Discriminacion y clasificacion.

Objetivo: separar datos (observcaciones) y posicionarlos en grupos diferentes previamente definidos.

Discriminación: Normalmente exploratorio y usado para investigar/ratificar diferencias en objetos representados en la muestra.

Clasificación: menos exploratorio, se fijan reglas de antemano para clasificar nuevos objetos u observaciones.

Concretamente:

1. Describir graficamente o algebraicamente las diferencias entre objetos de distintas poblaciones conocidas.
2. Organizar observaciones en 2 o más grupos.
3. Derivar una regla de clasificacion para asignar nuevos objetos.

Separación y clasificación de 2 poblaciones

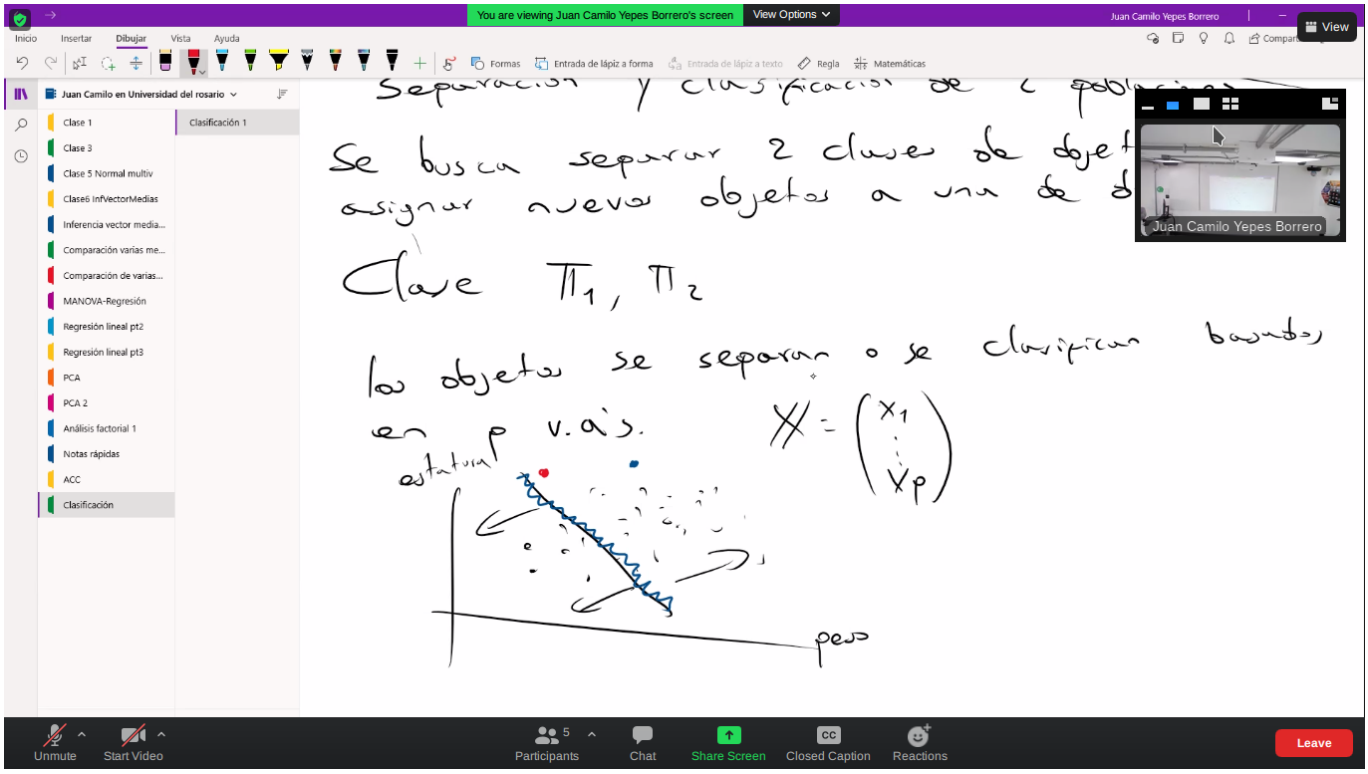
Se buscan separar 2 clases de objetos o asignar nuebos objetos a una de dos clases.

