



# Clasificadores Probabilísticos en Aprendizaje Automático

## Calibración y Rendimiento Independiente de Aplicación

**Daniel Ramos Castro**

Contribuciones de Segio Álvarez Balanya (Estudiante de Máster UAM)

[daniel.ramos@uam.es](mailto:daniel.ramos@uam.es)

Audias – Audio, Data Intelligence and Speech  
Universidad Autónoma de Madrid

<http://audias.ii.uam.es>

< audias >

Audio, Data Intelligence and Speech

UAM

# Sumario del Día

- Teoría de la Decisión (2 clases)
  - Ejemplo: comparación forense de voces
  - Marco probabilístico de decisión
  - Reglas de decisión óptima
- Calibración (2 clases)
  - Ejemplos de calibración extrínseca (basados en *scores*)
- Medida de rendimiento en calibración (2 clases)
- Calibración extrínseca multiclase

# El problema

- Grabación incriminatoria (**dubitada**)
  - Pinchazo telefónico
  - Llamada anónima
  - Micrófono oculto
  - ...
- La policía arresta a un sospechoso
- Se realiza una toma de voz del sospechoso (**indubitada**)
  - En dependencias policiales
  - Pinchazos cuya autoría se reconoce
  - ...
- El contenido lingüístico no se conoce a priori en ambos casos
  - **Independiente de texto**



**Criminal  
(Identidad C)**



**Sospechoso  
(Identidad S)**

# Evidencia

- La evidencia es la relación entre la toma dubitada y la toma indubitada
  - La evidencia nos da información sobre la relación de ambas fuentes
    - Ambas fuentes están relacionadas
    - Ambas fuentes no están relacionadas
- Valorar la evidencia es evaluar esa información



# Planteamiento

- Hipótesis que se manejan:
  - Hipótesis **del fiscal**:  $H_p$ 
    - Ejemplo: “ambas tomas pertenecen a la misma fuente”  
(ventana en la escena del crimen)
  - Hipótesis **del defensor**:  $H_d$ 
    - Ejemplo: “ambas tomas pertenecen a fuentes diferentes”  
(ventanas diferentes)
- Pregunta sobre la que se basa la decisión del juez
  - ¿Cuál es la probabilidad de que, a la luz de **la evidencia** ( $E$ ) y del **resto de información acerca del caso**, el sospechoso sea el autor del robo?

$$¿P\left(H_p \mid E, I\right)?$$

# Solución: Teorema de Bayes


$$P(H_p | E, I) = \frac{P(E | H_p, I) P(H_p | I)}{P(E | I)}$$

$$P(H_d | E, I) = \frac{P(E | H_d, I) P(H_d | I)}{P(E | I)}$$


$$\frac{P(H_p | E, I)}{P(H_d | E, I)} = \frac{P(E | H_p, I) P(H_p | I)}{P(E | H_d, I) P(H_d | I)}$$

# Separación de Roles


$$\frac{P(H_p | E, I)}{P(H_d | E, I)} = \frac{P(E | H_p, I)}{P(E | H_d, I)} \frac{P(H_p | I)}{P(H_d | I)}$$


$$\frac{P(H_p | E, I)}{P(H_d | E, I)}$$




$$\frac{P(E | H_p, I)}{P(E | H_d, I)}$$




$$\frac{P(H_p | I)}{P(H_d | I)}$$



# ¿Rol del científico forense?



- Calcular el *likelihood ratio (LR)*

$$LR = \frac{P(E|H_p, I)}{P(E|H_d, I)}$$

LR>1: apoyo la hipótesis del fiscal

LR<1: Apoyo la hipótesis del defensor

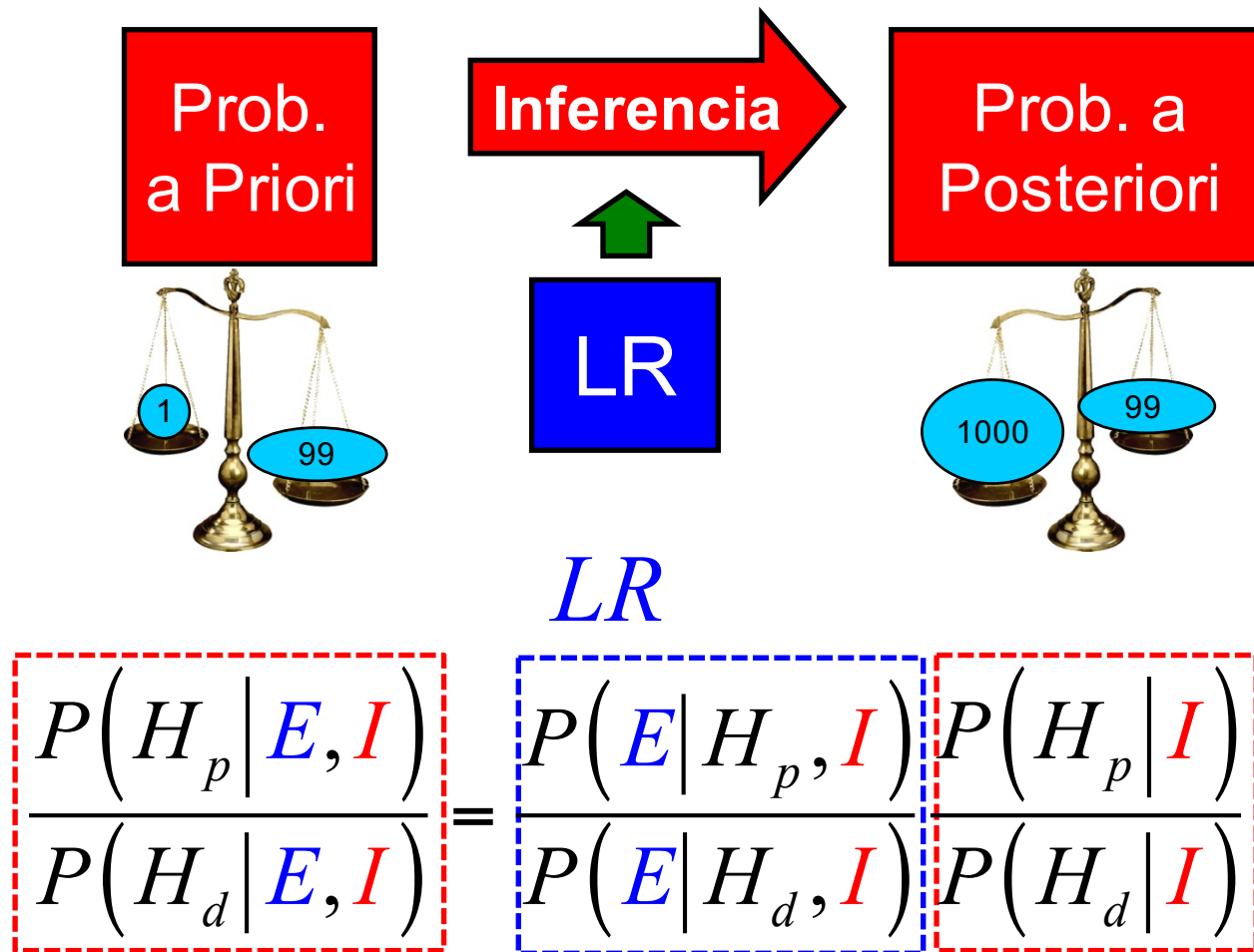
LR=1: No apoyo a nadie

- Cuanto **mayor** (**menor**) el valor del LR, más apoyo a la hipótesis del **fiscal** (**de la defensa**)
- **Clave: ¿cómo calcular el LR?**

