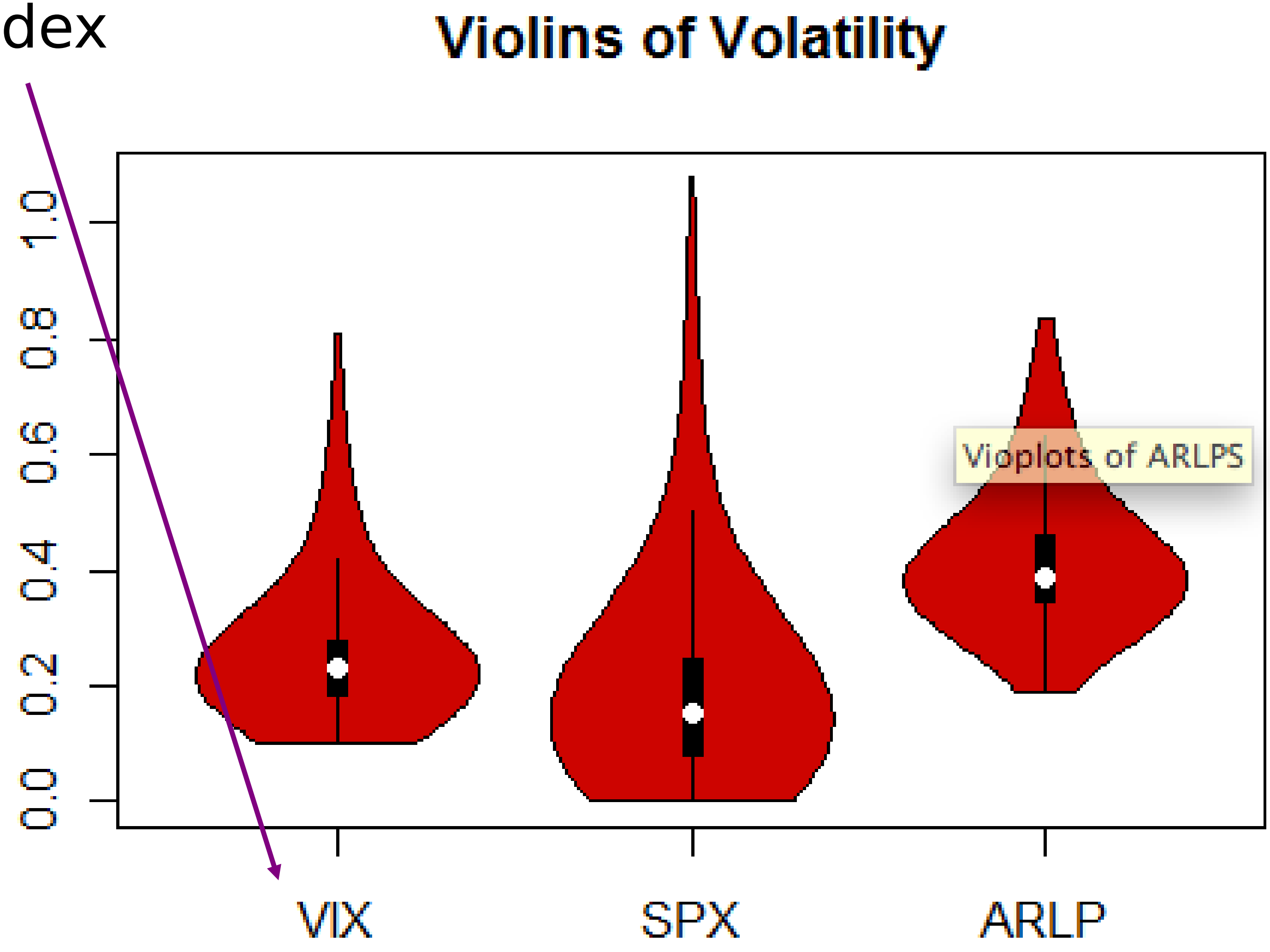
Boxplot condicionaado a una variable continua

Dos variables

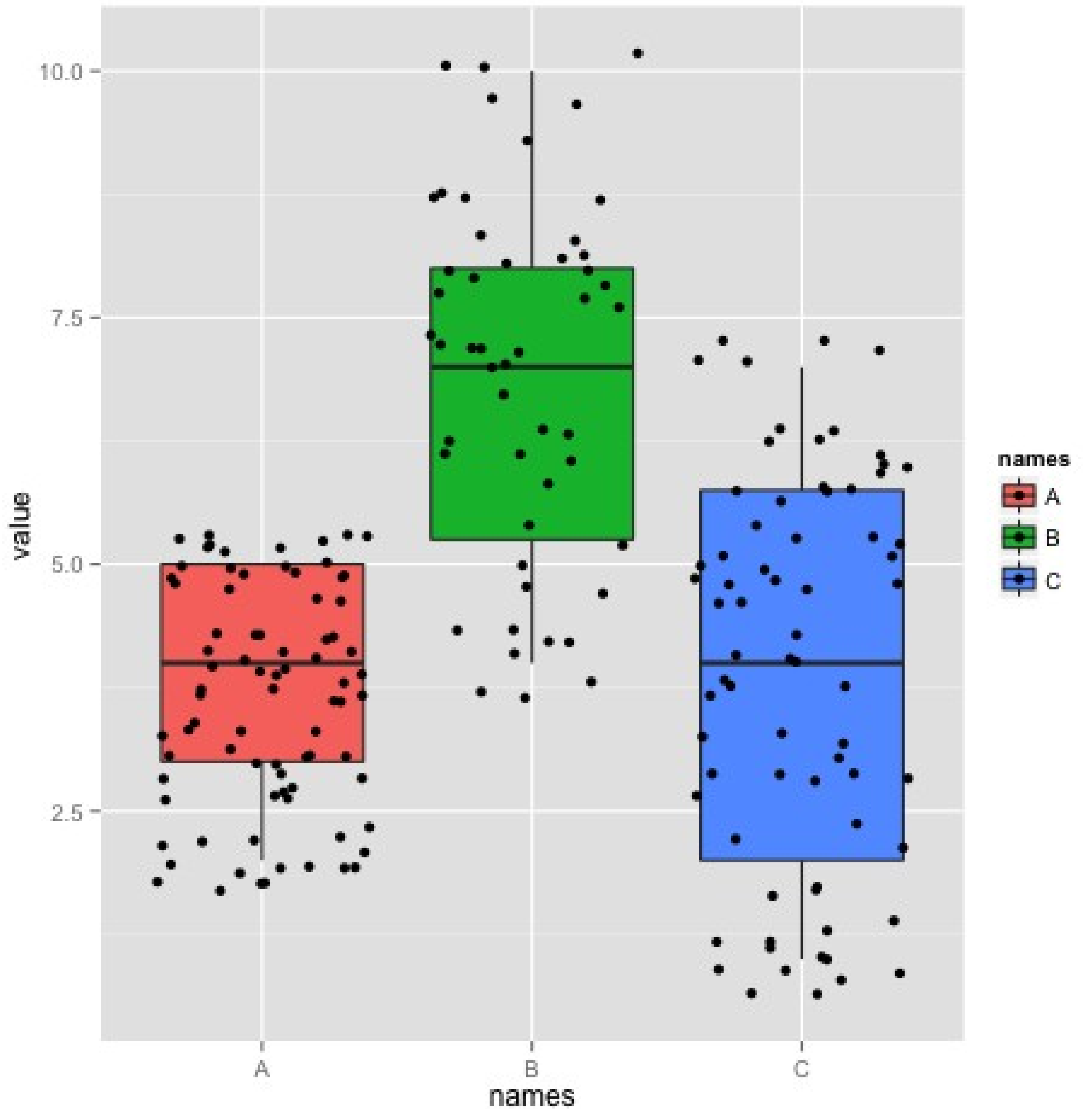


Gráficos de Violin

Stock index



Boxplot + Scatterplot



Grafos

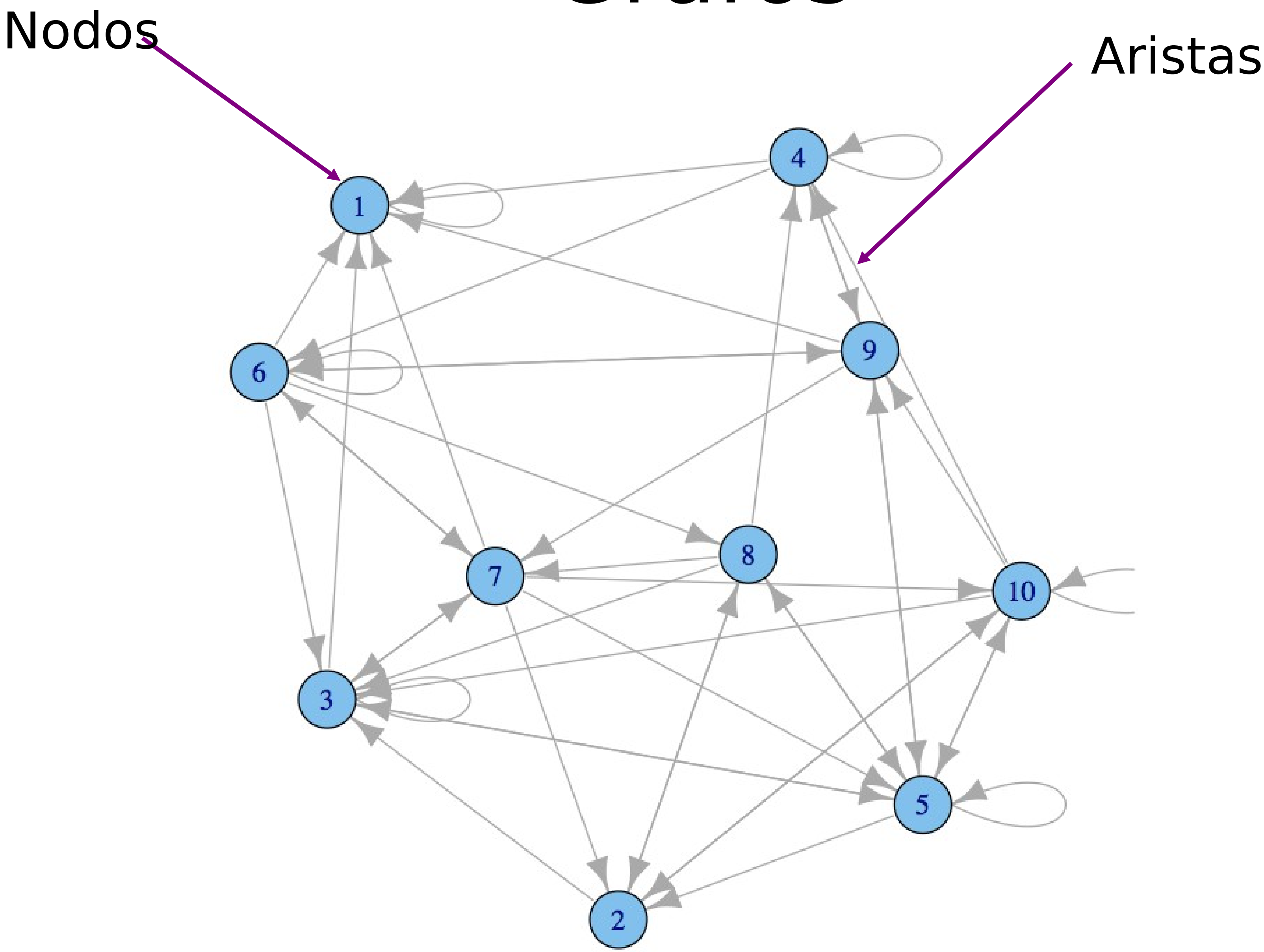
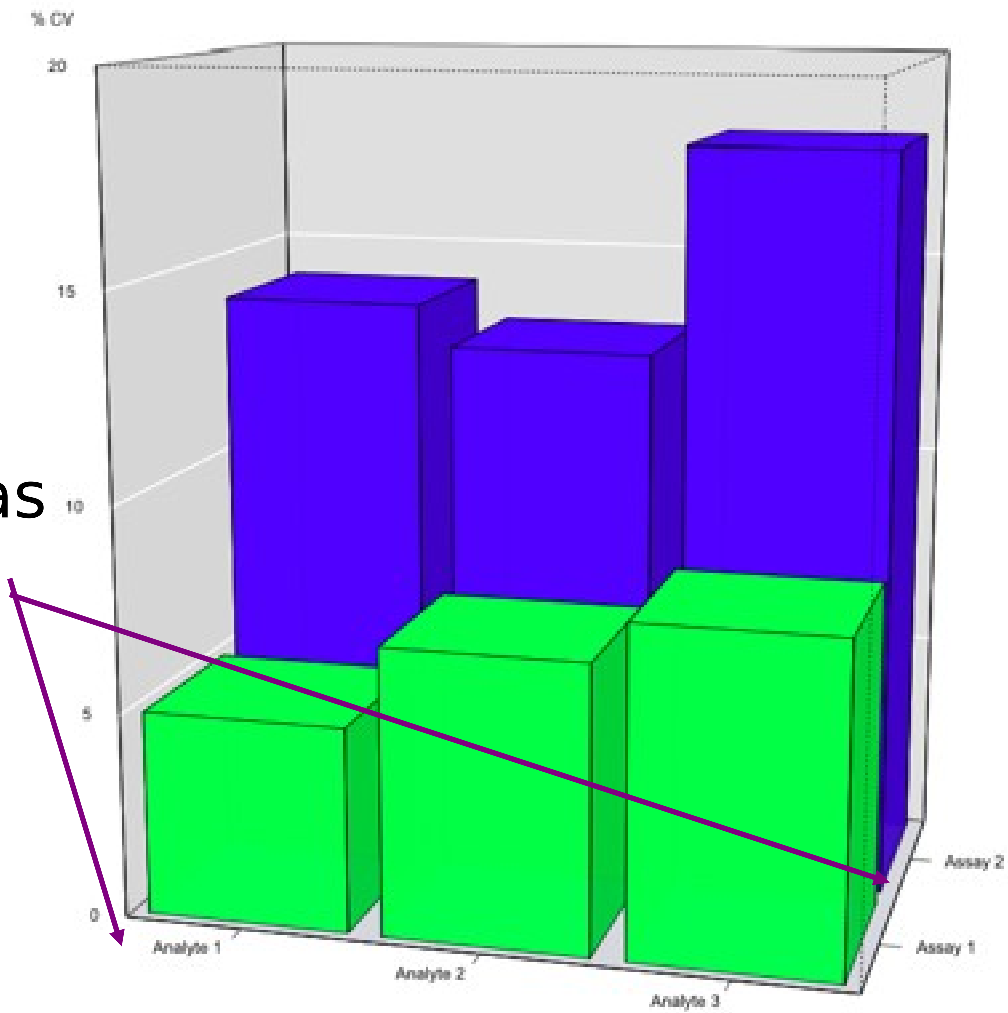