Clases:

$$\Pi_1,\ \Pi_2$$

Los objetos se separan o se clasifican basados en p v.a's.

$$\mathbb{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_2 \end{bmatrix}$$



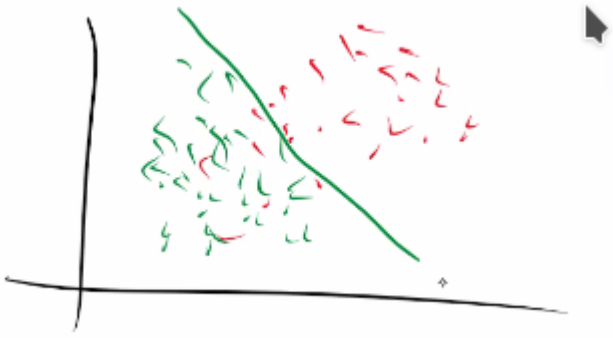
Los valores observados deberian tener alguna diferencia asociada a las clases. *ej: los hombres suelen pesar un poco más que las mujeres.*

La clase Π_1 se asocia a una población 1 con PDF $f_1(\mathbb{X})$, la clase Π_2 se asocia a una población 2 con PDF $f_2(\mathbb{X})$.

Clasificación:

método general:
Se usan reglas "aprendidas" de muestras de aprendizaje *(o entrenamiento)*.

Plot de pesos en hombres (rojo) y mujeres (verde), separados por un plano que ayuda a la clasificación

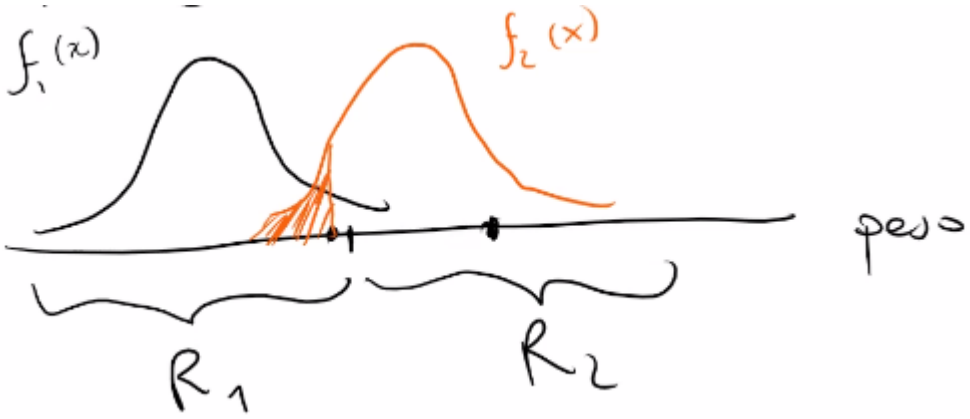


Se examinan dos aleatorias cuya población es conocida y se estudian las diferencias.

Buscamos dividir el conjunto de todas las observaciones en 2 regiones R_1, R_2 tales que si una nueva observación cae en R_1 se clasifica como Π_1 (igual con R_2 y Π_2).

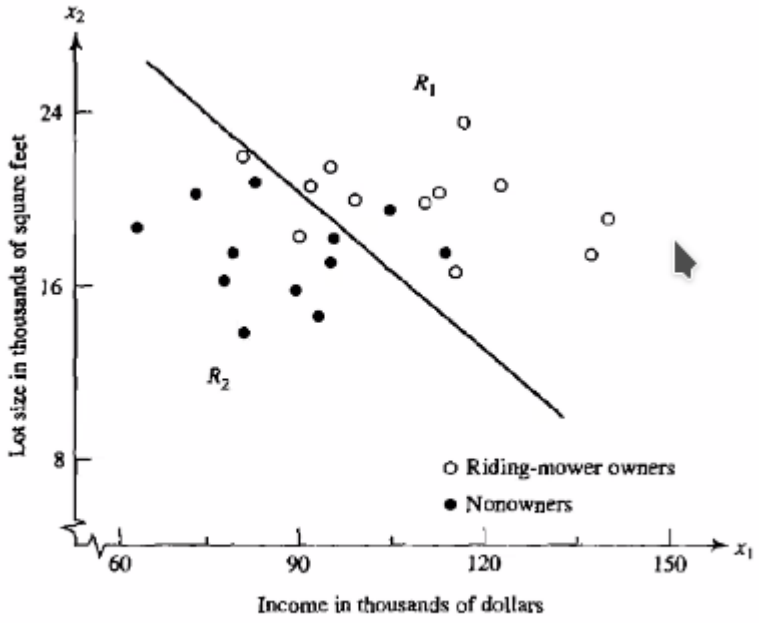
Observación: en general las clasificaciones tienen errores *(la distribución entre Π_1 y Π_2 no es perfecta dada las mediciones)*.