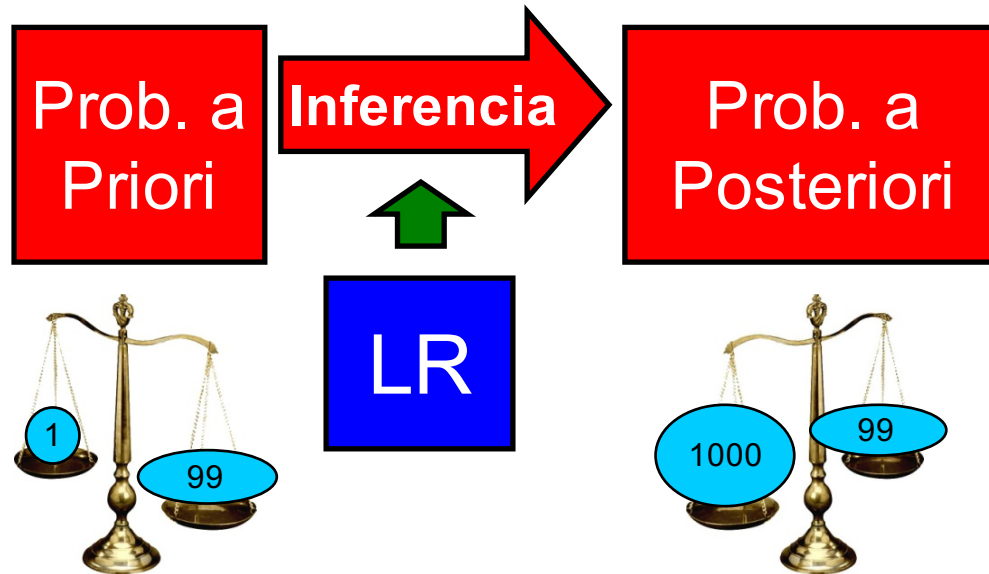


# Inferencia en Ciencia Forense

- Razón de Verosimilitud: valor probabilístico de la evidencia



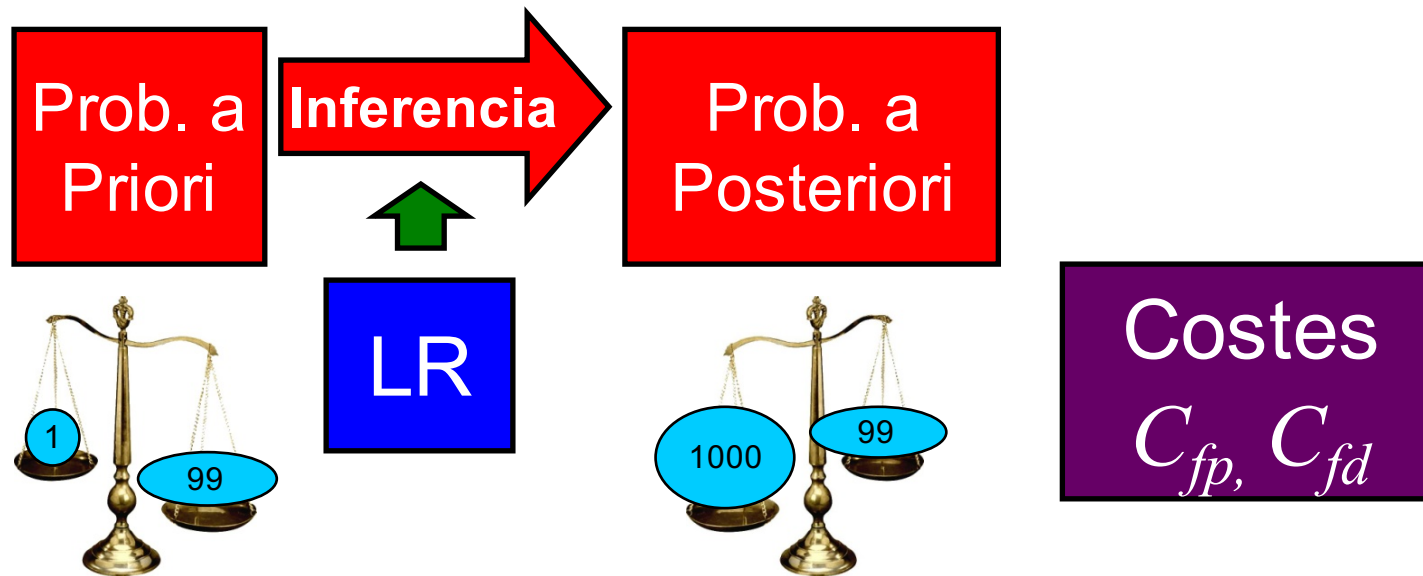
# Decisión en un Caso: Elementos



## ■ Inferencia

- ❑ Probabilidad (apuesta) a priori, sin conocer la prueba
- ❑ Probabilidad (apuesta) a posteriori, una vez conocida la prueba
- ❑ LR: valor de la prueba

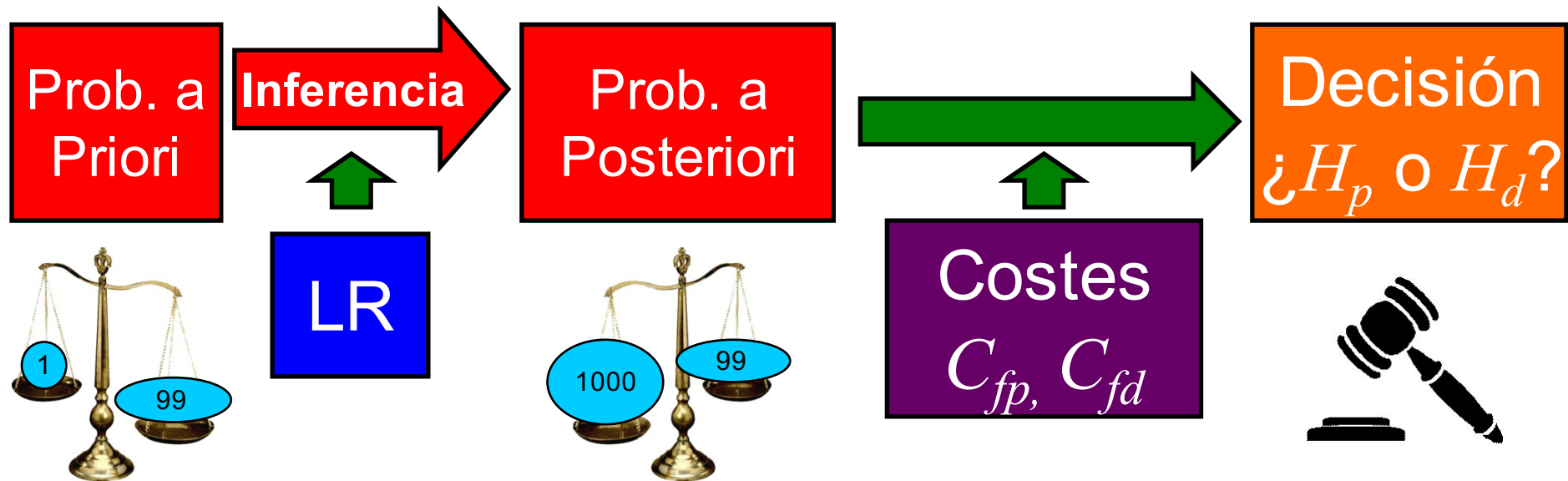
# Decisión en un Caso: Elementos



## ■ Costes

- Castigos por tomar decisiones **incorrectas** en favor de  $H_p$  ( $C_{fp}$ ) o de  $H_d$  ( $C_{fd}$ ).
  - Pueden ser diferentes
    - Ejemplo a nivel de ofensa: ¿es mejor condenar a un inocente (con coste  $C_{fp}$ ) o liberar a un culpable (con coste  $C_{fd}$ )?