Gráfico de barras 3D

Dos variables categóricas



Gráficos de elipses

Correlaciones

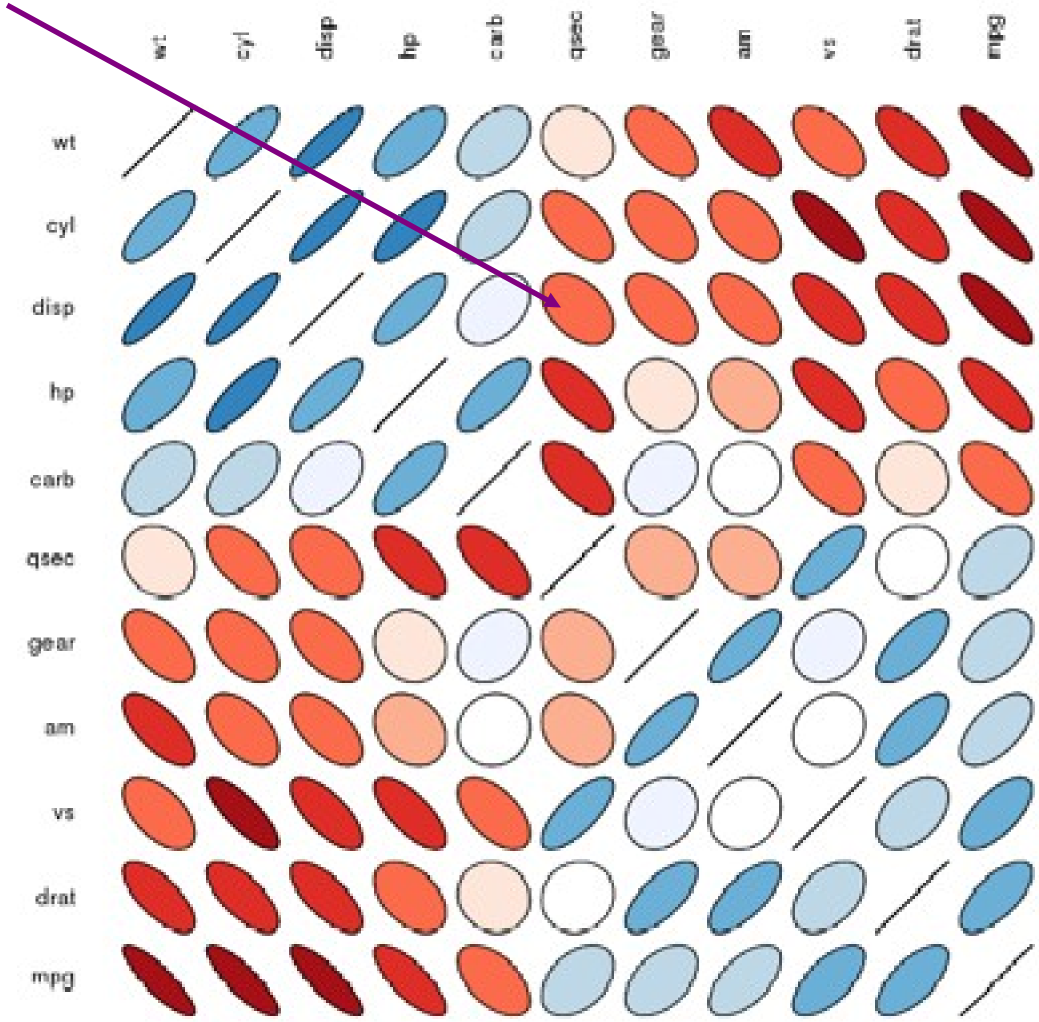
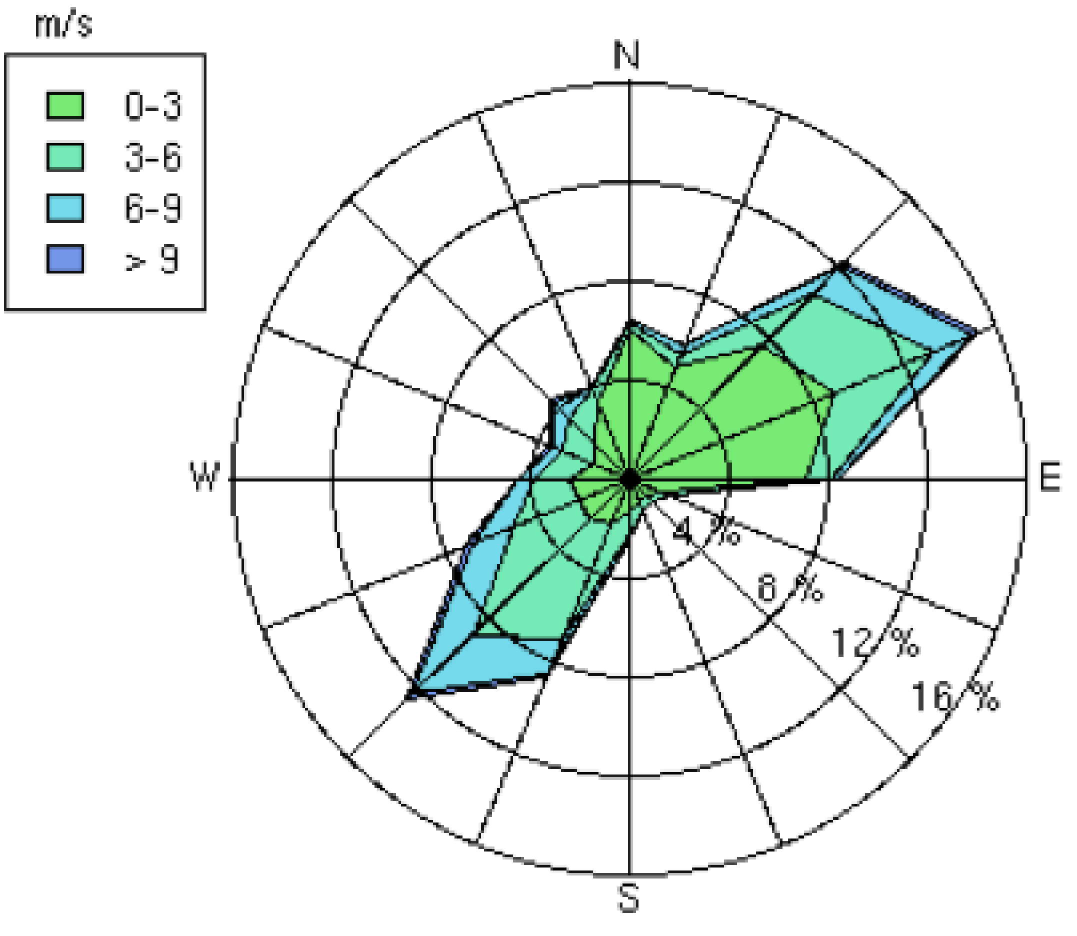
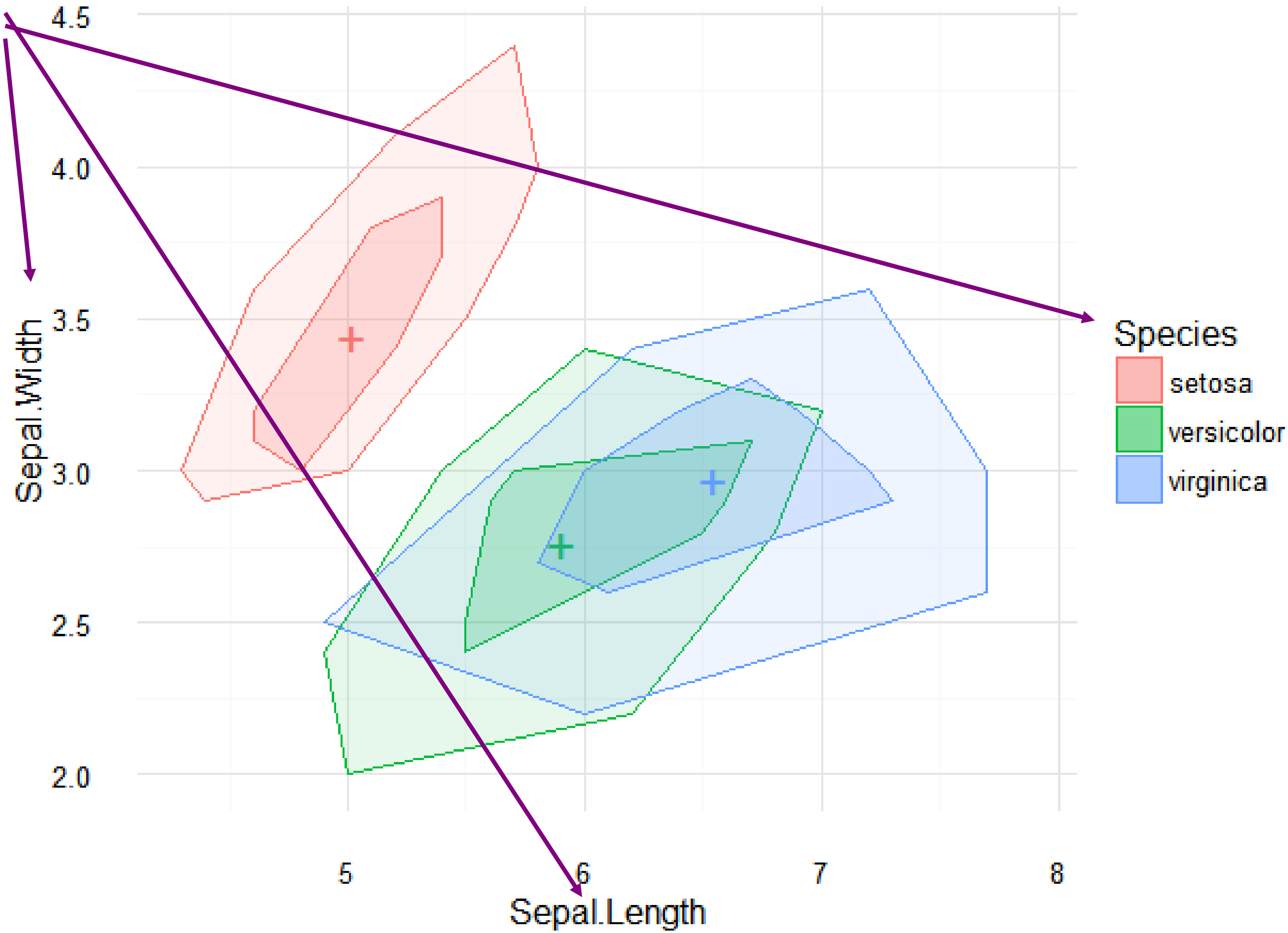


Gráfico de densidad para datos angulares (Viento)



Bagplot (2D Boxplot)

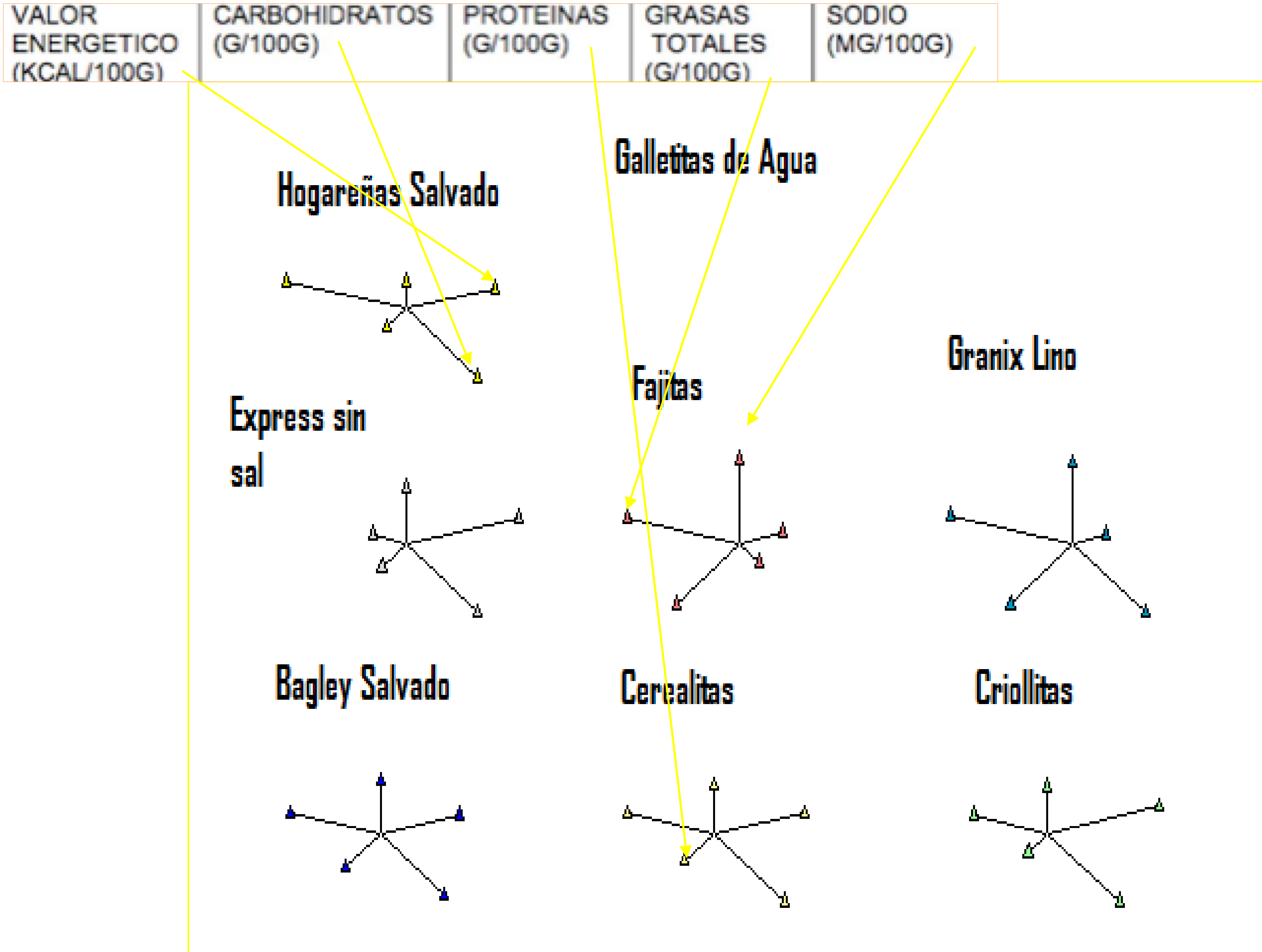
Tres variables



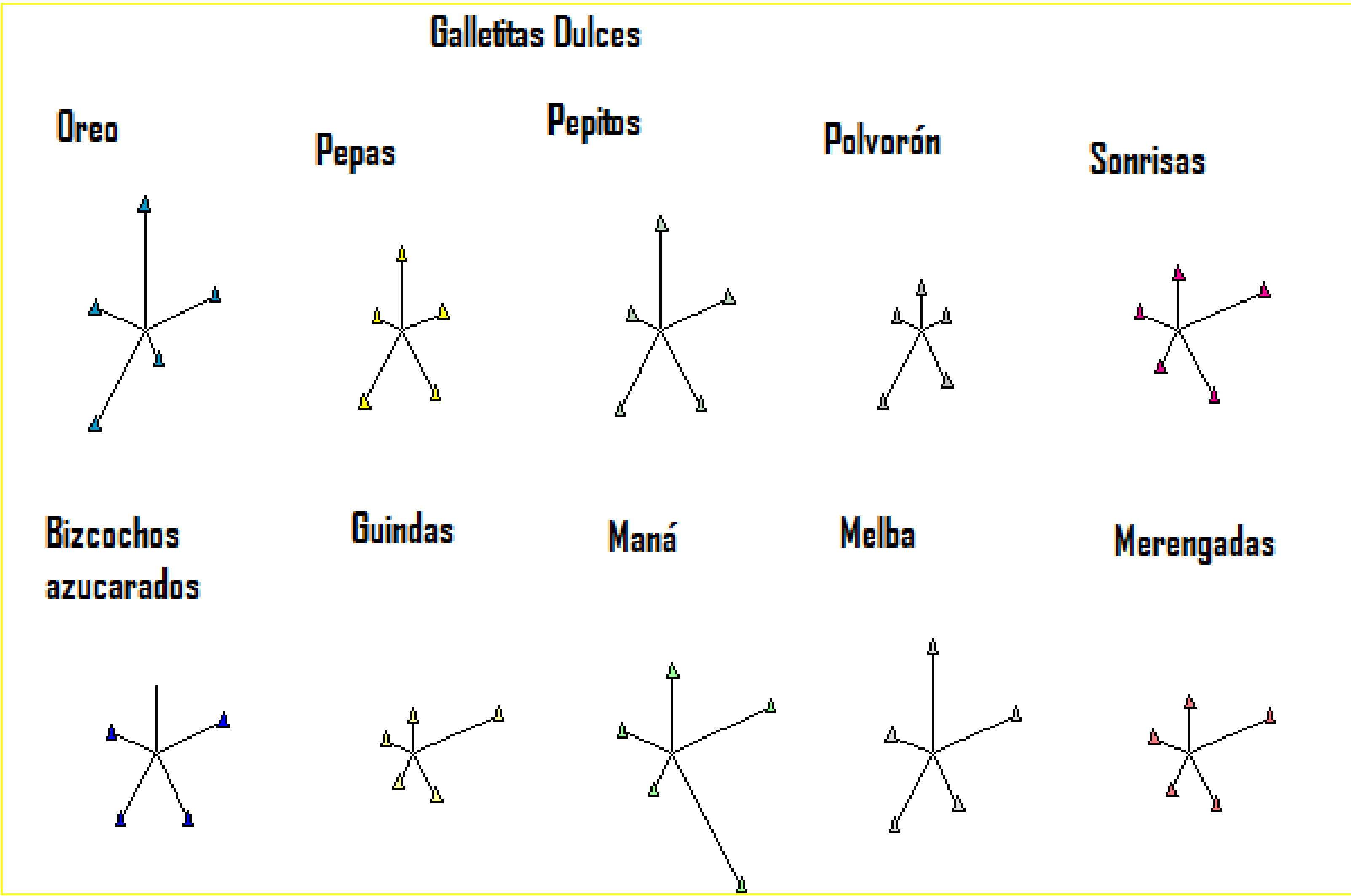
Datos de galletitas

MARCA	VALOR ENERGETICO (KCAL/100G)	CARBOHIDRATOS (G/100G)	PROTEINAS (G/100G)	GRASAS TOTALES (G/100G)	SODIO (MG/100G)
cerealitas	439	65	11	15	574
fajitas	466	57	10	22	828
express s/sal	445	69	11	14	12
oreo	478	67	5,6	21	363
melba	464	70	6,3	18	263
pepitos	463	66	7,1	19	136
criollitas	438	69	11	13	431
merengadas	418	69	6,3	13	201
sonrisas	423	70	6,8	13	241
maná	444	73	9	13	375
guinditas	407	70	6	12	106,7
pepas	437	60	6,7	18	76,67
Polvorón	410	56,7	6,3	18	66,7
bizcoch.grasa.azuc	493	60	7,6	24	1066
hogareñas.salvado	424	65	11	13	892
granix.con.lino	462	55	11	22	931
bagley salvado	421	63	11	14	624