Pude haber un error, ejemplo el punto medio de la gráfica, es más probable que sea de la clase R_2 , esto es como el error tipo II, clasificar como R_1 cuando es realmente de R_2. Si queremos minimizar el error tipo II, recuerde que necesitamos una muestra más grande.



Ejemplo:

Note que la variable de ingresos discrimina mucho mejor que la de tamaño del terreno. Pero juntas hacen una combinación más clara para clasificar.



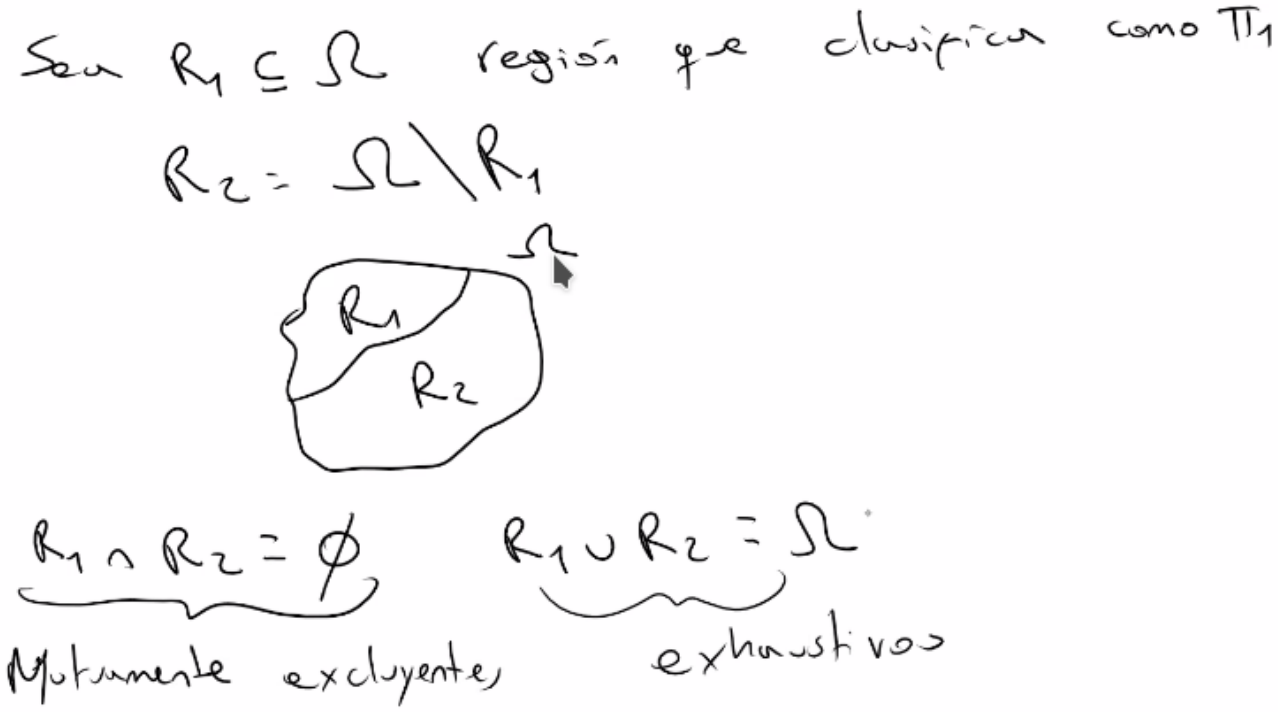
- **Obs:**

1. si es muy improbable ser de Π_2, no se debería clasificar como Π_2.
2. Si es muy costoso clasificar un Π_1 como Π_2 pero no un Π_2 como Π_1 se debería ser cauto.