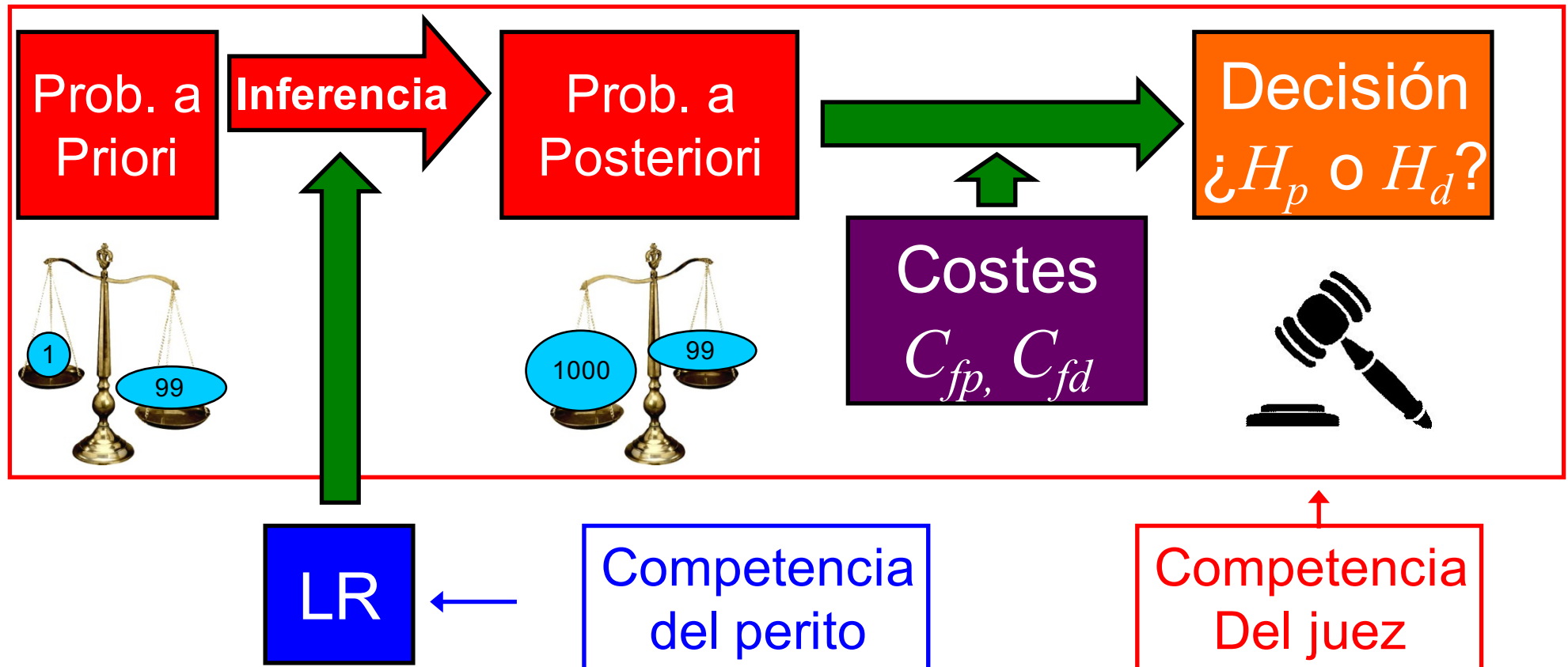
# Decisión en un Caso: Elementos



## ■ Decisión

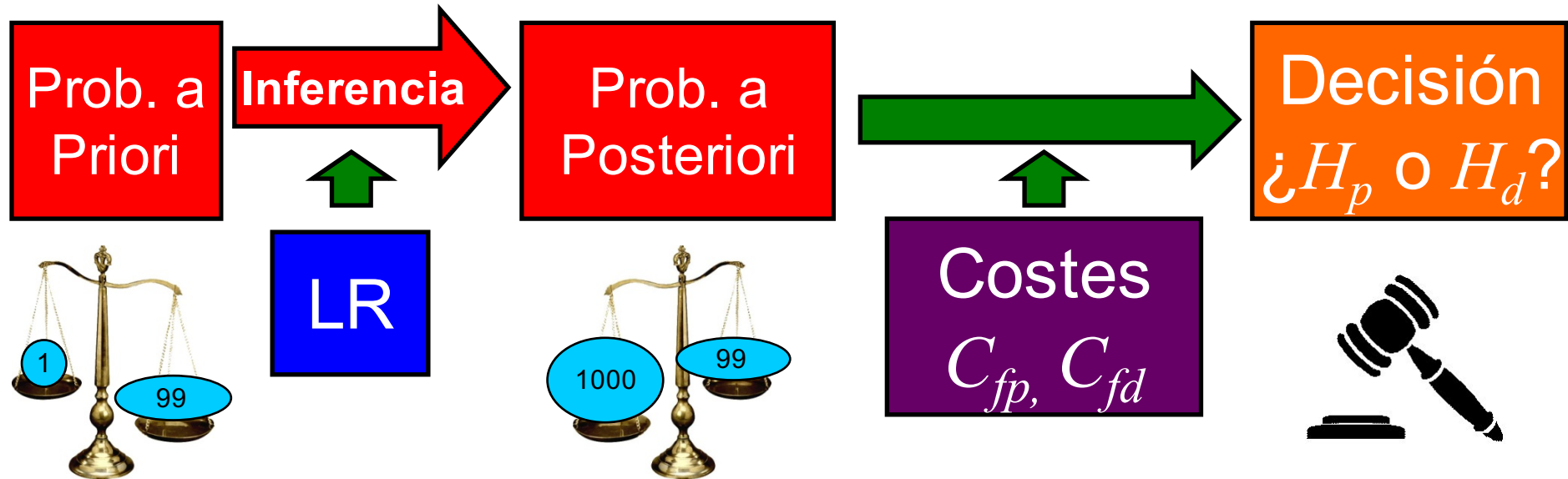
- ❑ Categórica, decide entre proposiciones  $H_p$  o  $H_d$
- ❑ Basada en la probabilidad a posteriori y...
- ❑ ¡También en los costes!

# Decisión en un Caso: Elementos



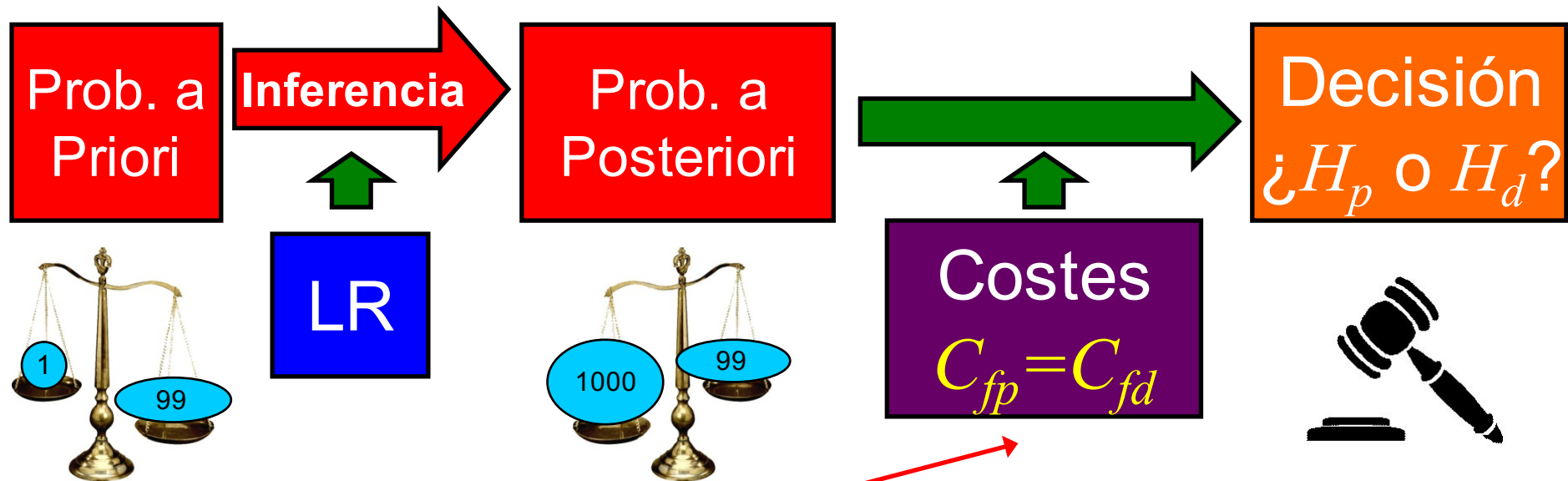
- Todo el proceso es competencia del juez, excepto...
  - ▣ El LR, competencia del perito

# El Perito Influye en las Decisiones



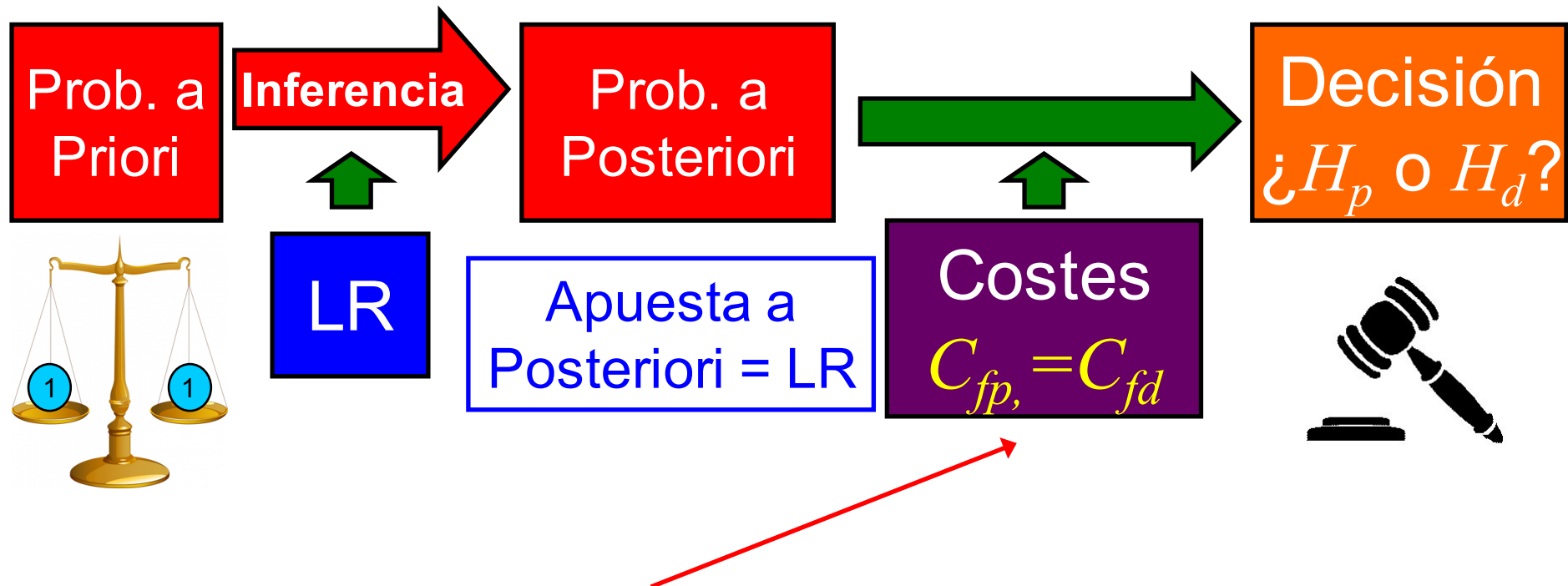
- El juez toma la decisión final
- Pero el perito influye en esa decisión
  - Con el valor del LR

# El Perito Influye en las Decisiones



- Ejemplo (**costes iguales**): apuesta a priori 1 sobre 99
  - Daría lugar a una decisión a favor de  $H_p$  (con costes iguales)
- El perito arroja un  $LR=1000$
- Apuesta a posteriori se calcula como 1000 sobre 99
  - Decisión a favor de  $H_d$  (**el LR cambia la decisión final**)

# El Perito Influye en las Decisiones



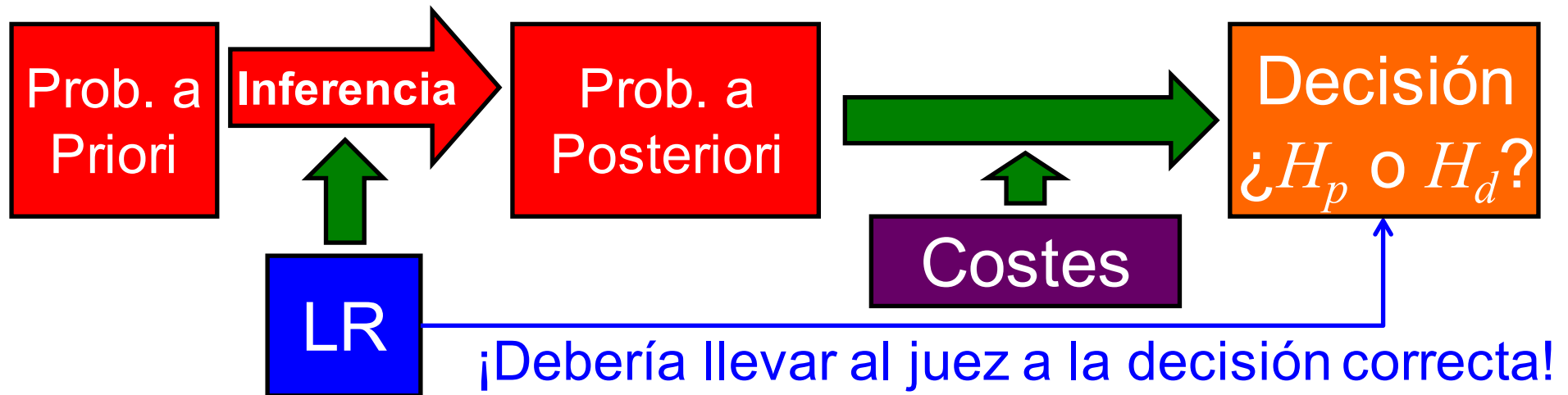
- Ejemplo (**costes iguales**): apuesta a priori no informativa (apuesta=1, probabilidad a priori=0.5)
- El LR domina completamente la apuesta a posteriori
  - Influencia máxima del perito en la decisión final