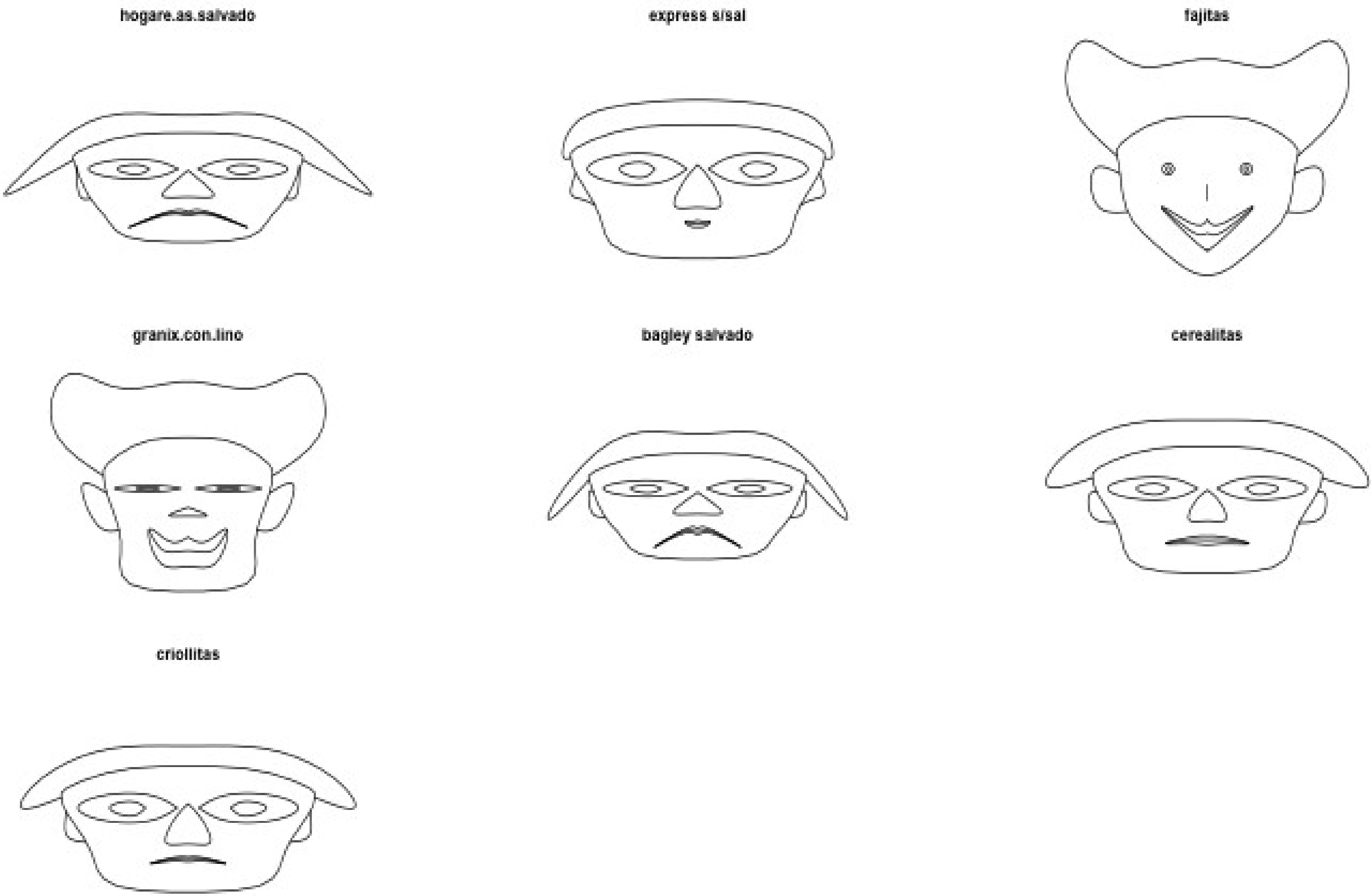
Gráficos de estrellas (1)



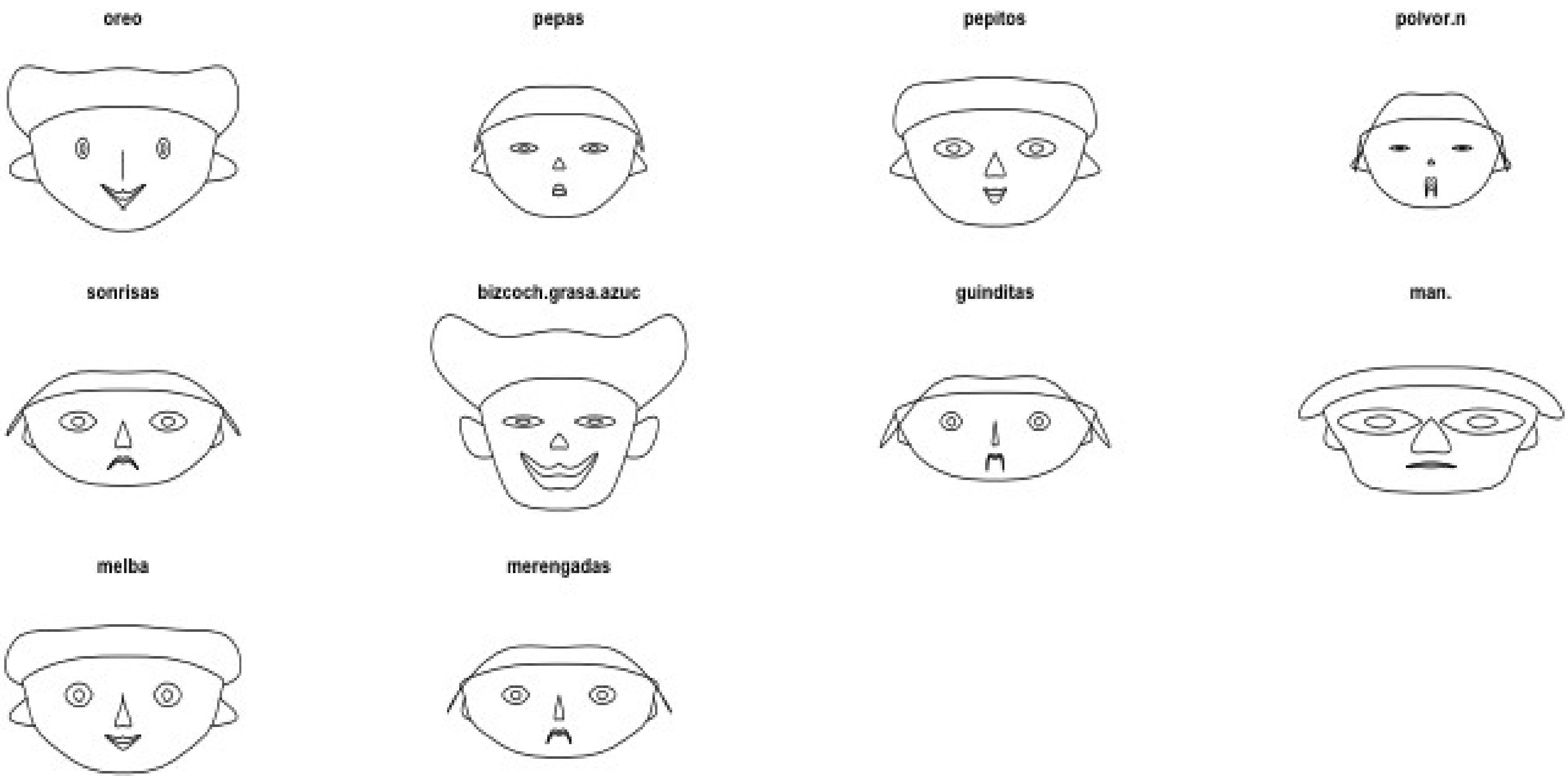
Gráficos de estrellas (2)



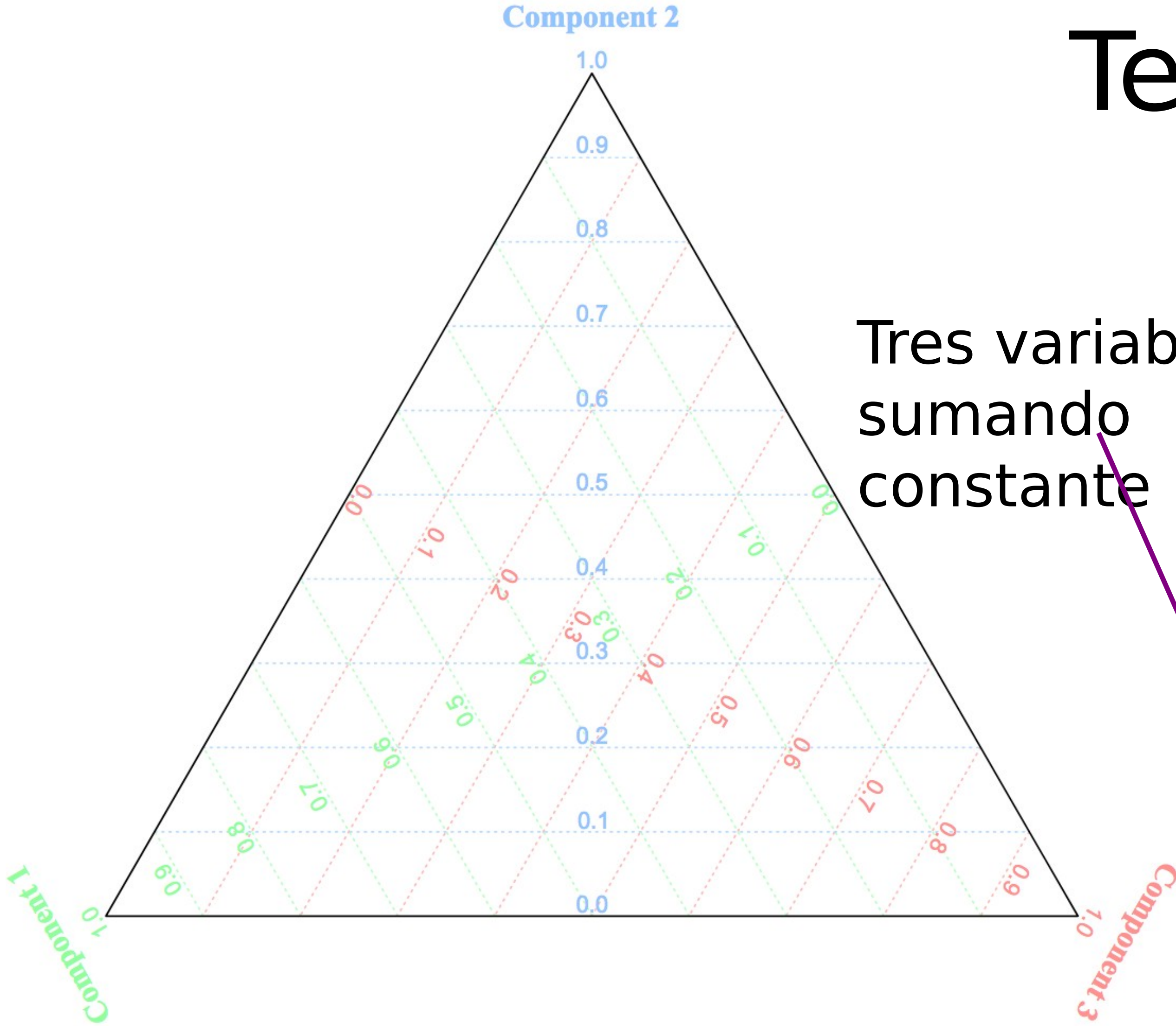
Gráficos de caras (1)



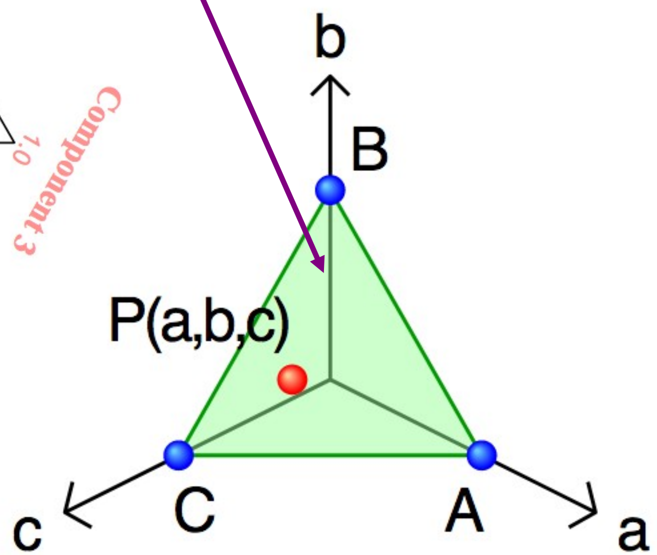
Gráficos de caras (2)