Suele existir un error que es más grave de cometer que otro.

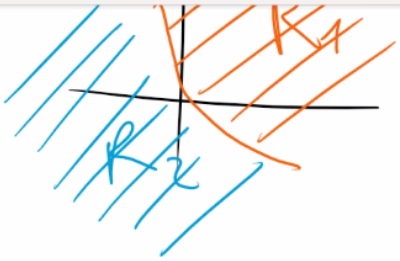
Sea $f_1(\mathbb{x})$, $f_2(\mathbb{x})$ las PDF's de $\mathbb{X}_{p\times 1}$ (vector aleatorio) de las poblaciones Π_1 Π_2.

Sea Ω el espacio muestral *(todos los posibles valores de $\mathbb{X}$)*.



ej: si $p = 2, \ \ \Omega = \mathbb{R}^2$

Formas Entrada de lápiz a forma Entrada de lápiz a texto Regla Matemáticas
Talking: Juan Camilo Yepes Borr...

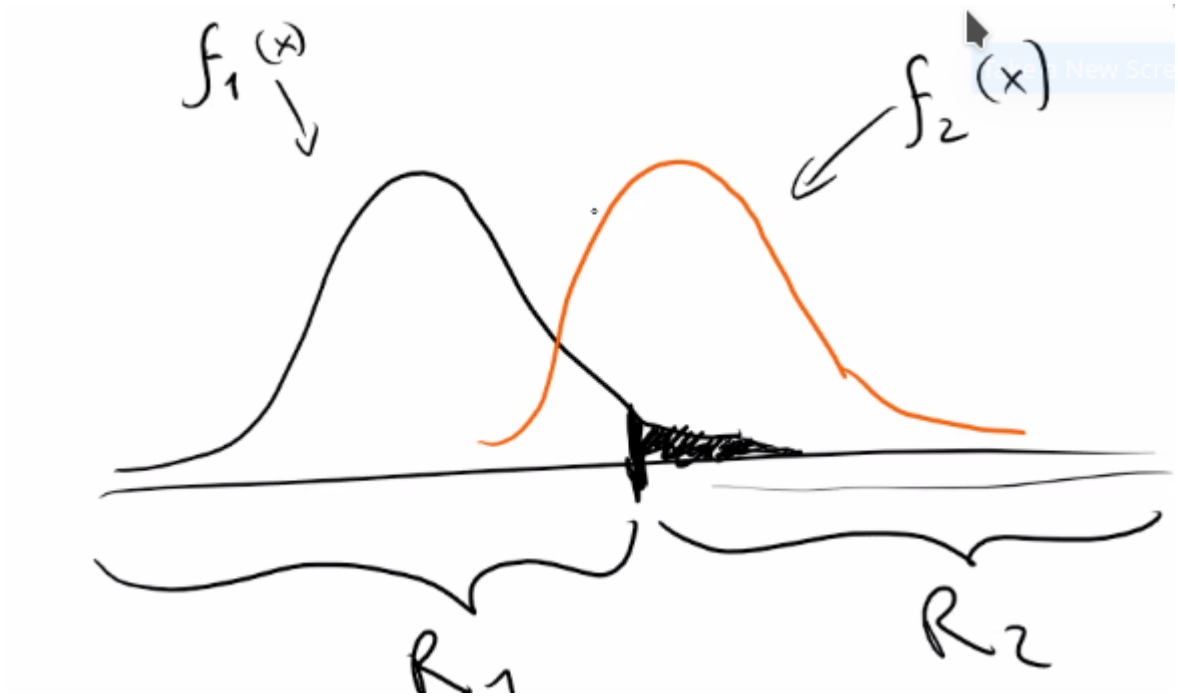