# Decisión en un Caso: Hechos

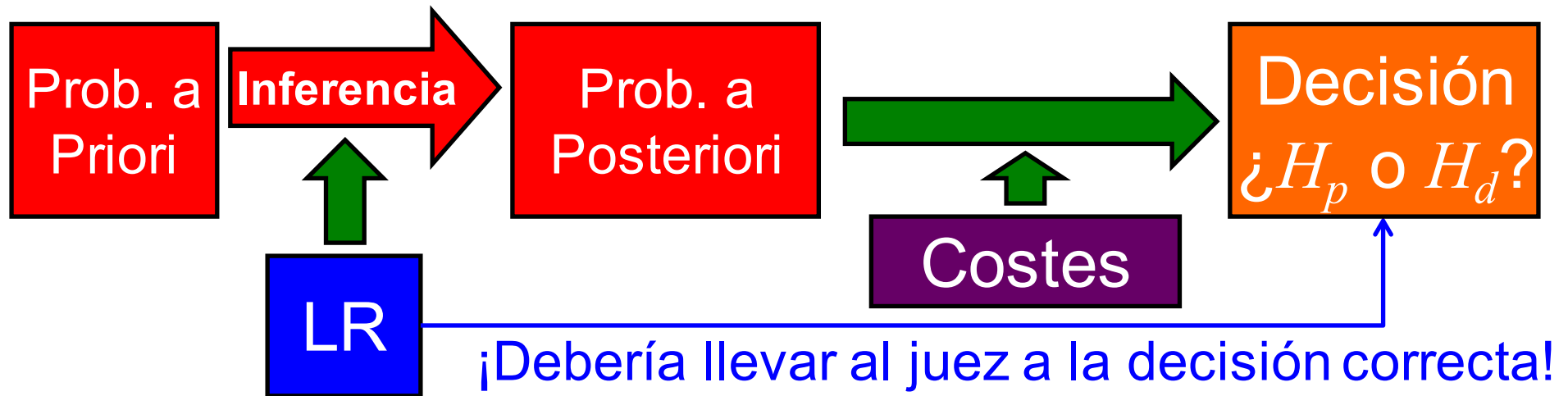
- En cualquier caso en el que entre en juego la prueba pericial, inevitablemente se debe tomar una decisión
  - No es competencia del perito
  - Pero está fuertemente influida por el perito a través del LR que calcula
- Por lo tanto, el perito no debe actuar sin tener en cuenta que es parte de ese proceso de decisión

# Decisión en un Caso: Consecuencias



- El perito debe calcular valores de LR que lleven a decisiones correctas
  - El LR debería apoyar  $H_p$  cuando  $H_p$  es realmente cierta
  - El LR debería apoyar  $H_d$  cuando  $H_d$  es realmente cierta
- Por tanto, si queremos saber si un LR es *bueno*...
  - Hemos de saber la “proposición correcta” en el caso
    - ¿Realmente ocurrió  $H_p$  o  $H_d$ ?

# Decisión en un Caso: Consecuencias



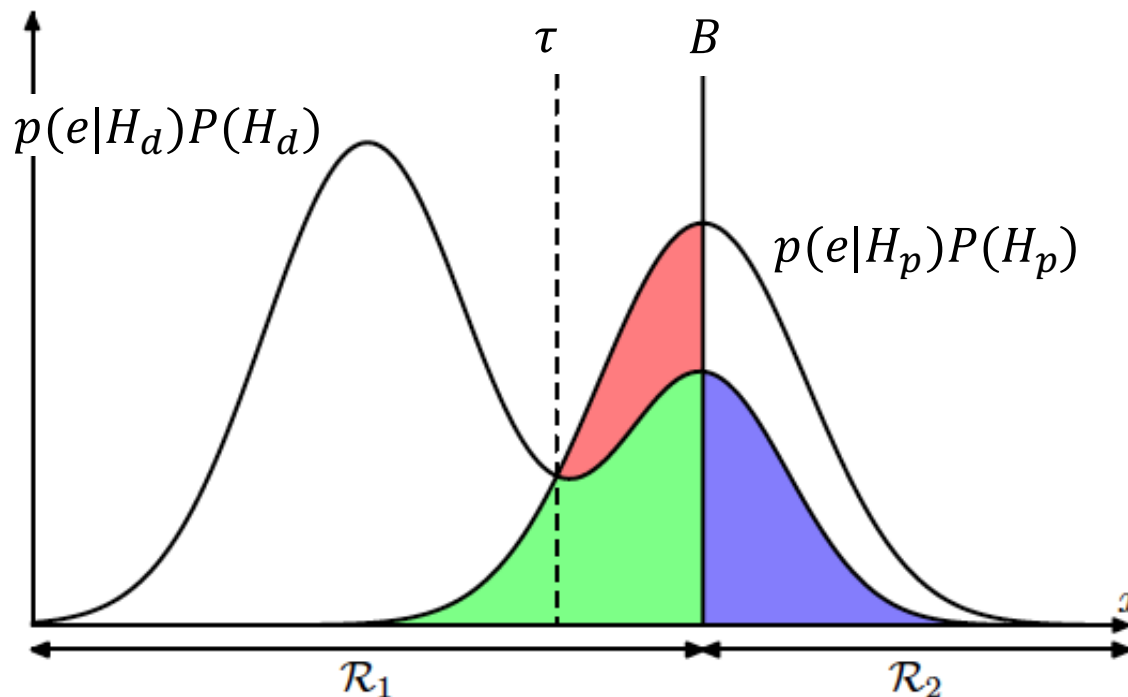
- Por tanto, ¿cuándo es *bueno* un valor de LR?
  - Cuando da lugar a probabilidades a posteriori cercanas a 1 cuando  $H_p$  es cierta
  - Cuando da lugar a probabilidades a posteriori cercanas a 0 cuando  $H_d$  es cierta
- Buscamos decisiones con el coste mínimo posible

# Regla Óptima de Decisión

- Umbral de decisión óptima ( $E = e$ : score)

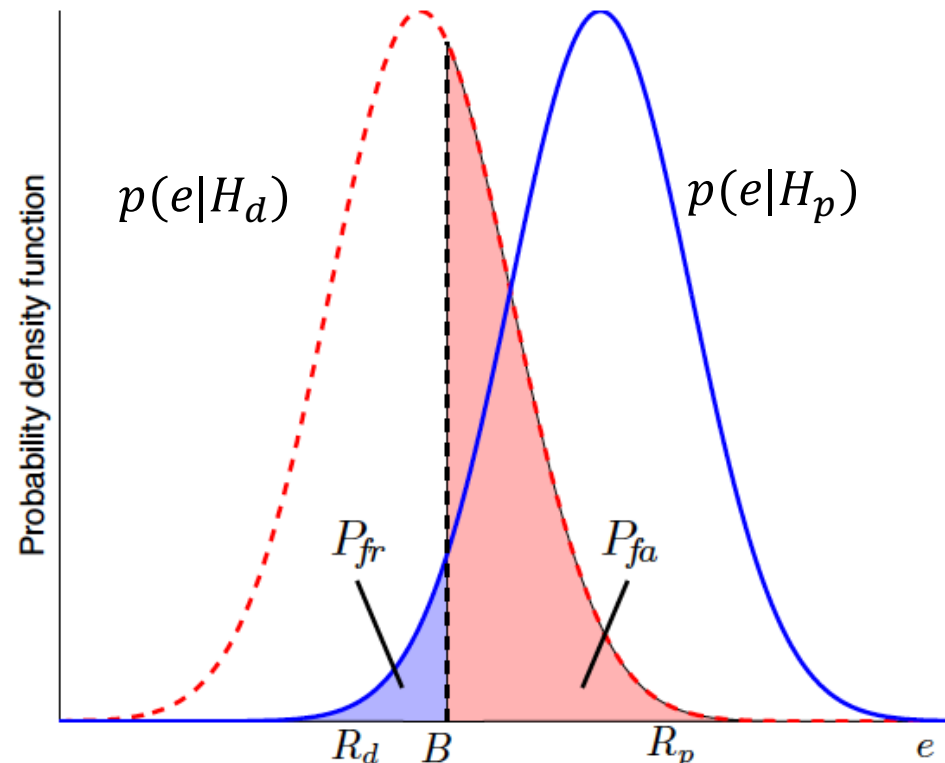
$$C_{fd}p(H_p|e) \leq C_{fd}p(H_d|e) \Rightarrow LR = \frac{p(e|H_p)}{p(e|H_d)} \leq \frac{p(H_p)C_{fd}}{p(H_d)C_{fp}} \equiv \tau$$

- $\tau$  : llamado “umbral de Bayes”
- Ejemplo  $C_{fp} = C_{fd}$ :



# Otra Forma de Verlo: Función de Coste

- Hay que minimizar la siguiente función de coste
- Cada umbral de decisión  $B$  dará lugar a un par de probabilidades de falsa aceptación y falso rechazo