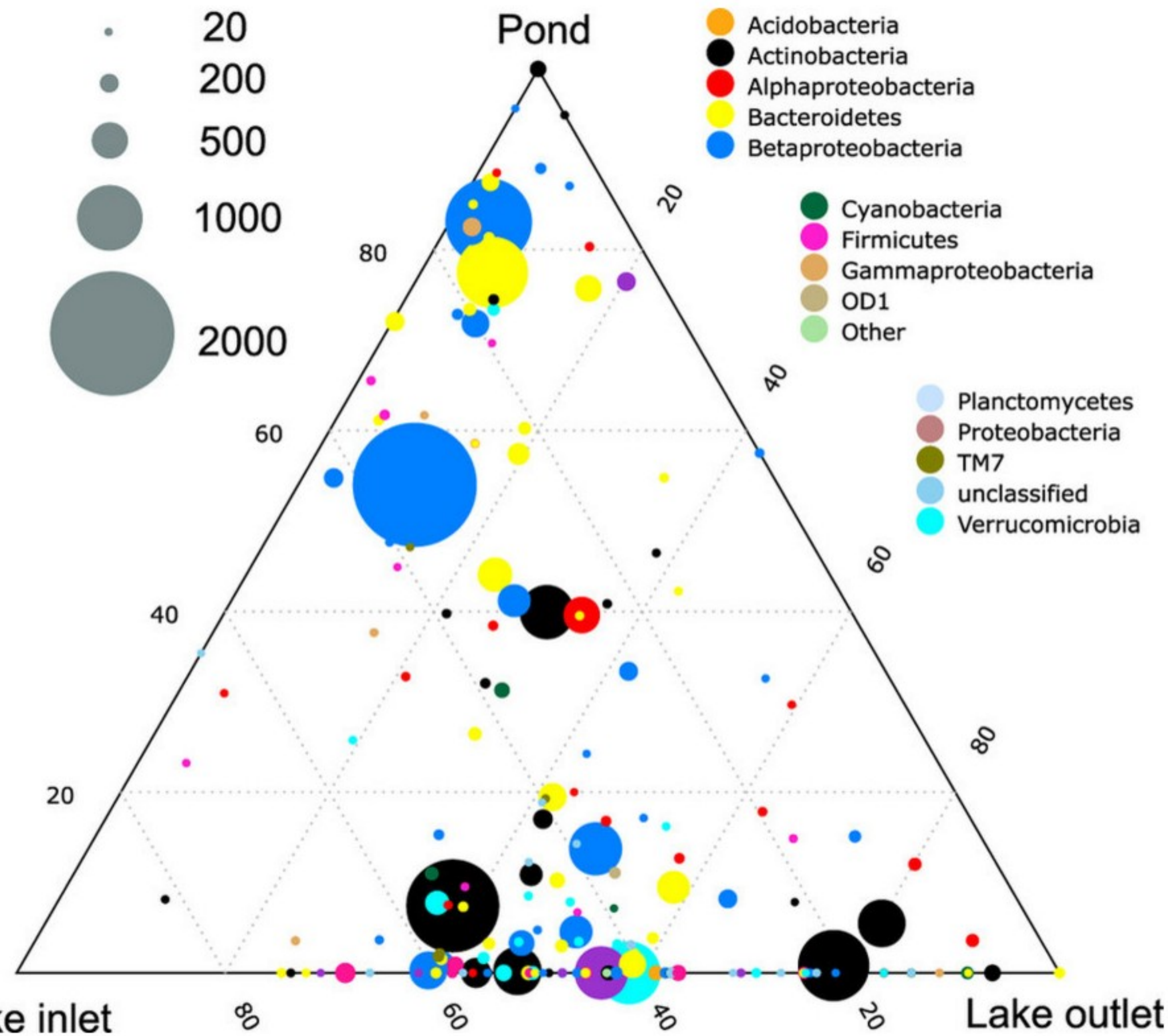
Ternary Plot



Tres variables sumando constante



Presencia de bacterias en 3 habitats



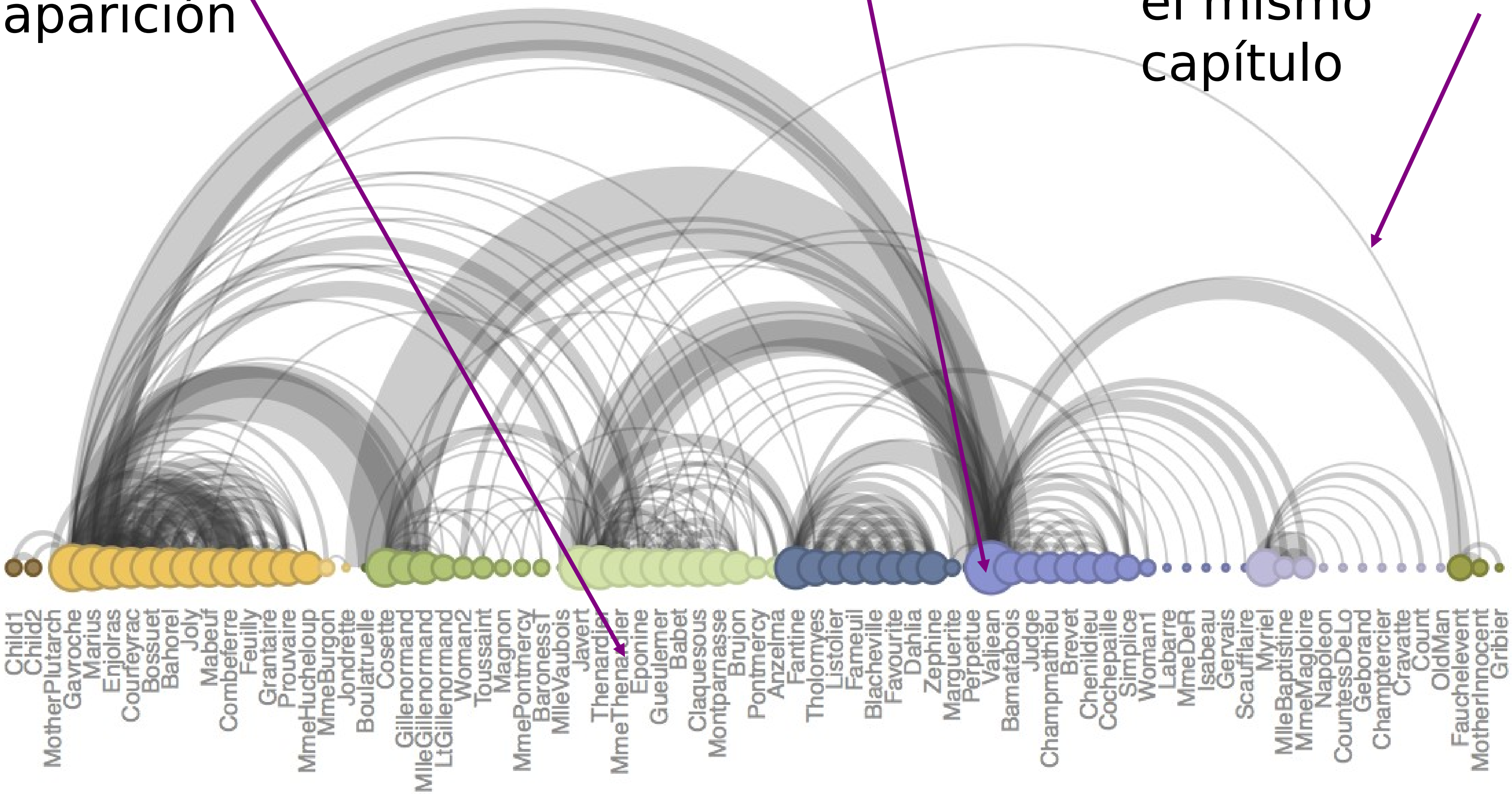
Axes represent the pond, inlet and outlet and the percentage of reads associated with each environment. The size of the symbol indicates number of reads associated with each OTU and taxonomic affiliations are indicated by colors. All OTUs with at least 20 reads were included into the plot.

Diagramas de Arco
(Los Miserables, Victor Hugo)

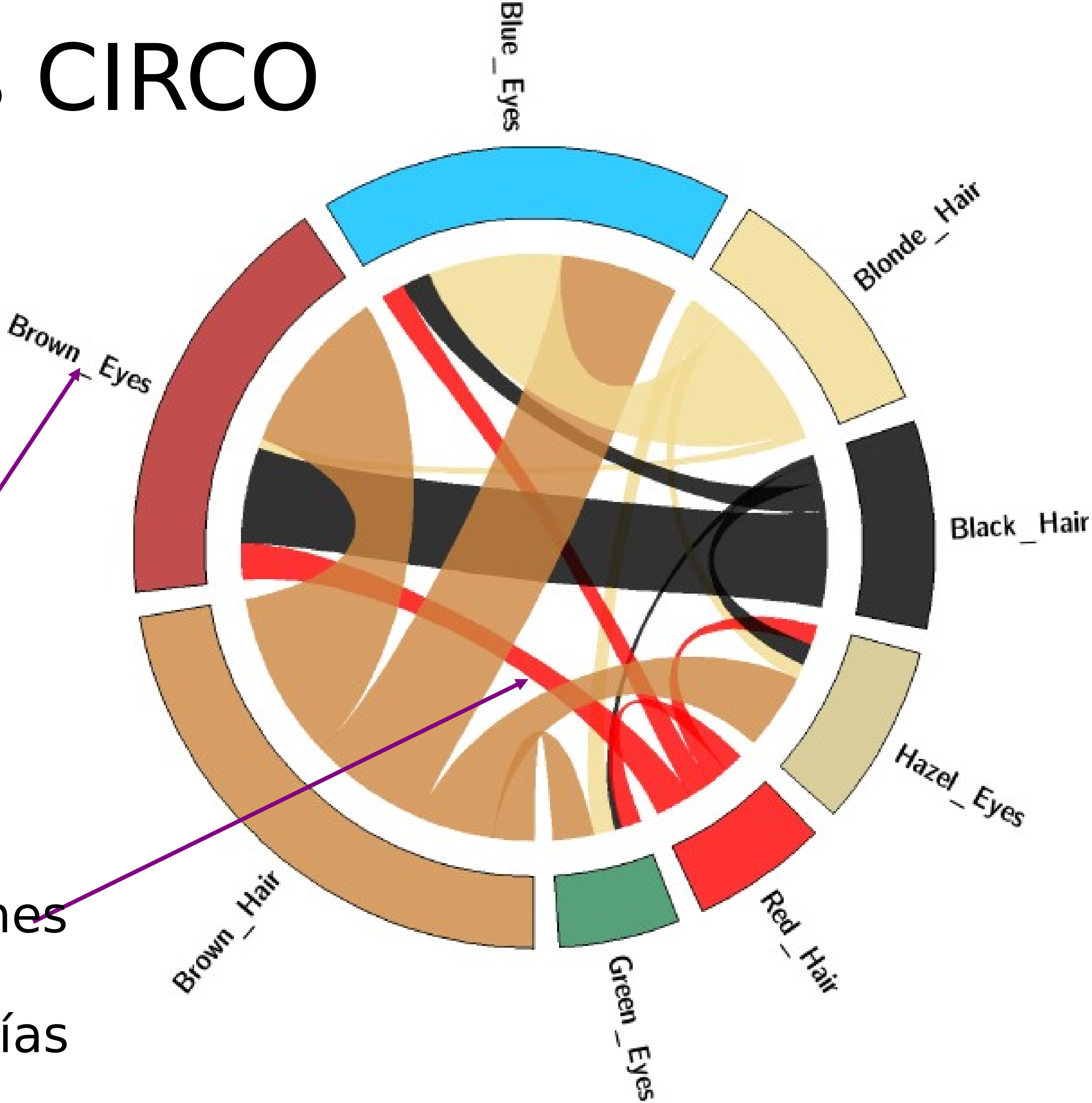
Protagonistas en el libro por orden de aparición

Capítulos del libro

Conexión por aparición en el mismo capítulo



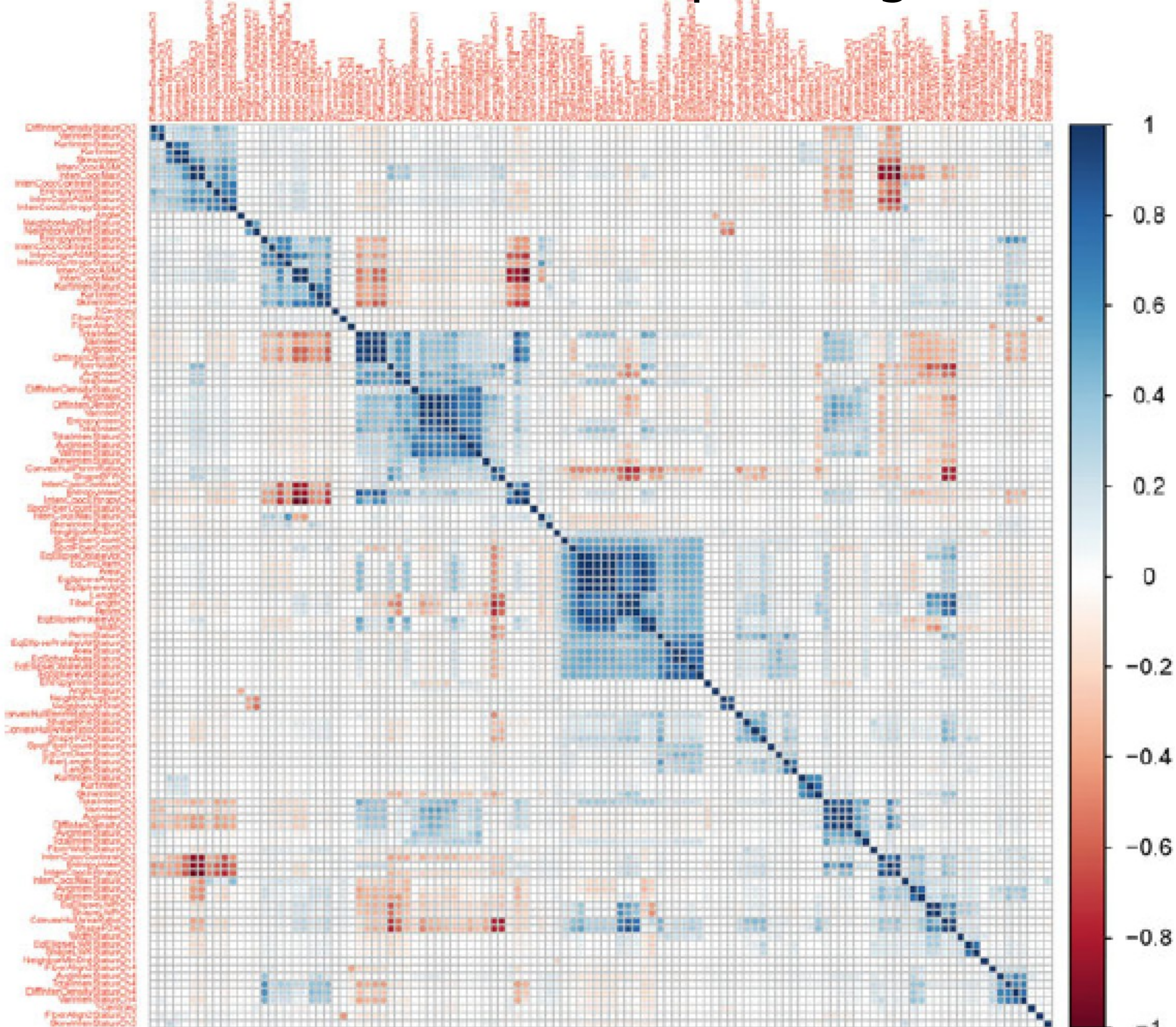
Gráficos CIRCO



Categorías

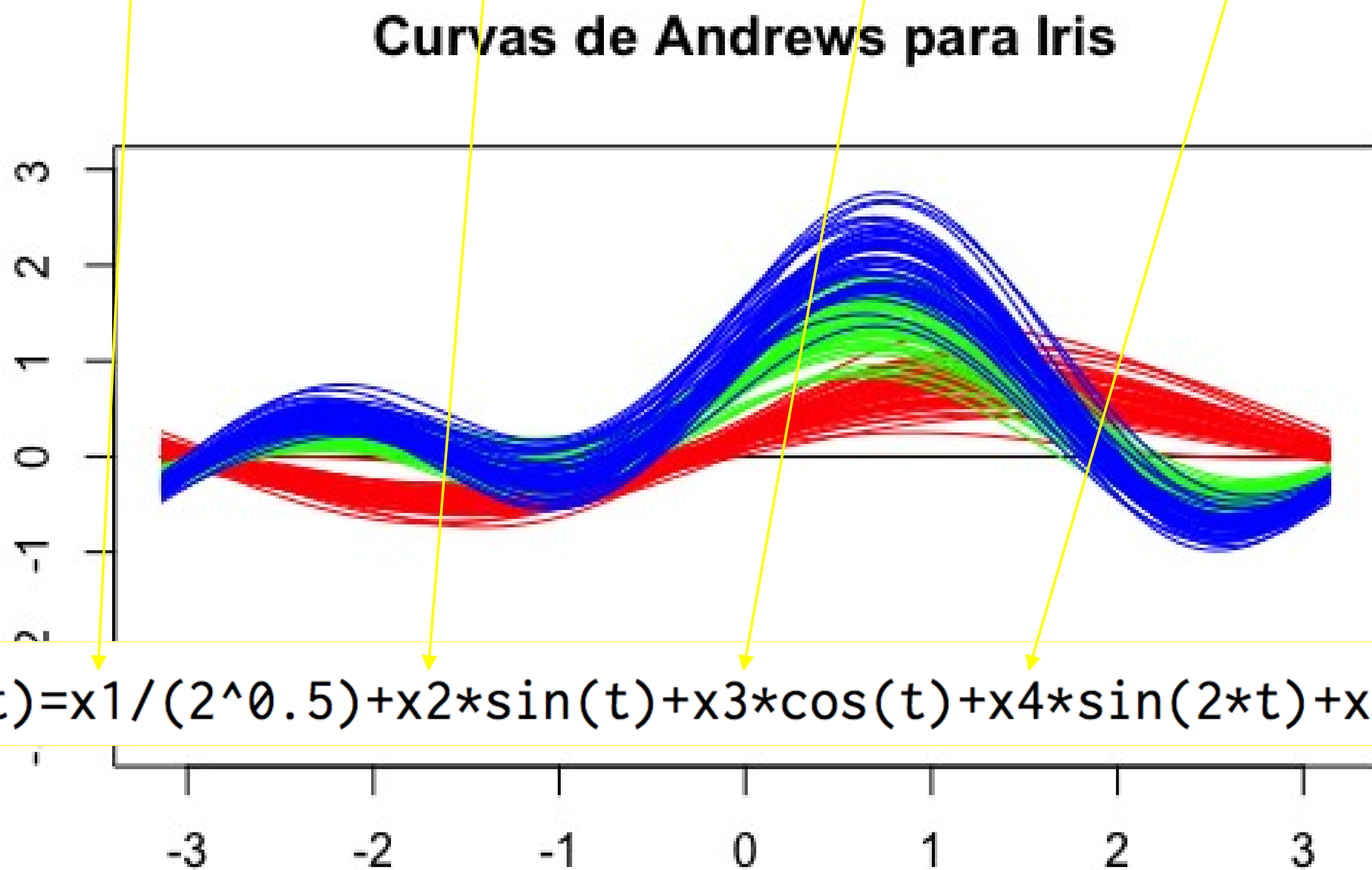
Relaciones entre Categorías

Correlaciones Ordenadas por Segmentación



Curvas de Andrews

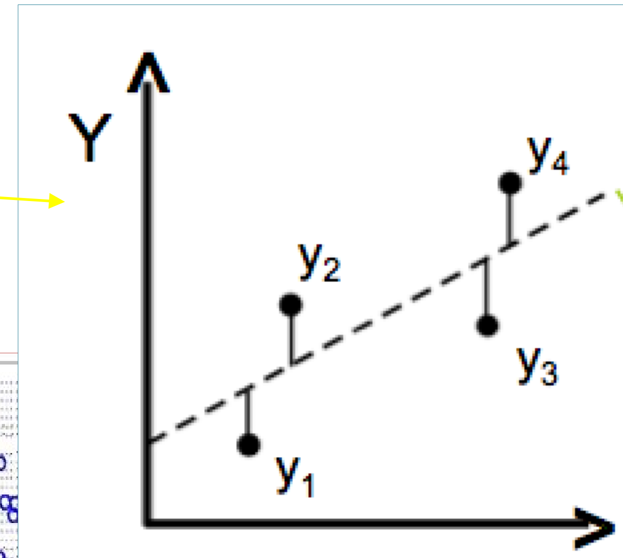
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
5.1	3.8	1.5	0.3	setosa
5.6	2.5	3.9	1.1	versicolor
				versicolor
				virginica
				virginica