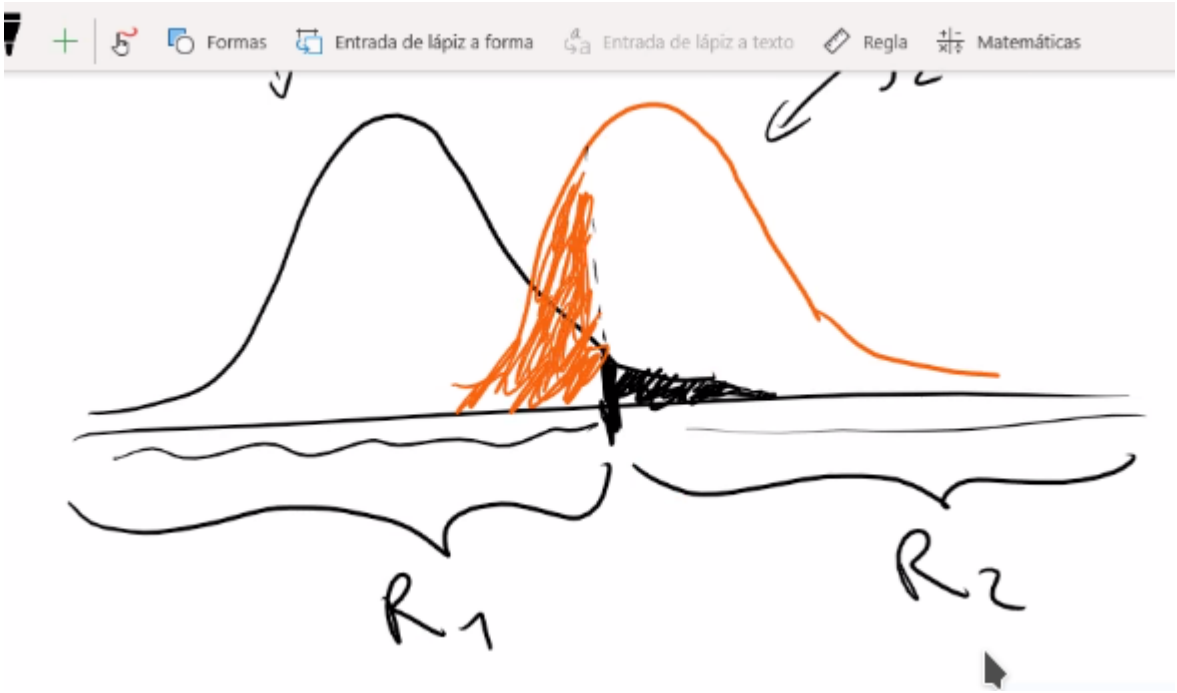
Denotamos $p(2|1)$ prob. de clasificar obj como Π_2 siendo Π_1

$$p(2|1) = P(\mathbb{X} \in R_2 | \Pi_1)$$

$$= \int_{R_2} f_1(\mathbb{X})\, d\mathbb{X}$$

$$p(1|2) = P(\mathbb{X} \in R_1 | \Pi_2) = \int_{R_1} f_2(\mathbb{X}) dx$$





Sean p_1, p_2 las probabilidades previas de Π_1 y Π_2 entonces: $p_1 + p_2 = 1$.

1. $$P(obs \in \Pi_1 \ \ y \ \ clasif \in \Pi_1) = P(\mathbb{x} \in R_1|\Pi_1)P(\Pi_1)$$
$$= P(1|1)p_1$$