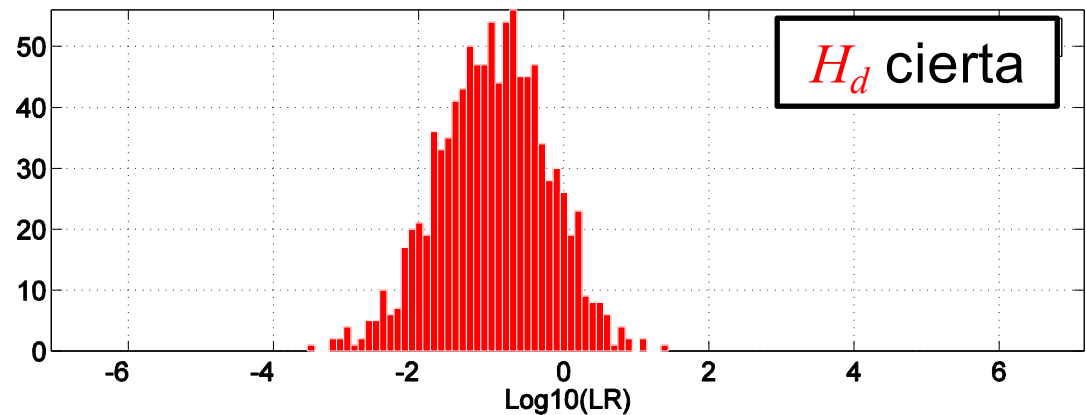
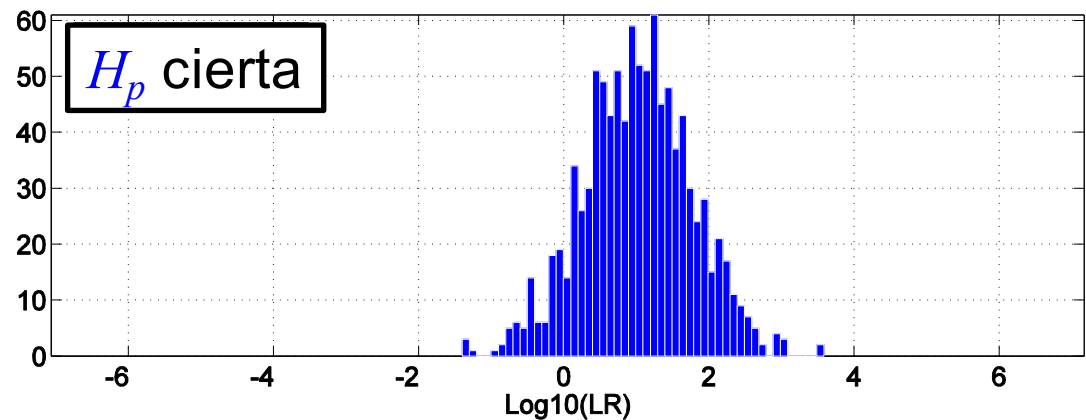
- Encontrar  $e^*$ : minimizar  $C_s = P_{fa}C_{fp}P(H_d) + P_{fr}C_{fd}P(H_p)$

# Validación Empírica

- Utilización de una base de datos (vidrios, locuciones, etc.)
  - ▣ Sabemos a qué fuente pertenece cada toma
    - Dubitada
    - Indubitada
  - ▣ Por tanto, sabemos las respuestas correctas
- Hacemos muchas comparaciones diferentes: generamos muchos valores de LR
  - ▣ Podemos separar los LR obtenidos en
    - LR para los que es cierta  $H_p$
    - LR para los que es cierta  $H_p$

# Validación Empírica

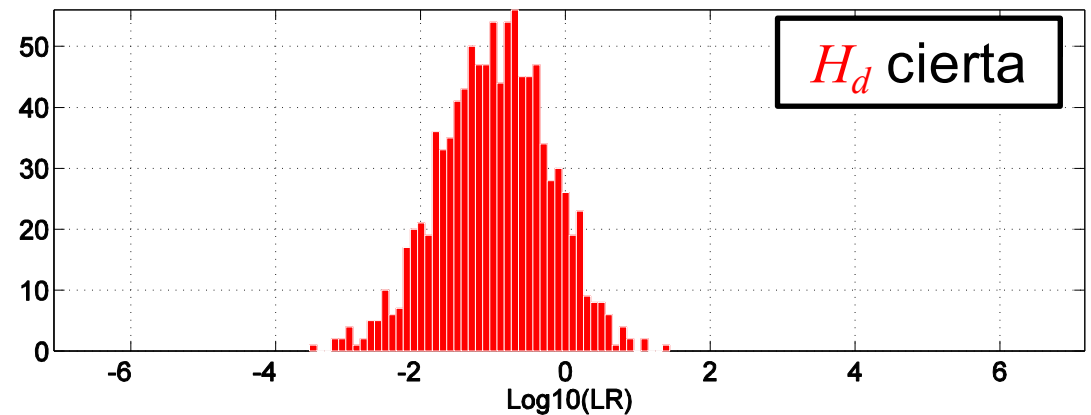
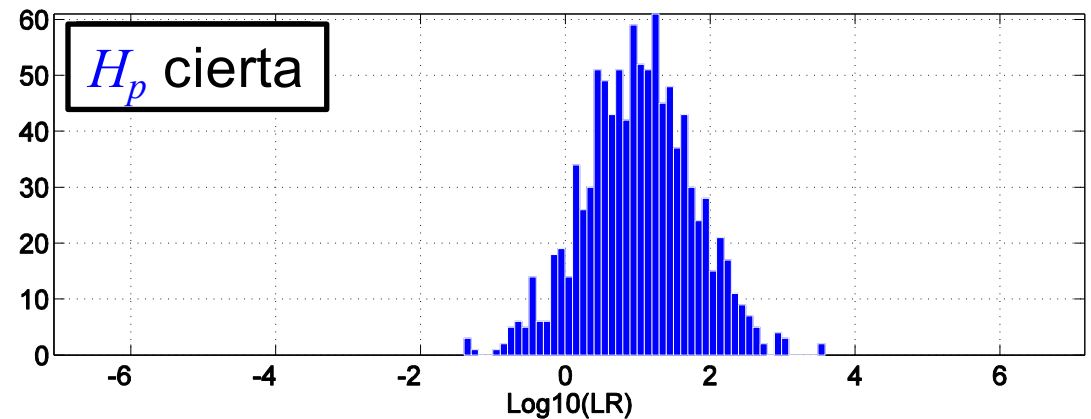
- Histogramas de LR generados respectivamente con  $H_p$  o  $H_d$  ciertas



# ¿Criterio de bondad?

- La separación entre ambos tipos de LR deseable

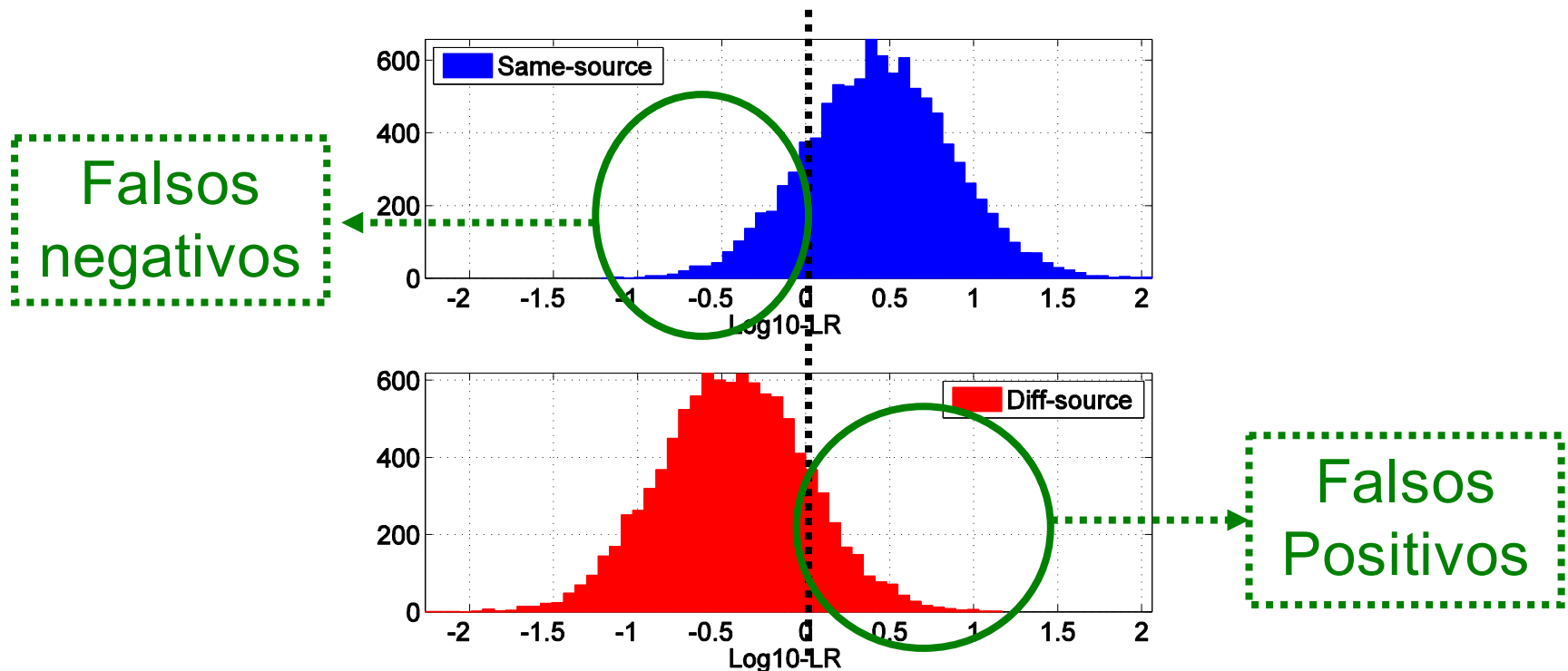
- Cuanto más separados, mejor distingue el método de evaluación entre casos en los que cada proposición es cierta