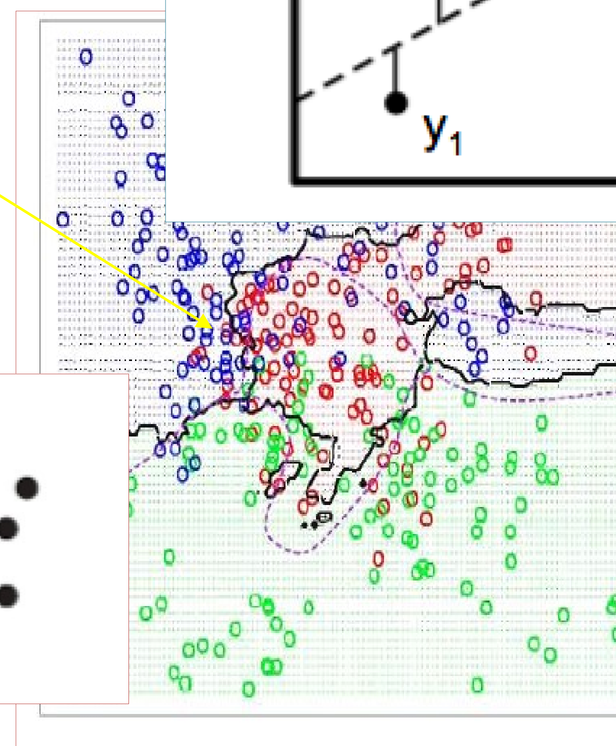


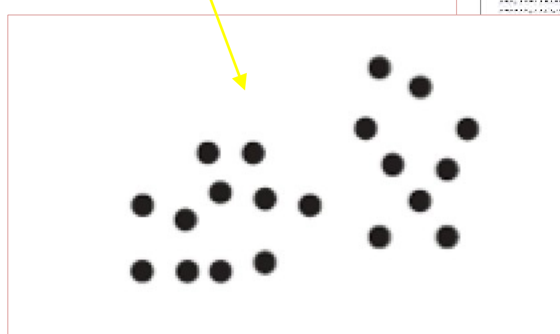
f(t)=x1/(2^0.5)+x2*sin(t)+x3*cos(t)+x4*sin(2*t)+x5*cos(2*t)+...

Benchmarking

- Como medimos la efectividad de los métodos ?
- En regresión ?
- En clasificación ?
- En segmentación ?
- Que métricas utilizamos ?
- Que técnicas empleamo ?







Benchmarking en Regresión

$$\sum (y_i - \widehat{y}_i)^2$$

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Proportional Mean Absolute Error (PMAE)
- Coeficiente de Dterminación (R2)
- Correlación – Pearson - Spearman

$$\frac{1}{n}\sum_{i=1}^{n} |e_i|$$

$$R^2 = \frac{SCR_{eg}}{SCT} = 1 - \frac{SCE}{SCT} \quad ; \quad 0 \leq R^2 \leq 1$$

$$\sum (\widehat{y}_i - \bar{y})^2 \qquad \sum (y_i - \bar{y})^2$$

Benchmrking en Clasificación

- Accuracy

$$ACC = (TP + TN)/(P + N)$$

- Sensibilidad

$$TPR = TP/P = TP/(TP + FN)$$

- Especificidad

$$SPC = TN/N = TN/(FP + TN)$$

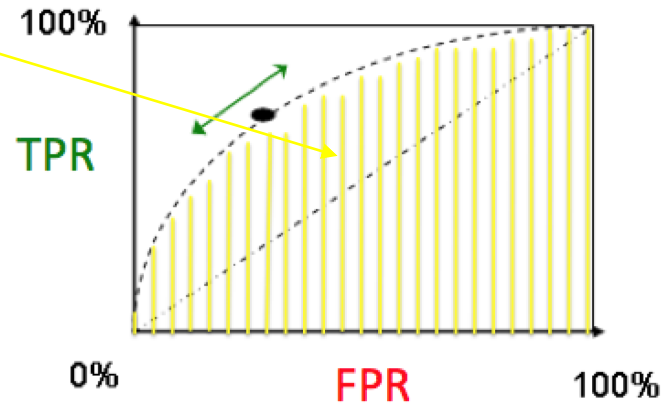
- False Discovery Rate

$$FDR = FP/(FP + TP) = 1 - PPV$$

$$PPV = TP/(TP + FP)$$

- Presición

- Area bajo la curva



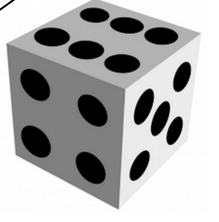
Coeficiente Kappa de Cohen

Accuracy
Obtenida

$$ACC = (TP + TN)/(P + N)$$

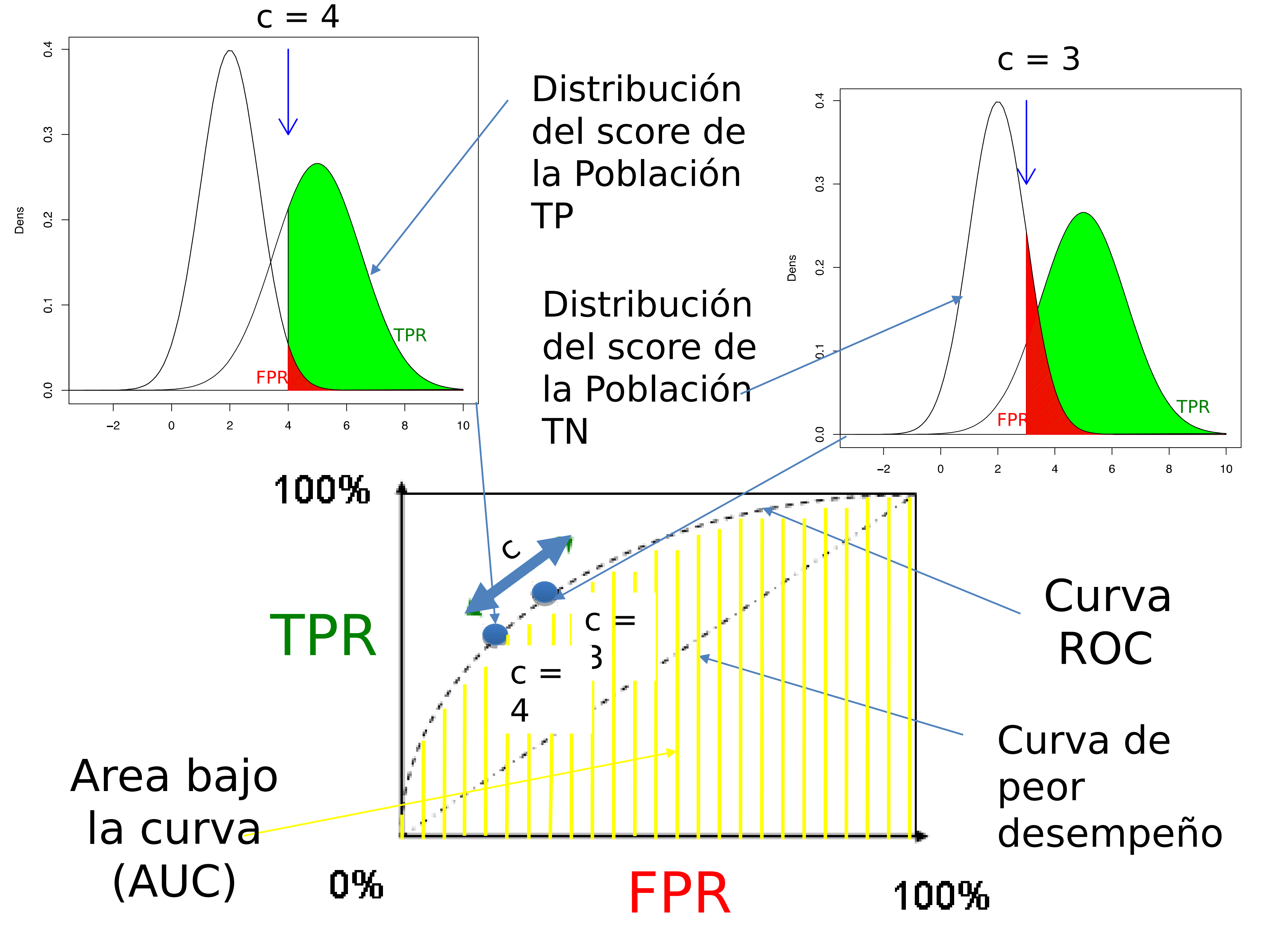
$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Accuracy que se
obtendría por
AZAR



Curvas ROC

- Método gráfico que muestra el desempeño de un procedimiento de Clasificación.
- Contempla la asimetría en el error.
- Sólo viable en casos de procedimientos con Scores continuos
- Interpretación Probabilística: Es la Probabilidad que una Observación Positiva (tomada al azar) tenga un score mayor que una Observación Negativa (tomada al azar).



c = 4
Dens
TPR
FPR
Distribución del score de la Población TP
c = 3
Dens
TPR
FPR
Distribución del score de la Población TN
100%
TPR
c
c = 3
c = 4
Curva ROC
Curva de peor desempeño
Area bajo la curva (AUC)
0%
FPR
100%

Trade-off Sesgo-Varianza

Variable a predecir

Variable predictora

$$\boxed{y = f_k(x) + \epsilon}$$

Error aleatorio

Complejidad del modelo

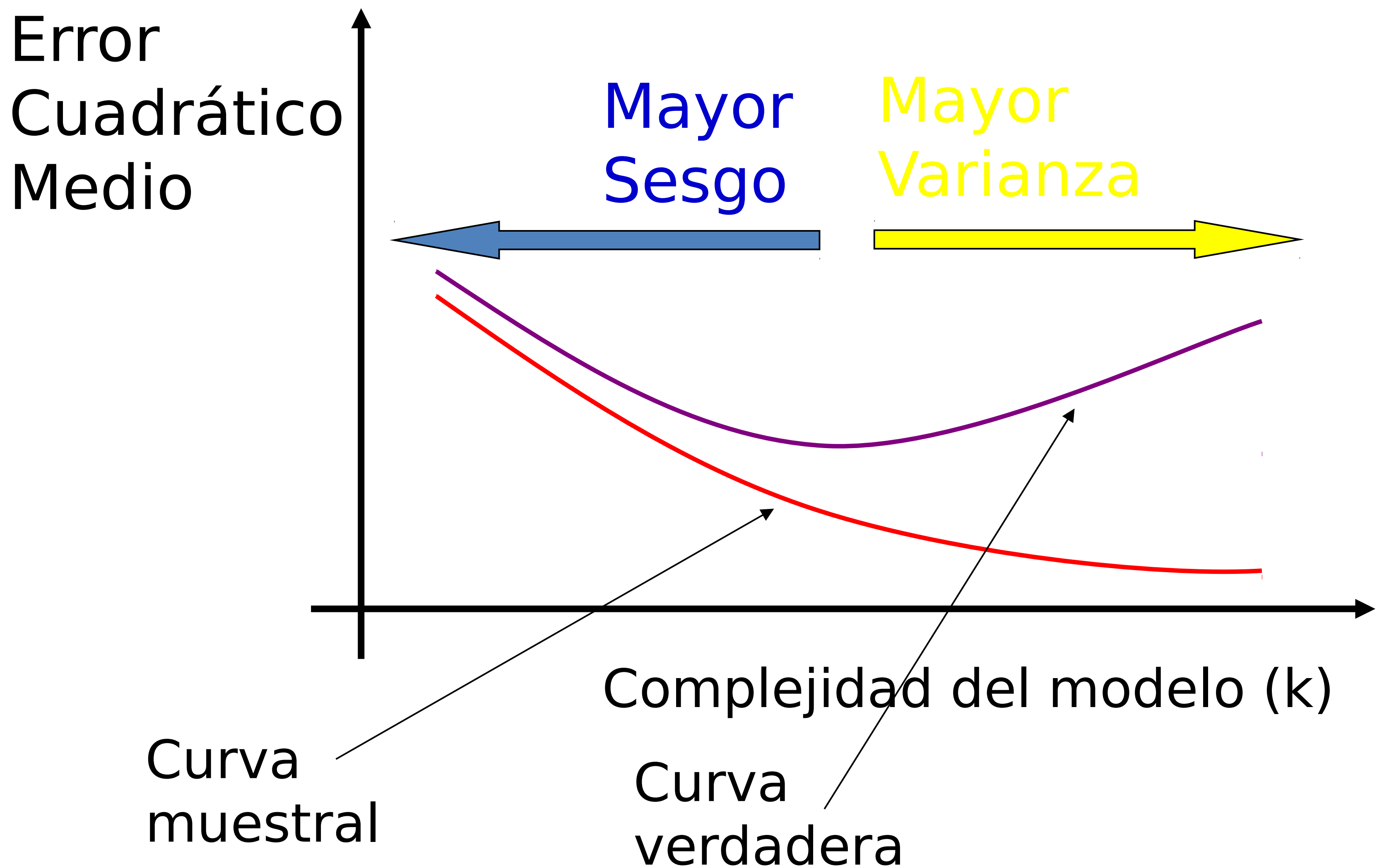
$$\mathbf{E}\Big[\big(y - \hat{f}(x)\big)^2\Big] = \mathrm{Bias}\big[\hat{f}(x)\big]^2 + \mathrm{Var}\big[\hat{f}(x)\big] + \sigma^2$$

Variabilidad irreducible

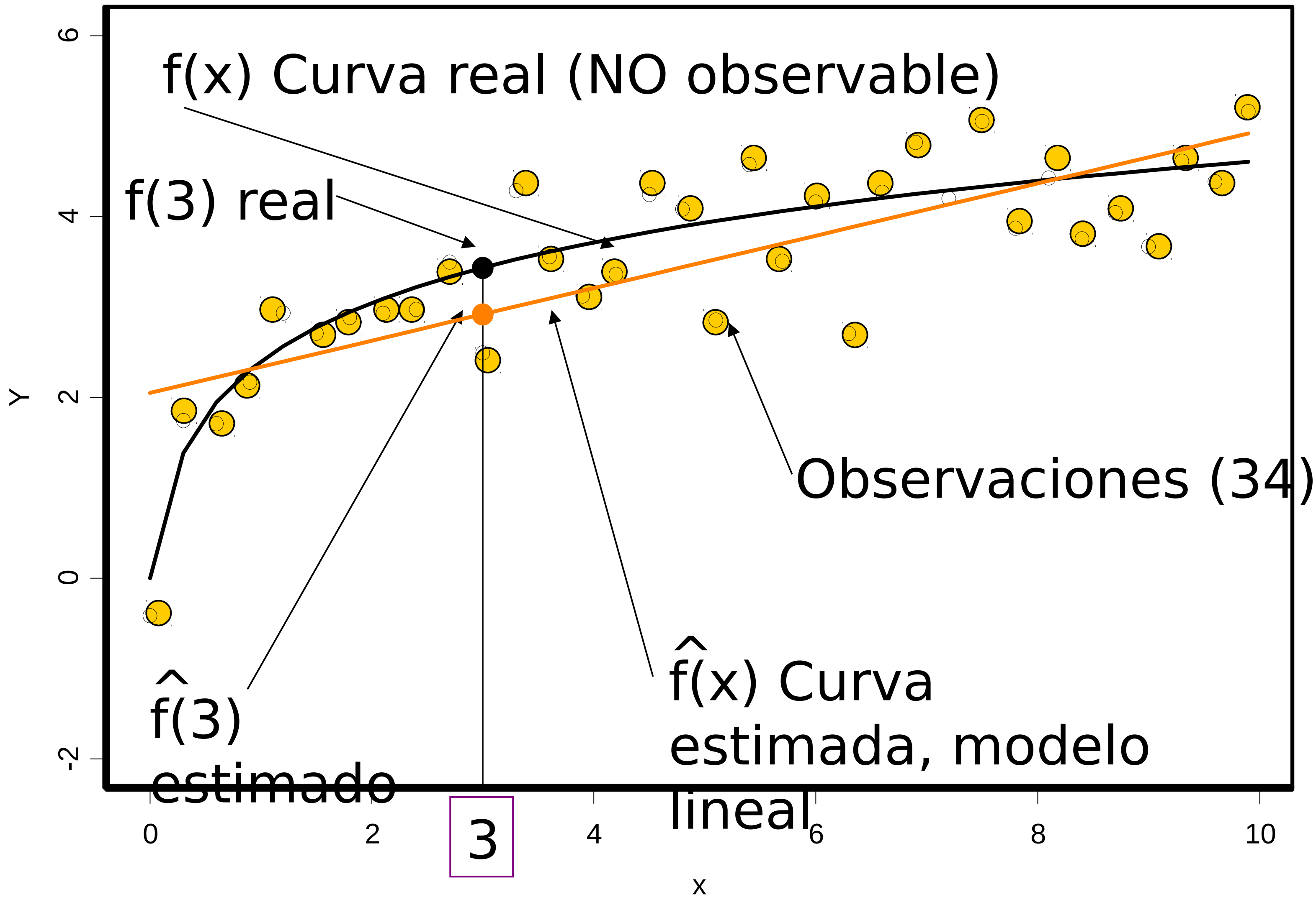
$$\mathrm{Bias}\big[\hat{f}(x)\big] = \mathbf{E}\big[\hat{f}(x) - f(x)\big]$$

$$\mathrm{Var}\big[\hat{f}(x)\big] = \mathbf{E}[\hat{f}(x)^2] - \mathbf{E}[\hat{f}(x)]^2$$