2. Clasificacion correcta de Π_2

$$P(\mathbb{x} \in R_2|\Pi_2)P(\Pi_2) = P(2|2)p_2$$

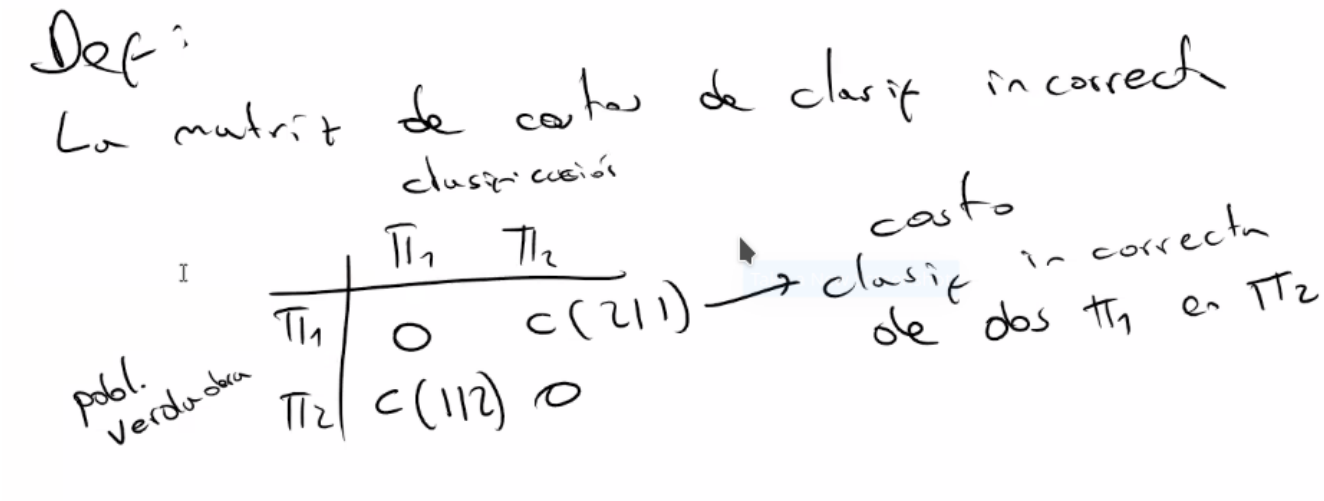
3. P(Clasificacion erronea de Π_1):

$$P(\mathbb{x} \in R_1|\Pi_2)P(\Pi_2) = P(1|2)p_2$$

4. P(Clasificacion erronea de Π_2):

$$P(\mathbb{x} \in R_2|\Pi_1)P(\Pi_1) = P(1|2)p_1$$

Def:



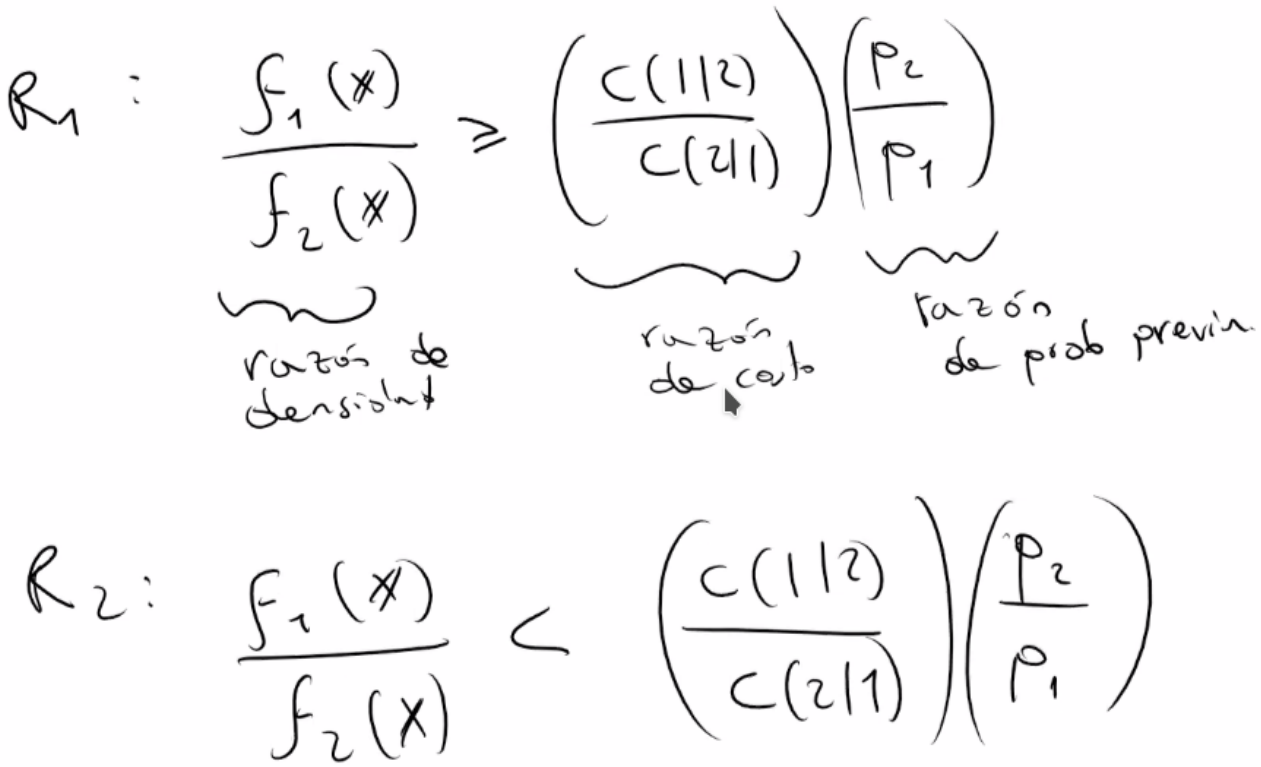
El costo esperado o promedio del costo de clasificacion. incorrecta (formulita de valor esperado aplicado):