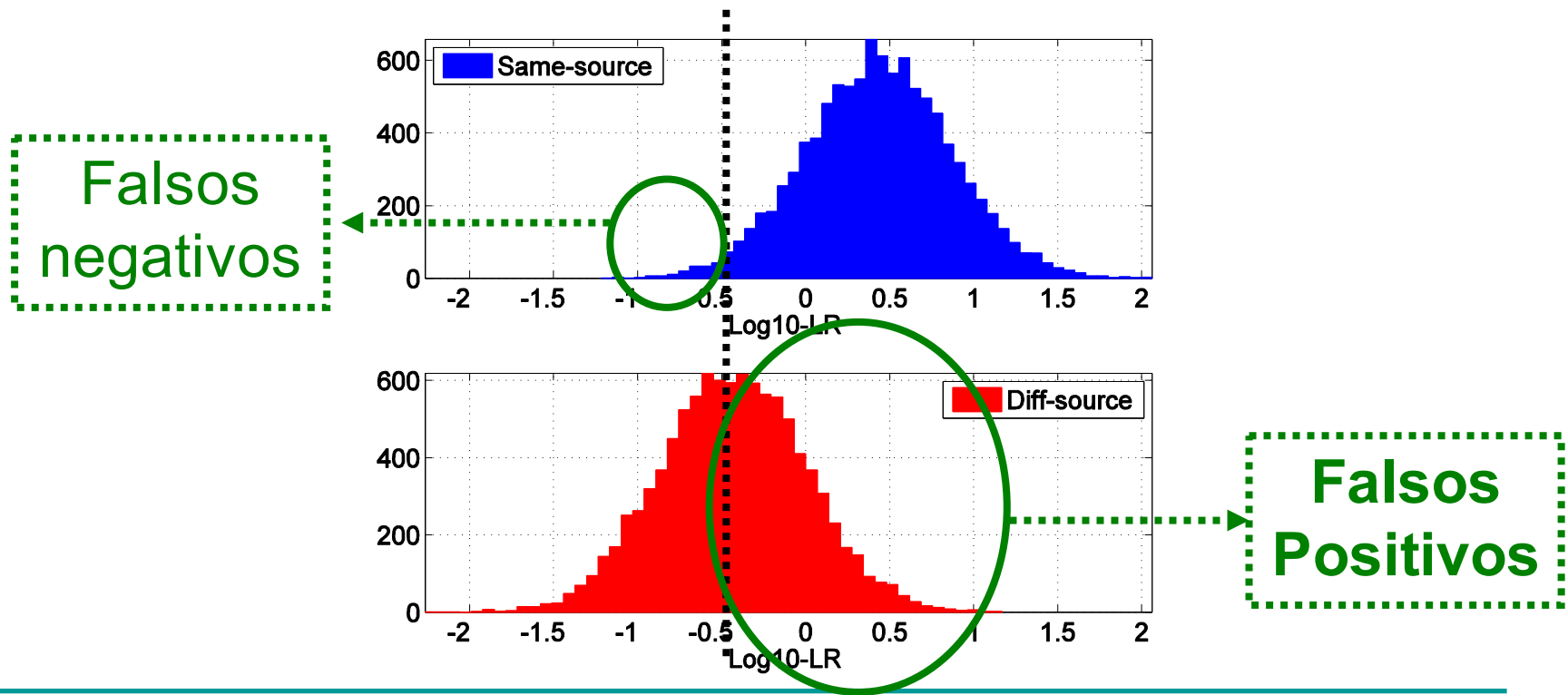
# Falsos Positivos y Falsos Negativos

- Medida de rendimiento clásica
- Se miden para valores concretos de los umbrales de decisión
  - Para  $\log(\text{LR}) = 0$  se llaman “tasas de evidencia errónea”
- Medida del poder de discriminación



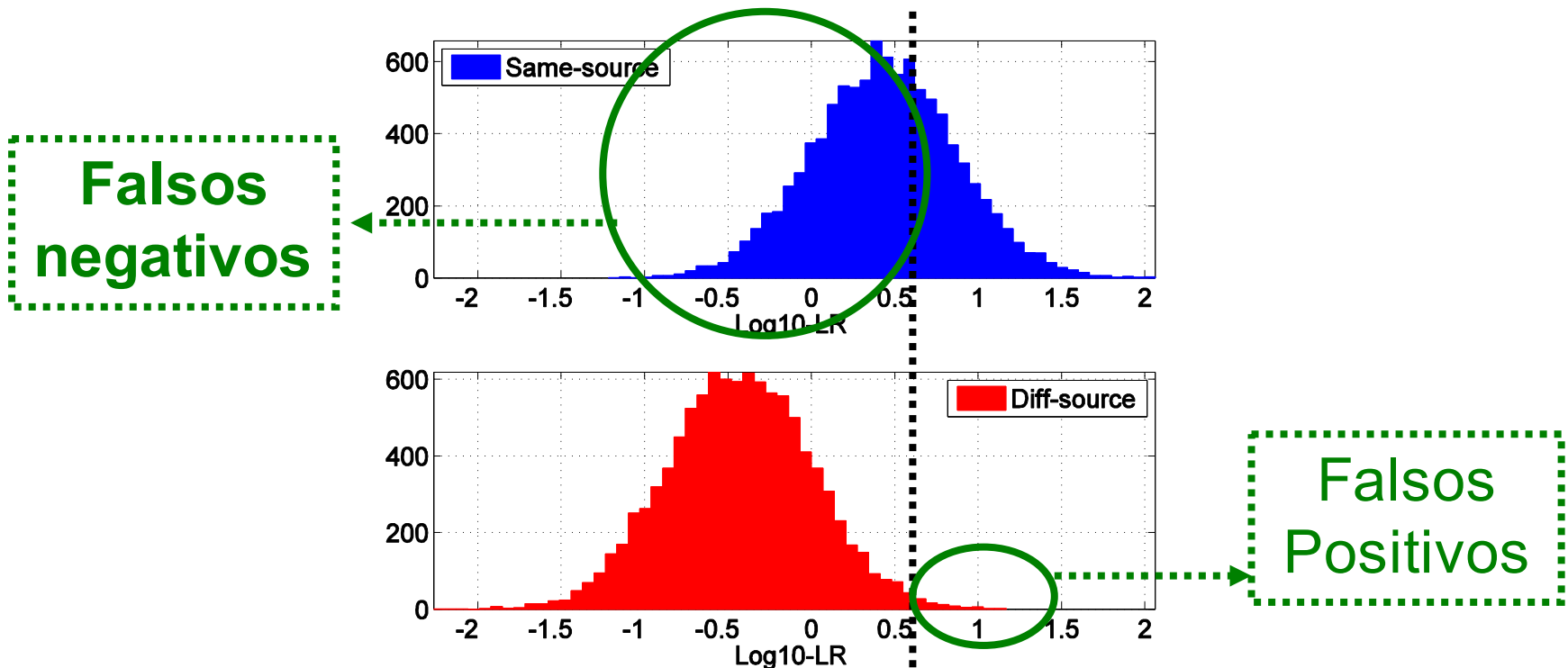
# Falsos Positivos y Falsos Negativos

- Medida de rendimiento clásica
- Se miden para valores concretos de los umbrales de decisión
  - Para  $\log(LR) = 0$  se llaman “tasas de evidencia errónea”
- Medida del poder de discriminación



# Falsos Positivos y Falsos Negativos

- Medida de rendimiento clásica
- Se miden para valores concretos de los umbrales de decisión
  - Para  $\log(\text{LR}) = 0$  se llaman “tasas de evidencia errónea”
- Medida del poder de discriminación





# Calibración

# Calibración

- Se tiene un conjunto de probabilidades a posteriori
  - Con sus correspondientes etiquetas (ground-truth)
- **Calibración** significa
  - $P(H_p|E)$  se aproxima a la proporción real de ocurrencia de  $H_p$  en el conjunto de probabilidades a posteriori