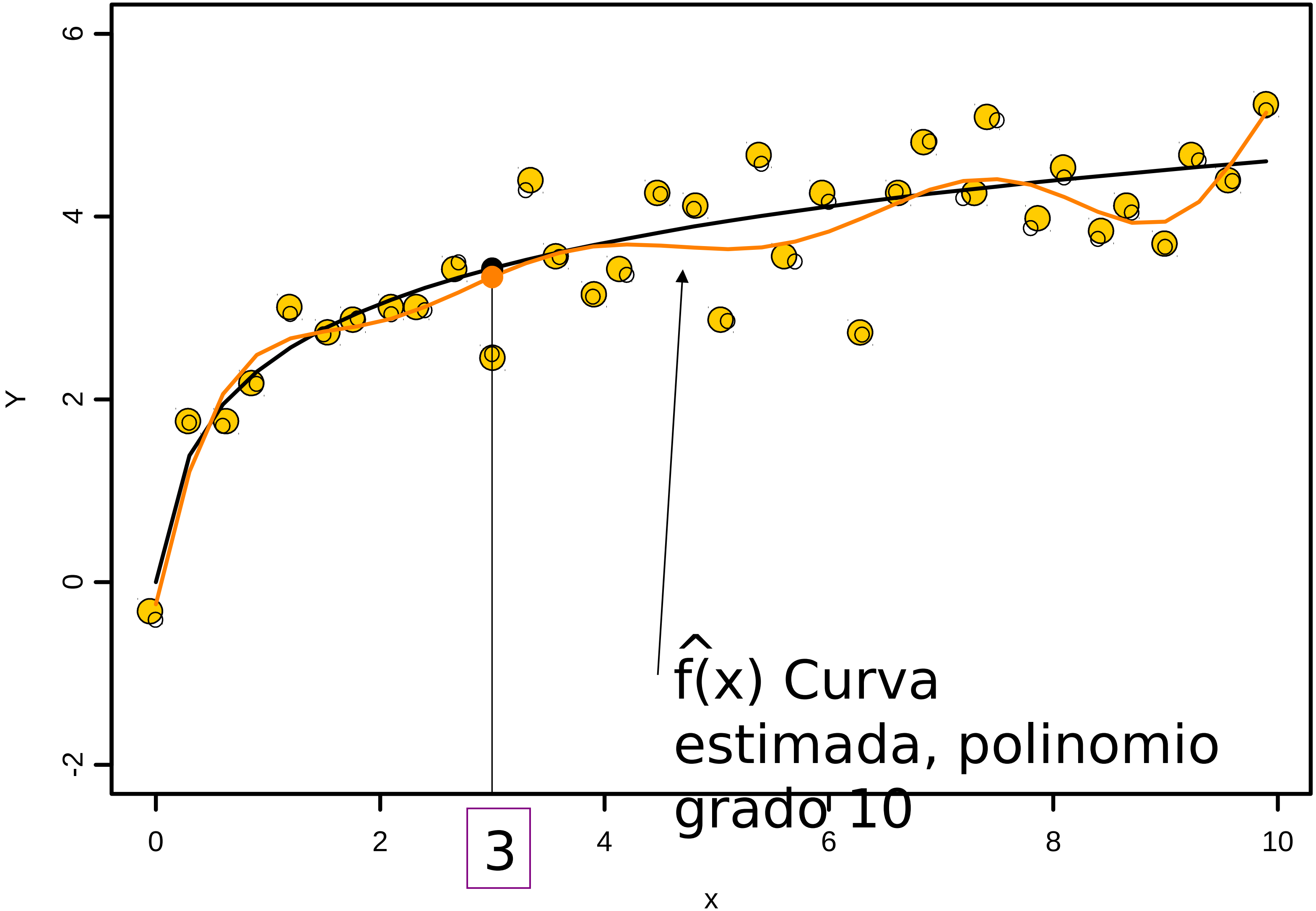
Tradeoff Sesgo - Varianza



Ajuste de $Y = f(X) + \varepsilon = \mathrm{Ln}(10*X+1) + \varepsilon$

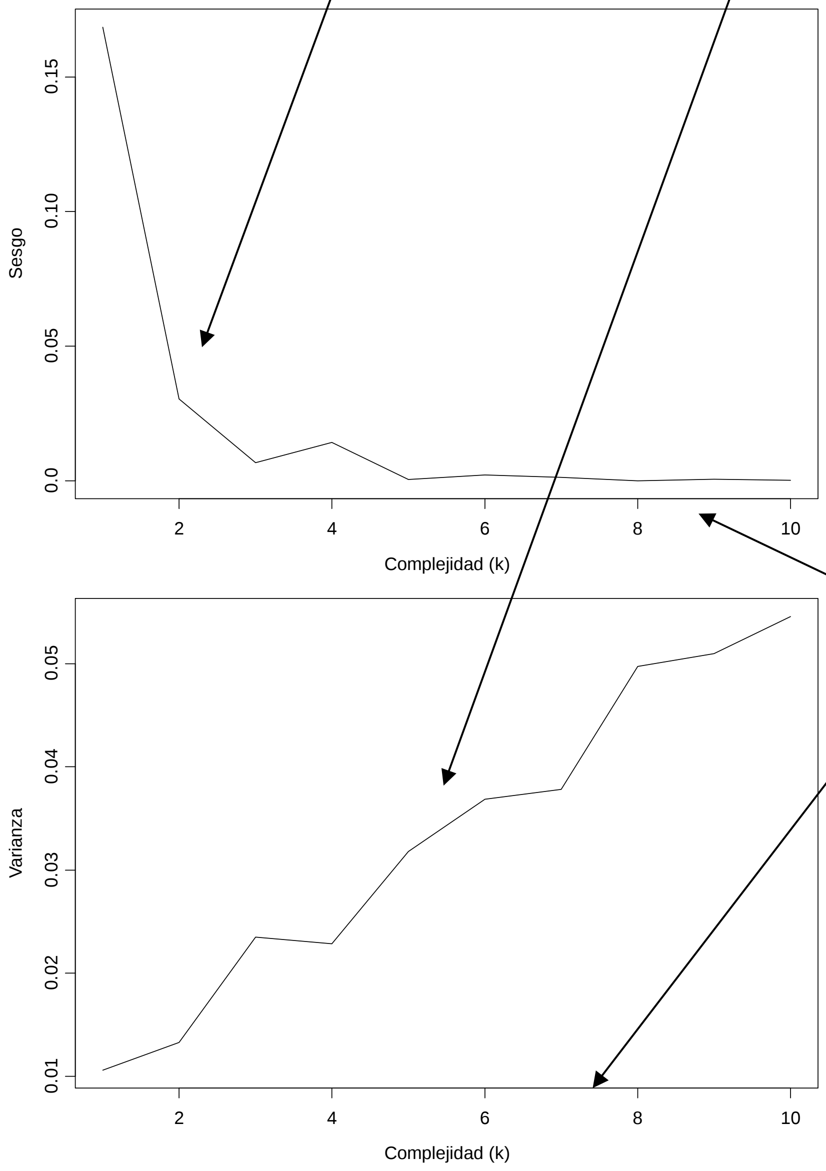


Ajuste de Y = f(X) + ε = Ln(10*X+1) + ε



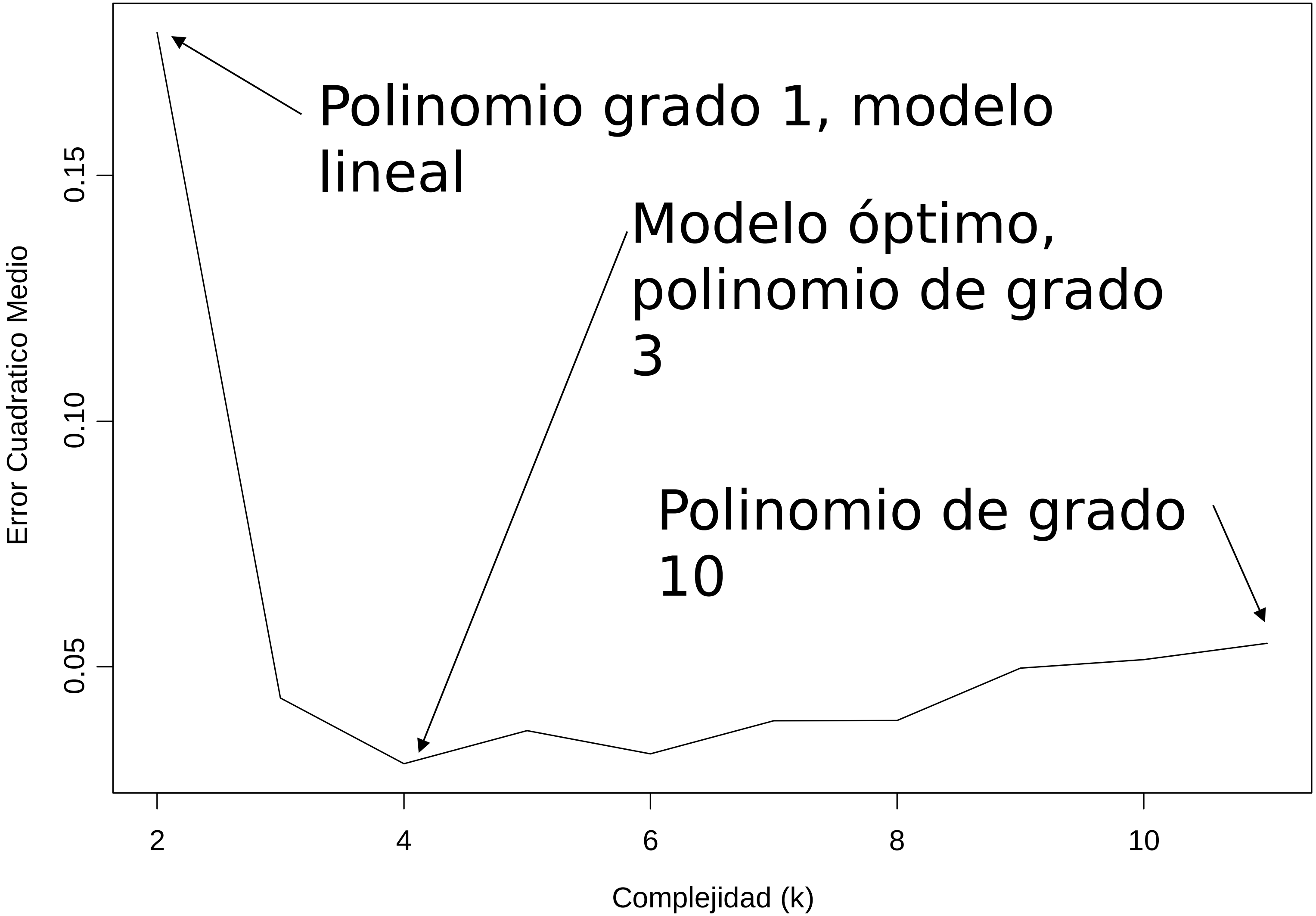
$\hat{f}(x)$ Curva estimada, polinomio grado 10

Tradeoff Sesgo - Varianza



Complejidad del Modelo = Grados del Polinomio

Curva de Error Cuadrático Medio

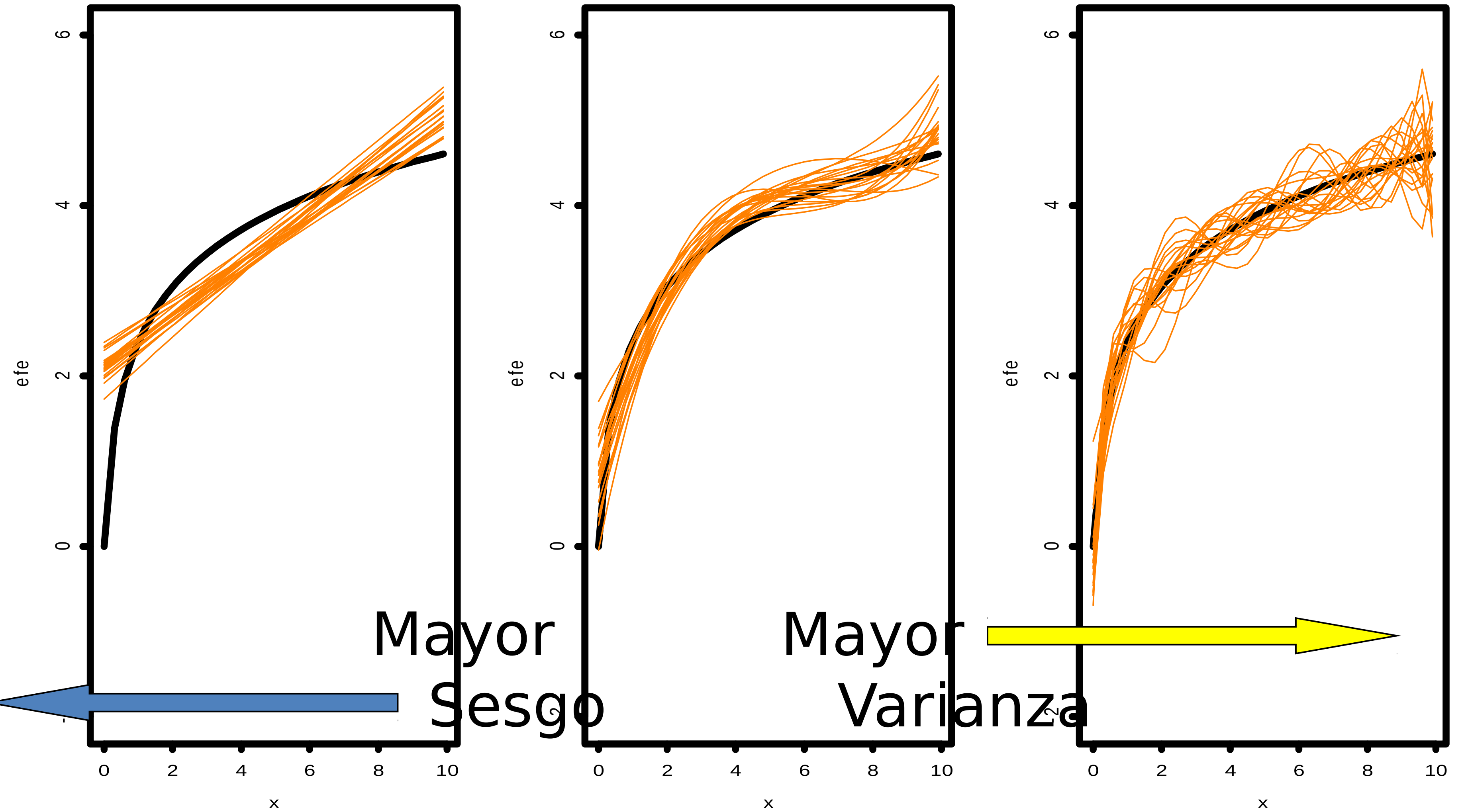


Tradeoff Sesgo-Varianza

Complejidad k=2
Polinomio grado 1

Complejidad k=4
Polinomio grado 3

Complejidad k=11
Polinomio grado 10



Overfitting (sobreajuste)

- Efecto nocivo que proviene de ajustar modelos con mas complejidad (menos parsimonia) que la que la cantidad de información muestral admite.
- Resultado de la ALTA VARIABILIDAD del estimador.
- Es MUCHO mas común que el SUBAJUSTE.
- Estamos geneticamente