$$(1) \quad ECM = c(2|1) \cdot p(2|1)p_1 + c(1|2) \cdot p(1|2)p_2$$

El objetivo es minimizar (1) ***al clasificar.***

Teorema:

Las regiones R_1, R_2 que minimizan (1) se definen por la $\mathbb{x}$ que satisfacen



Tarea
La Demostración es ejercicio pero está en → 11.3 del libro.

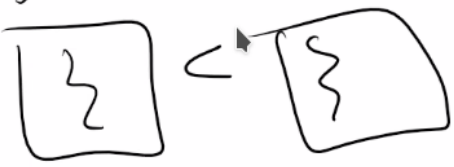
Observación:
Para usar esto necesitamos:

1. La densidad evaluada en una observación $\mathbb{x}_0$
2. Costos.
3. Probabilidad previa (p_1, p_2).

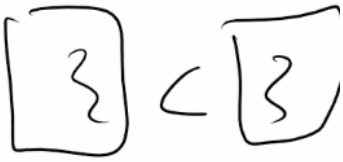
Casos especiales

1) si $p_2/p_1 = 1$ $(p_2 = p_1)$

$$R_1: \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)}$$

R_2: 

2) $c(1|2) = c(2|1) = 1$

$$R_1: \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1}, \quad R_2: \text{  }$$

---

3) $p_2/p_1 = 1 = c(1|2)/c(2|1) = 1$

$$R_1: \frac{f_1(x)}{f_2(x)} \geq 1, \quad R_2: \frac{f_1(x)}{f_2(x)} < 1$$

- *ejemplo:*

Example 11.2 (Classifying a new observation into one of the two populations) A researcher has enough data available to estimate the density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ associated with populations π_1 and π_2, respectively. Suppose $c(2|1) = 5$ units and $c(1|2) = 10$ units. In addition, it is known that about 20% of *all* objects (for which the measurements $\mathbf{x}$ can be recorded) belong to π_2. Thus, the prior probabilities are $p_1 = .8$ and $p_2 = .2$.

Given the prior probabilities and costs of misclassification, we can use (11-6) to derive the classification regions R_1 and R_2. Specifically, we have

$$R_1: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

$$R_2: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

Suppose the density functions evaluated at a new observation $\mathbf{x}_0$ give $f_1(\mathbf{x}_0) = .3$ and $f_2(\mathbf{x}_0) = .4$. Do we classify the new observation as π_1 or π_2? To answer the question, we form the ratio

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = \frac{.3}{.4} = .75$$

and compare it with .5 obtained before. Since

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = .75 > \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = .5$$

we find that $\mathbf{x}_0 \in R_1$ and classify it as belonging to π_1. ■