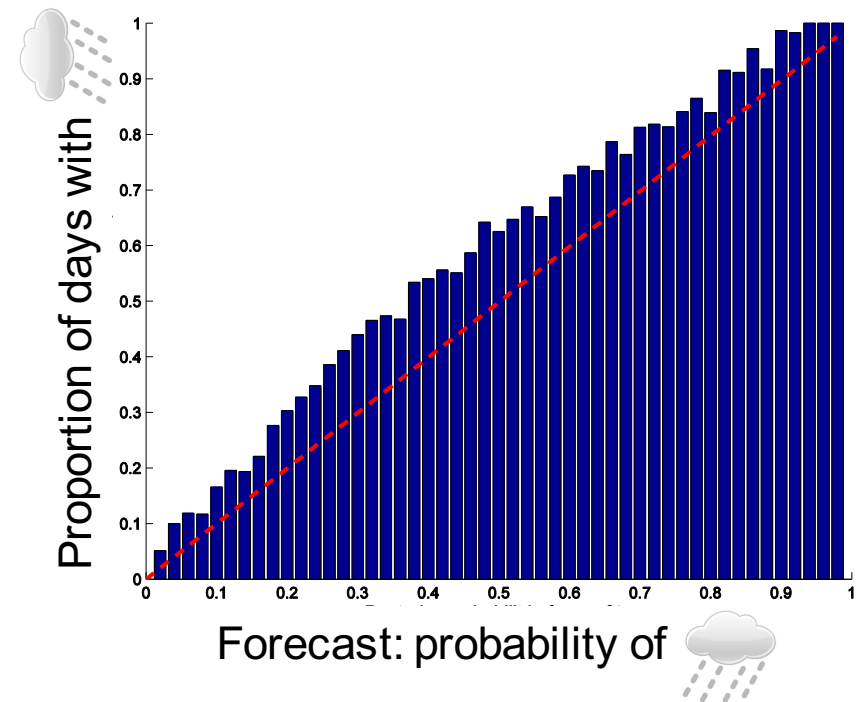
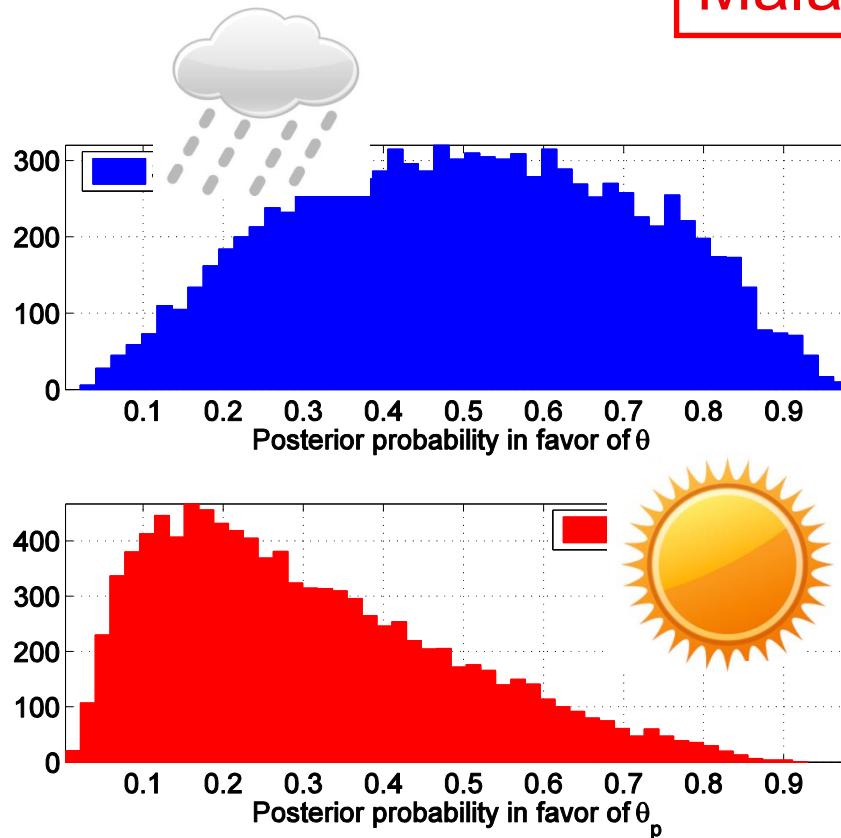
## LINDLEY, TVERSKY AND BROWN – *Reconciliation of Probability Assessments*

assessments in terms of a semantic criterion that pertains to the meaning of the probability scale. Clearly, there is no way of validating, for example, a meteorologist's single judgement that the probability of rain is 2/3. If the meteorologist is using the scale properly, however, we would expect that rain would occur on about two-thirds of the days to which he assigns a rain probability of 2/3. This criterion is called calibration. Formally, a person is calibrated if the proportion of correct statements, among those that were assigned the same probability, matches the stated probability, i.e. if his hit rate matches his confidence. If only half of the

# Calibración

- Ejemplo: conjunto experimental de probabilidades (de lluvia)
  - Separadas por el valor real (llovió, no llovió)

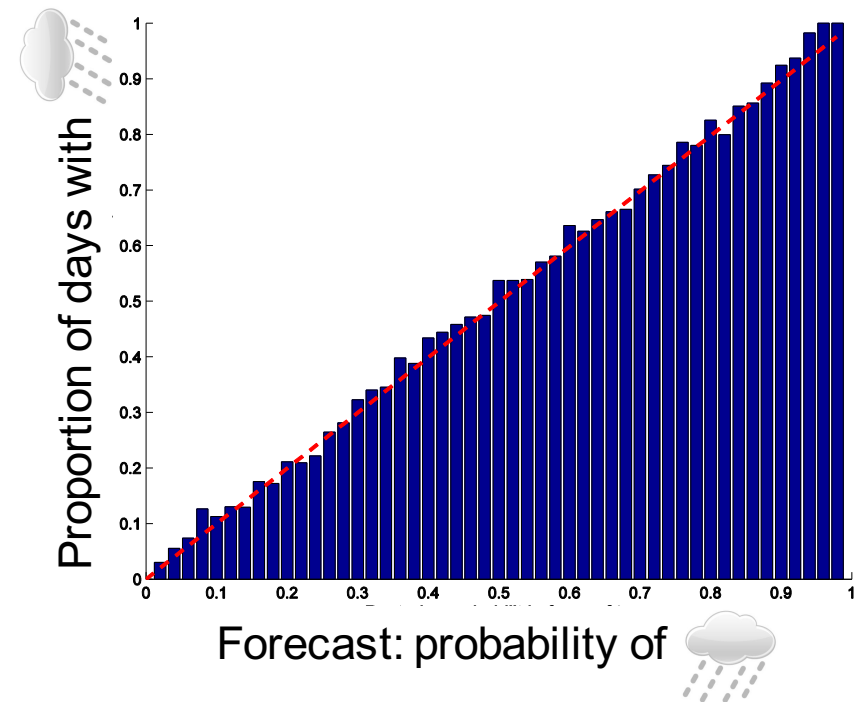
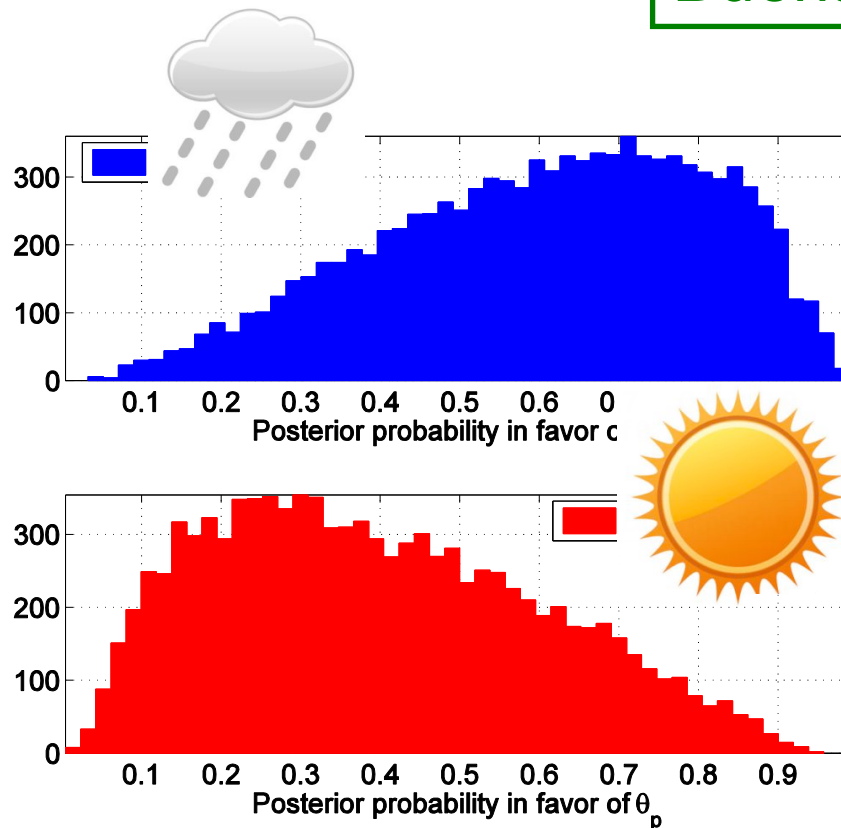
Mala Calibración



# Calibration

- Ejemplo: conjunto experimental de probabilidades (de lluvia)
  - Separadas por el valor real (llovió, no llovió)