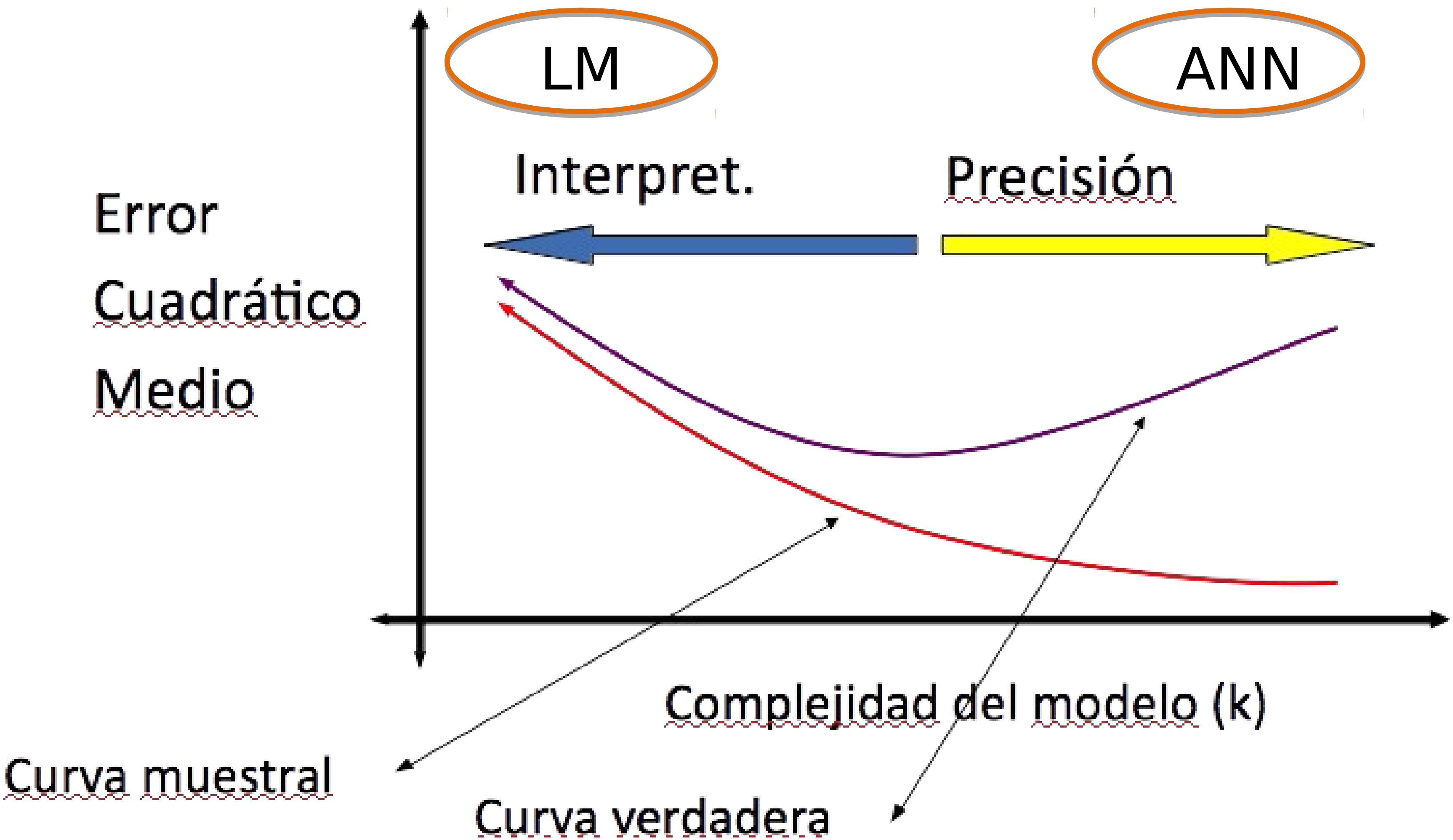
Progrmados

para el OVERFITING !

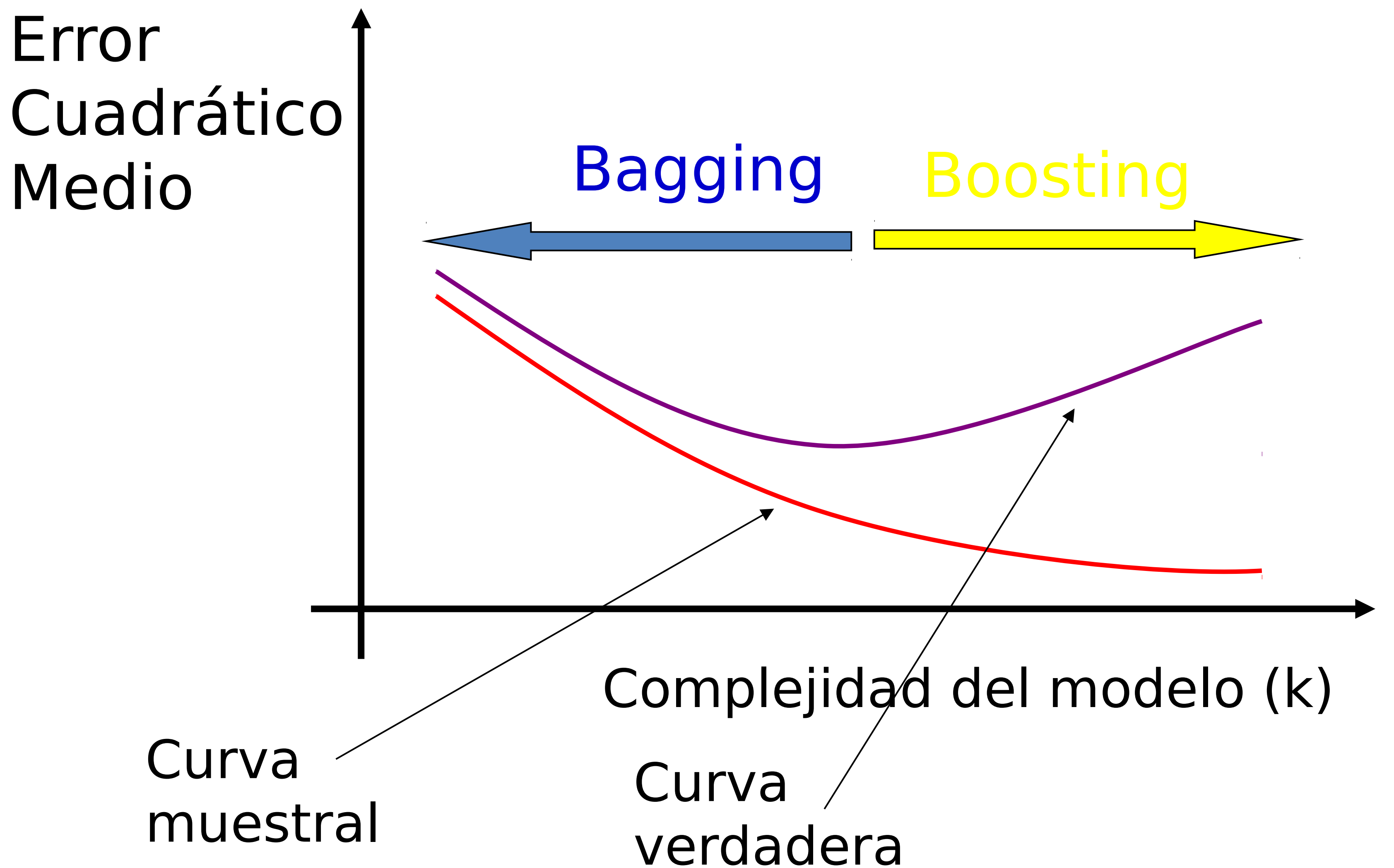
Solución al Overfitting

- Usar modelos poco complejos (mas parsimoniosos)
- Partir la muestra: Entrenamiento / Testeo
- Usar validación cruzada
- Bootstrap y Bagging
- Usar técnicas de “Shrinkage”, como:
  - Ridge Regression
  - LASSO Regression
  - Penalización o Regularización

Tradeoff Precisón - Interpretabilidad



Meta - Heurísticas



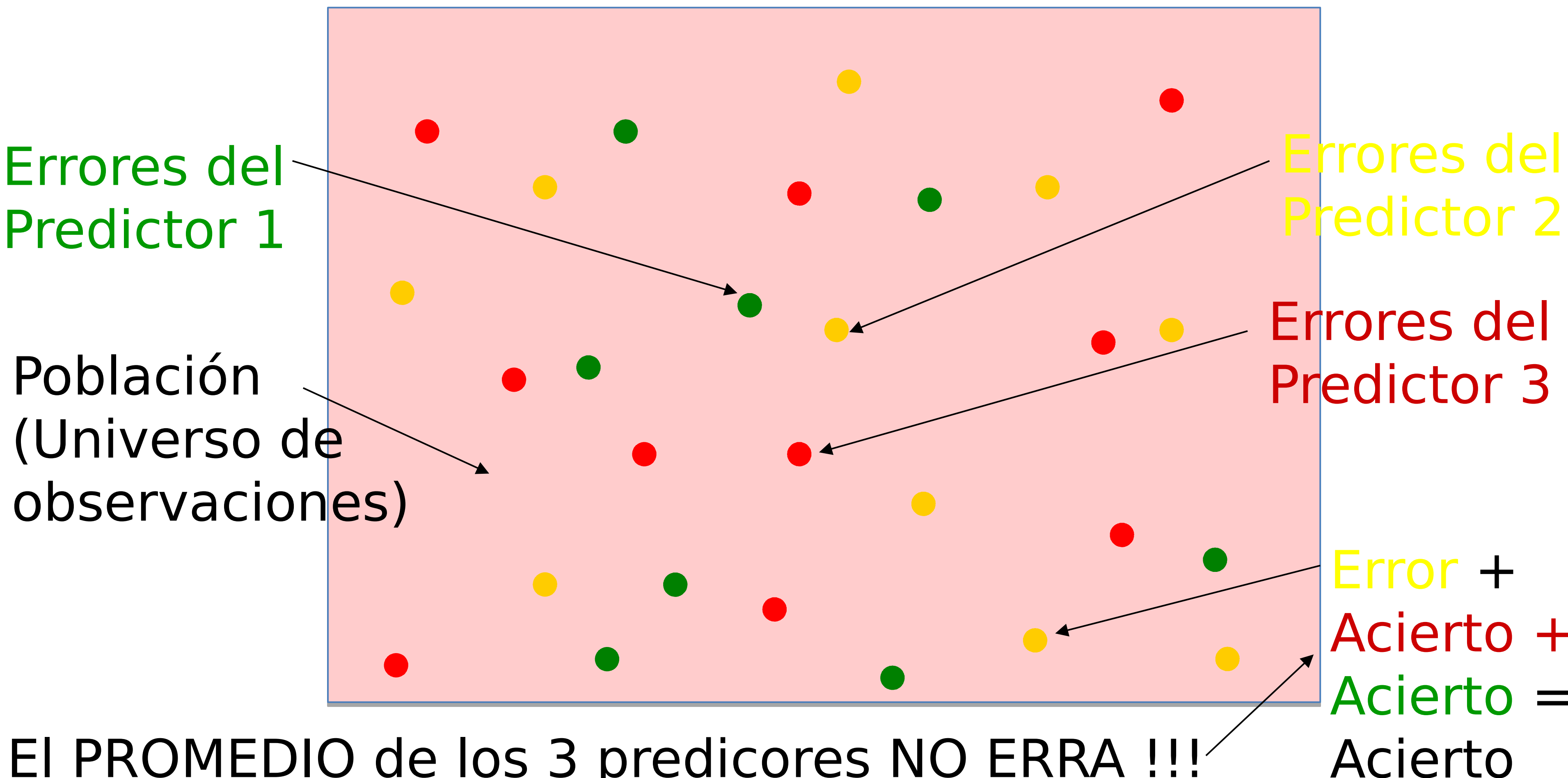
Bagging (Bootstrap Aggregating)

- Se basa en PROMEDIAR los resultados de iterar la aplicación de modelos COMPLEJOS, "bootstrapeando" la muestra de entrenamiento.
- Esta técnica reduce la VARIANZA típico de los modelos COMPLEJOS.

Intuición del Bagging en Clasificación

Los circulos de colores representan a las observaciones en las que los predictores erran.

Con 3 predictores COMPLEJOS (insesgados y muy variables) e INDEPENDIENTES



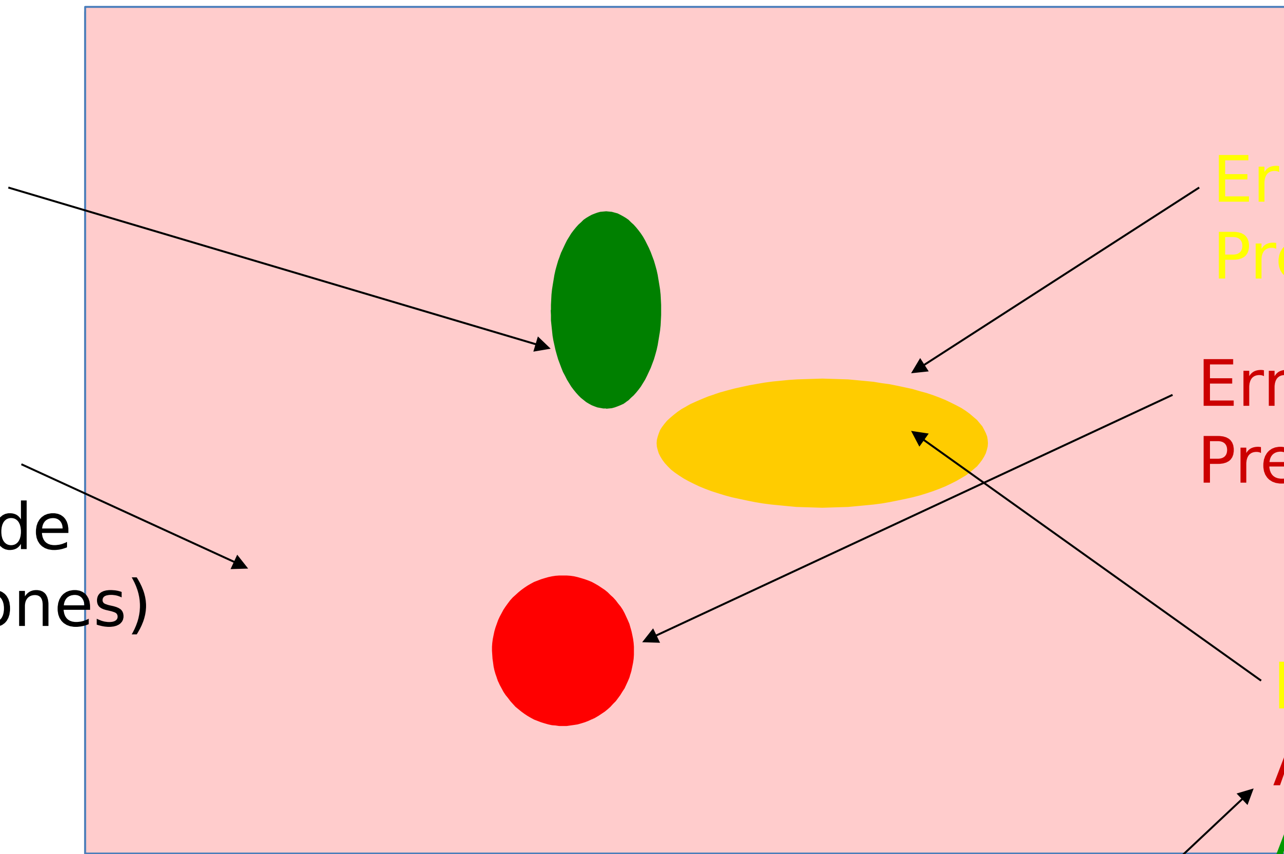
Boosting

- Se basa en RE-ENTRENAR iterativamente modelos SIMPLES aumentando la ponderación de las observaciones PEOR predichas.
- Esta técnica reduce el SESGO típico de los modelos SIMPLES.

Intuición del Boosting en Clasificación

Los circulos de colores representan a las observaciones en las que los predictores erran.

Con 3 predictores SIMPLES (sesgados y poco variables) e INDEPENDIENTES

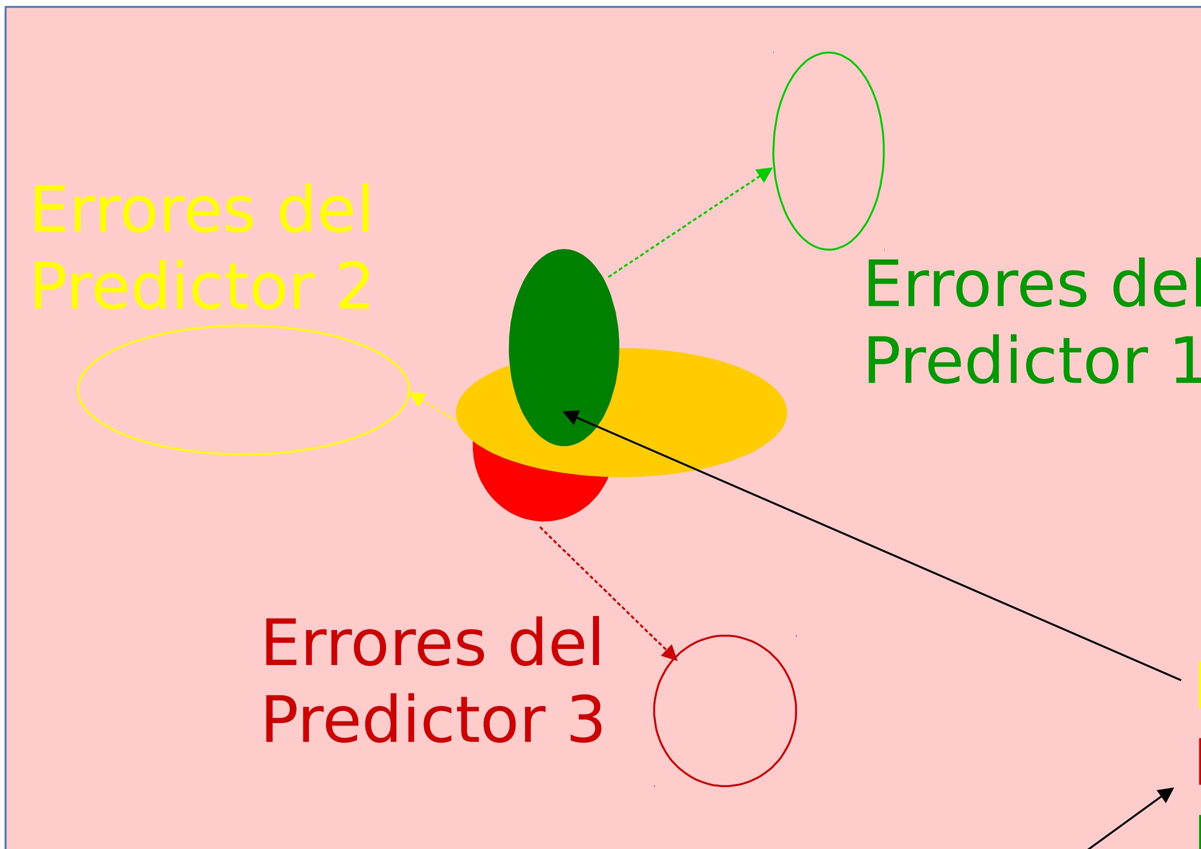


El PROMEDIO de los 3 predicores NO ERRA !!!

En la Práctica: Boosting en Clasificación

Con 3 predictores SIMPLES (sesgados y poco variables) que NO SON INDEPENDIENTES

¿ Como SEPARAMOS las regiones de ERROR de los predictores?

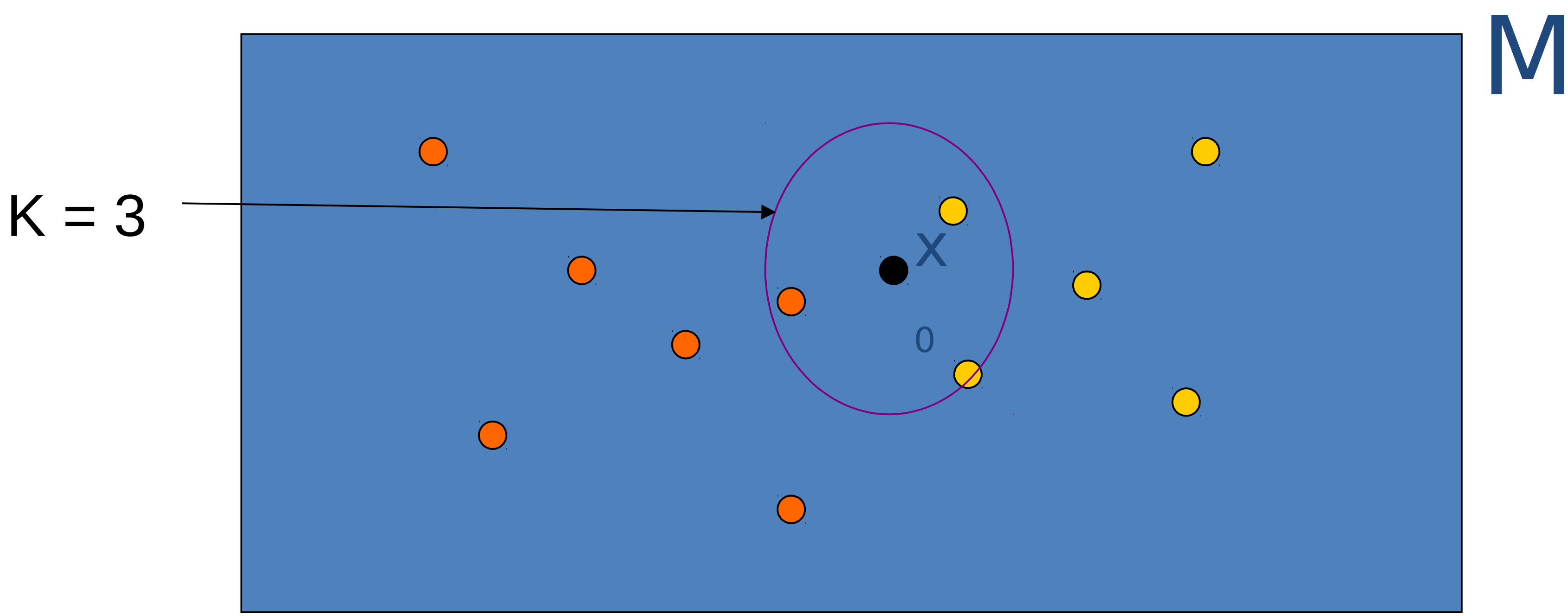


Error + Error + Error = Error

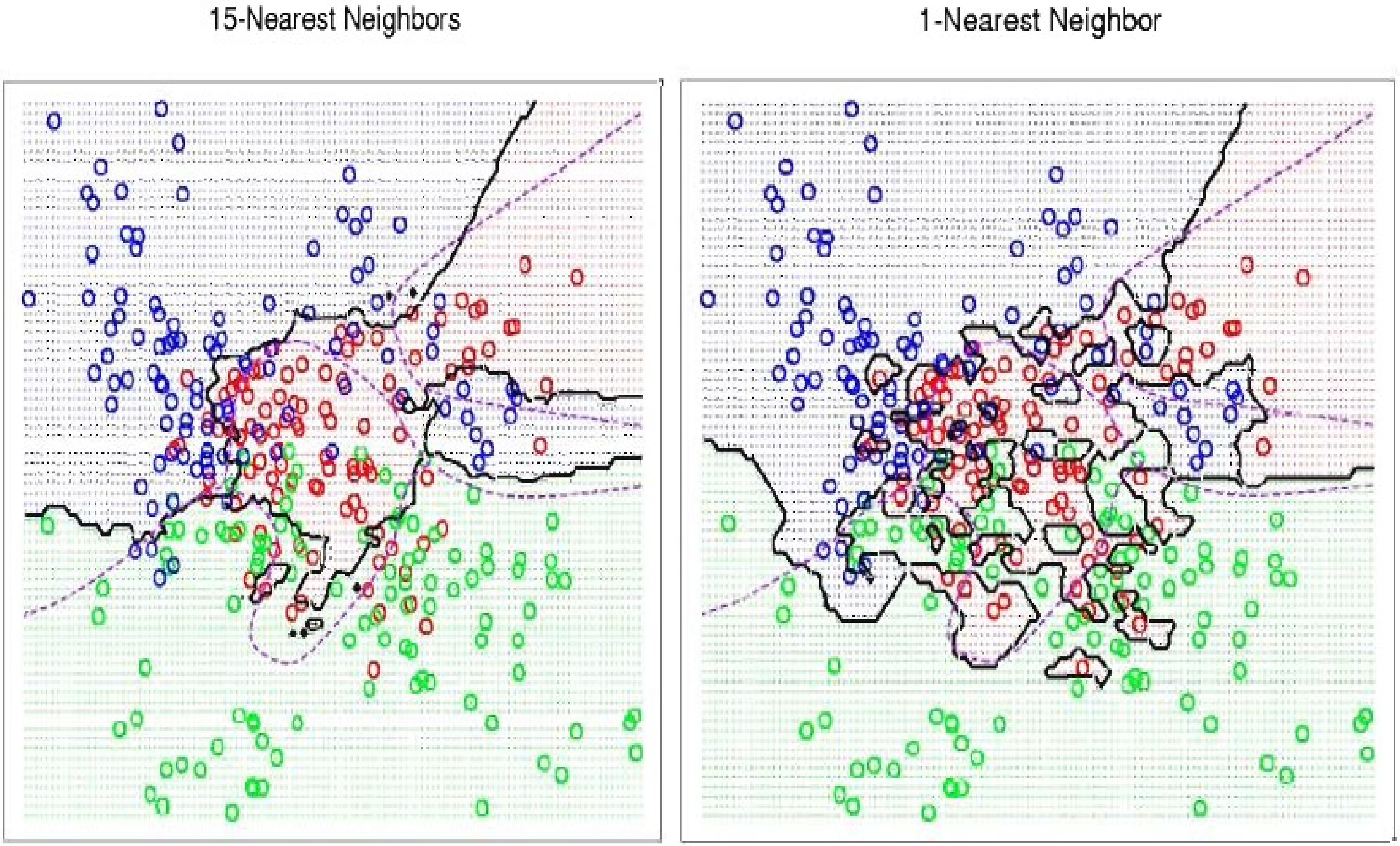
El PROMEDIO de los 3 predicores ERRA !!!

K vecinos mas cercanos (KNN)

- Dada una nueva observación X_0, la clasifico en aquella población que posee una representación mayoritaria entre los K vecinos mas cercanos a X_0.



Ejemplo gráfico



Estimando el error de testeo

TRAIN / TEST

Se divide la muestra en 2 partes: muestra de entrenamiento y muestra de testeo.

Se ajusta el modelo usando la muestra de entrenamiento y el modelo ajustado se usa para predecir las respuestas de la muestra de testeo

El error cuadrático medio calculado con las observaciones de la muestra de testeo es un estimador del error de testeo.

Validación cruzada *Leave one out*

LOOCV

Una sola observación se usa como muestra de testeo y todas las demás como muestra de entrenamiento. Supongamos que sacamos $(\mathbf{x}_1, y_1)$

$$ECM_1 = (y_1 - \hat{y}_1)^2$$

es un estimador "aproximadamente insesgado" del error de testeo

Este procedimiento se repite n veces, dejando afuera una observación cada vez. Obtenemos

$$ECM_1, ECM_2, \ldots, ECM_n.$$

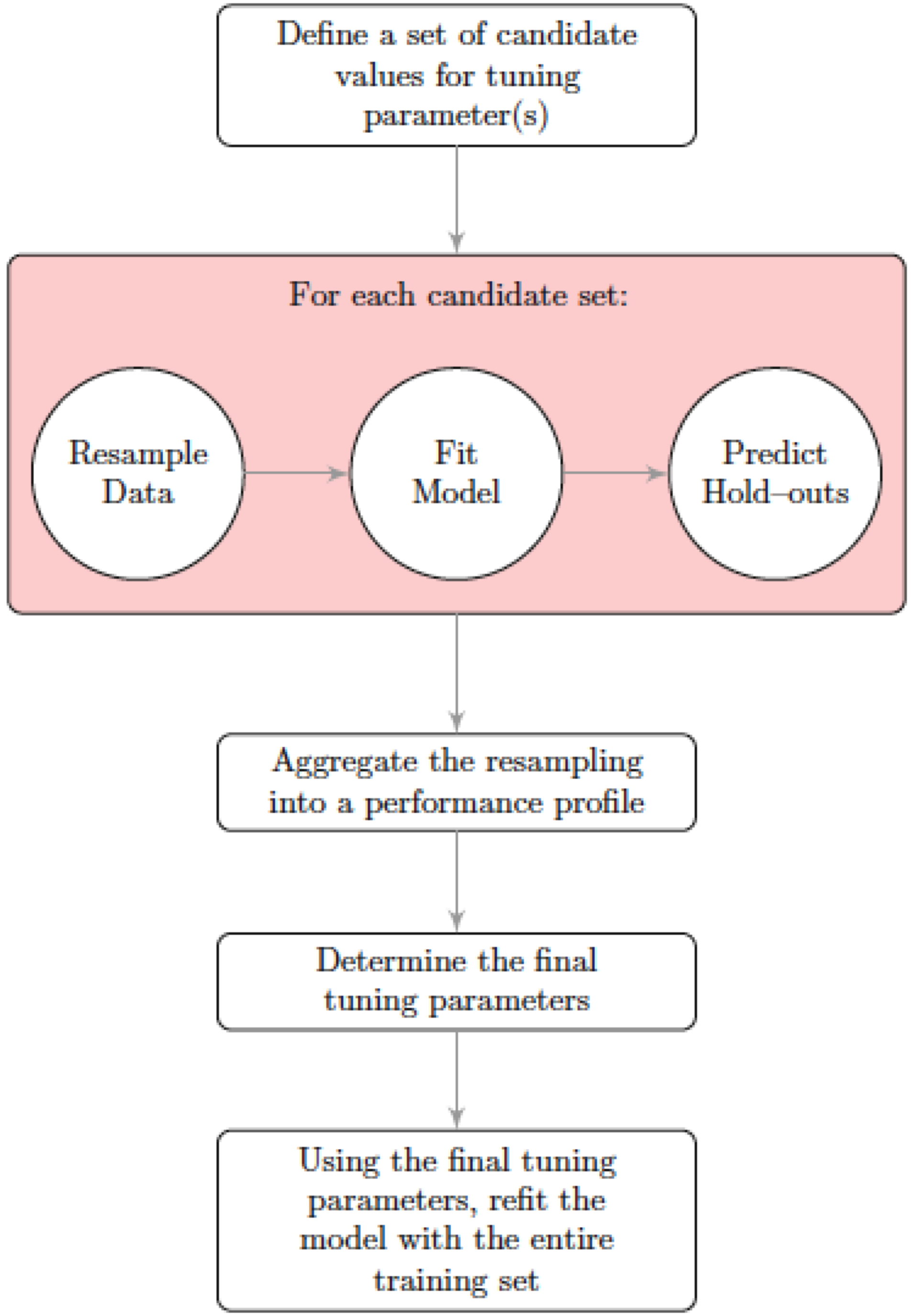
El estimador *LOOCV* del error de testeo es el promedio de estos :

$$ECM_k = \frac{1}{n}\sum_{i=1}^{n} ECM_i.$$

LOOCV Versus TRAIN/TEST

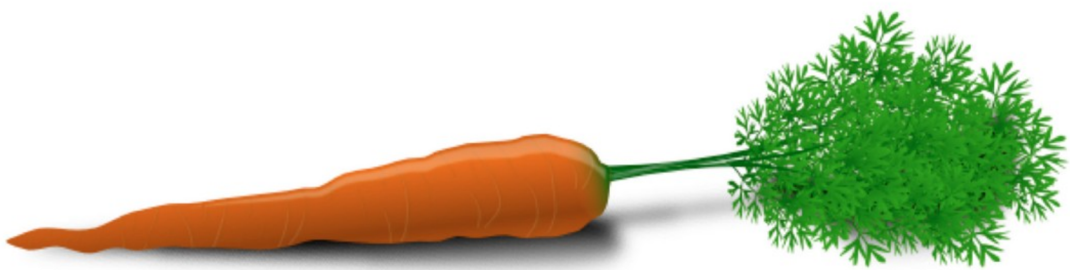
- Dividir a la muestra en submuestra de entrenamiento y de testeo da un estimador del error de testeo sesgado.
- LOOCV da estimadores aproximadamente insesgados del error de testeo, pero tienen mayor varianza.
- 5-fold o 10-fold cv es conveniente porque da un buen compromiso entre sesgo y varianza.

K-fold CV


Define a set of candidate values for tuning parameter(s)
For each candidate set:
Resample Data
Fit Model
Predict Hold–outs
Aggregate the resampling into a performance profile
Determine the final tuning parameters
Using the final tuning parameters, refit the model with the entire training set

The caret R package

the caret package



Links

`train` Model List

Topics

- Main Page
- Data Sets
- Visualizations
- Pre-Processing

The **caret** package (short for *C*lassification *A*nd *RE*gression *T*raining) is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

http://caret.r-forge.r-project.org/