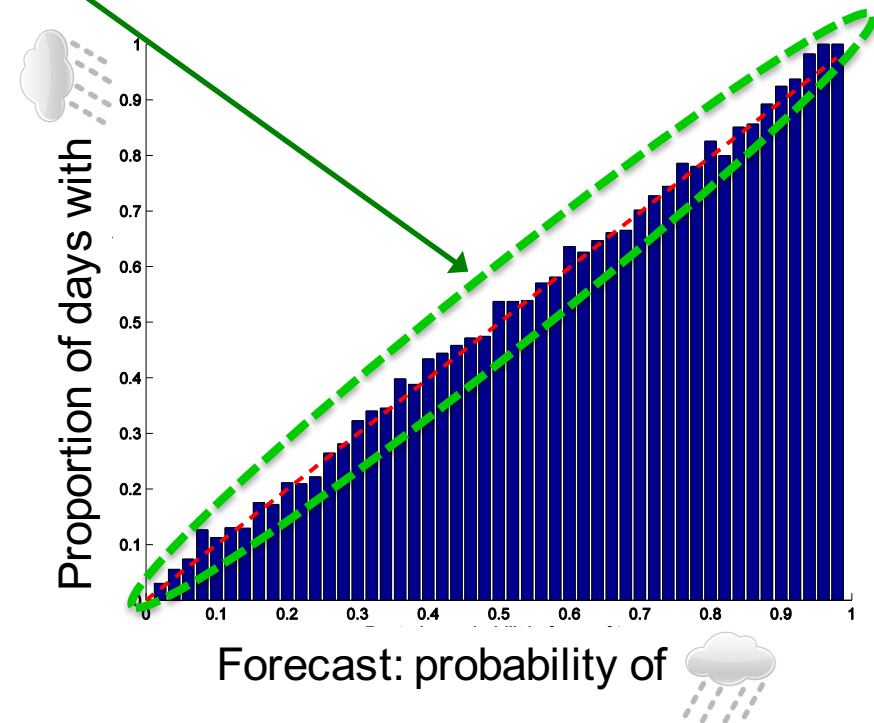
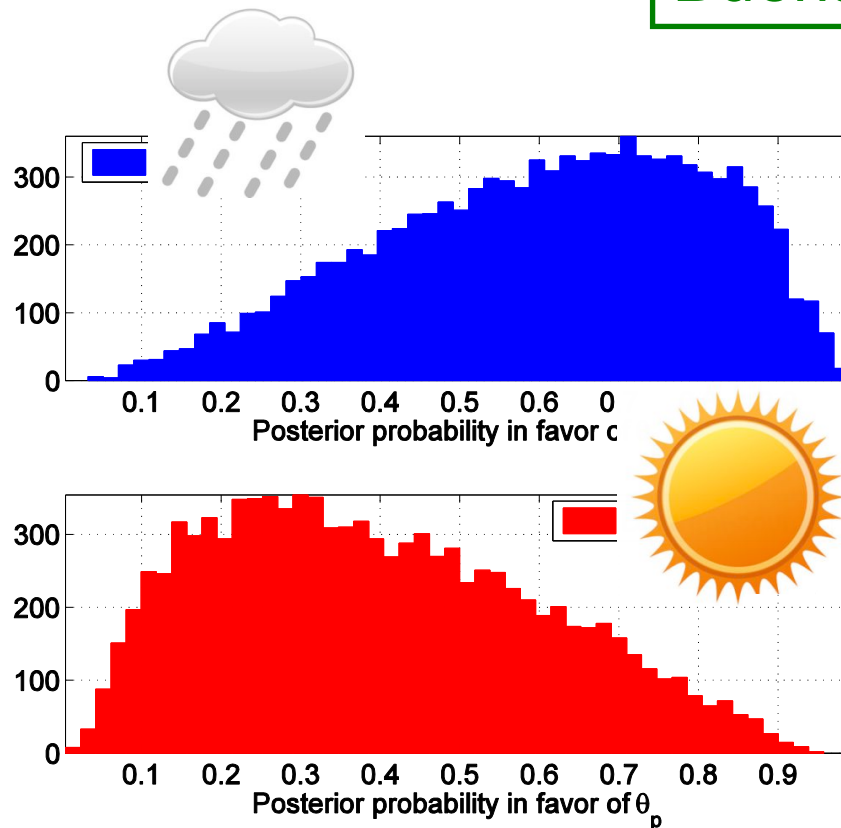
Buena Calibración



# Calibration

- Ejemplo: conjunto experimental de probabilidades (de lluvia)
  - Separadas por el valor real (llovió, no llovió)

Buena Calibración





# Propiedades de la Calibración de LRs

# Calibration y la Fuerza del LR

- Cuanto mejor es el poder de discriminación de un sistema
- Más fuerte tiende a ser el LR (valor de  $|\log(\text{LR})|$ )
- Y vice-versa

# Calibration y la Fuerza del LR

- Cuanto mejor es el poder de discriminación de un sistema
- Más fuerte tiende a ser el LR (valor de  $|\log(\text{LR})|$ )
- Y vice-versa
  - Si la calibración es buena, sólo métodos con alta discriminación podrán ofrecer altos valores de LR
  - Ejemplos:
    - ADN: generalmente arroja LR's muy fuertes
    - Voz: generalmente arroja LR's más moderados

# Calibration y la Fuerza del LR

- Cuanto mejor es el poder de discriminación de un sistema
- Más fuerte tiende a ser el LR (valor de  $|\log(\text{LR})|$ )
- Y vice-versa
  - Si la calibración es buena, sólo métodos con alta discriminación podrán ofrecer altos valores de LR
  - Ejemplos:
    - ADN: generalmente arroja LR's muy fuertes
    - Voz: generalmente arroja LR's más moderados
- Calibración: “Fiabilidad”
  - Gracias a esta y otras propiedades

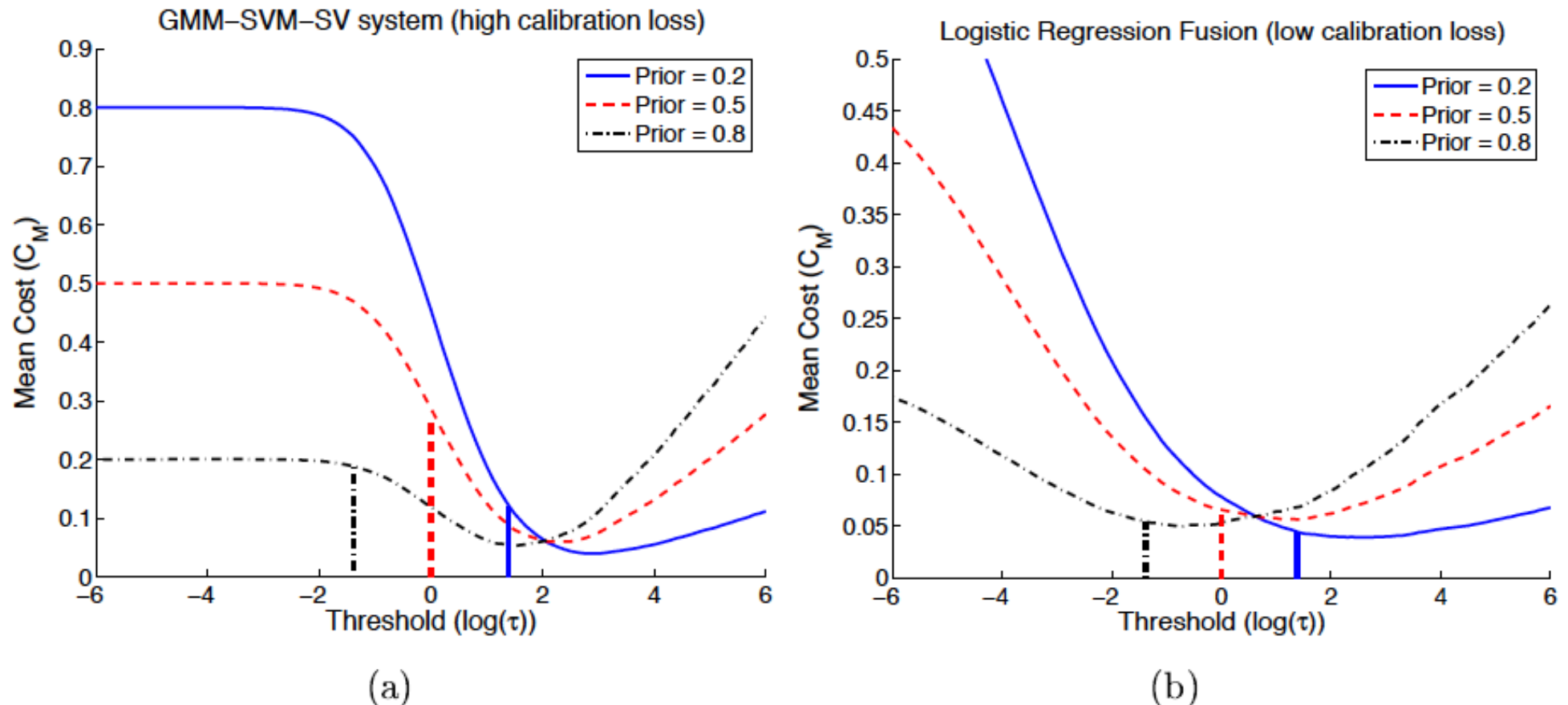
*The Statistician* 32 (1983)  
**The Comparison and Evaluation of Forecasters<sup>†</sup>**  
MORRIS H. DeGROOT and STEPHEN E. FIENBERG

Journal of the American Statistical Association  
September 1982, Volume 77, Number 379  
**The Well-Calibrated Bayesian**  
A. P. DAWID\*

# Calibración y Coste Mínimo

- Dos conjuntos de scores con una misma DET
  - Mismo poder de discriminación
- Y obtenemos  $C_s$  para múltiples valores de  $P_{fa}C_{fp}$ 
  - Usando el umbral de Bayes  $\tau$
- Conjunto de scores calibrados ( $\log(\text{LRs})$ ): siempre decisión óptima
- Conjunto de scores no calibrados ( $\log(\text{LRs})$ ): decisión subóptima

# Calibración y Coste Mínimo



**Figure 6.10:** Value of  $C_M$  (Equation 6.11) for different decision thresholds. (a) GMM-SVM-SuperVector system (calibration is not considered) and (b) Logistic regression fused system (calibration is considered).  $C_{fr} = C_{fa} = 1$ . Bayes thresholds (Equation 6.15) are shown as vertical lines.



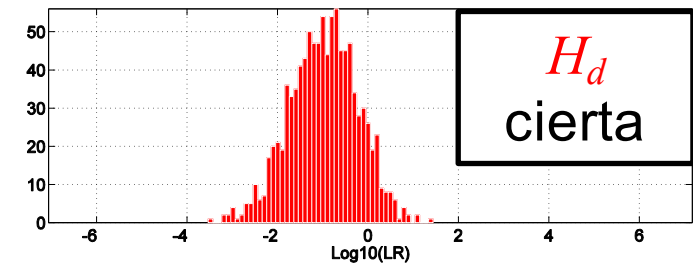
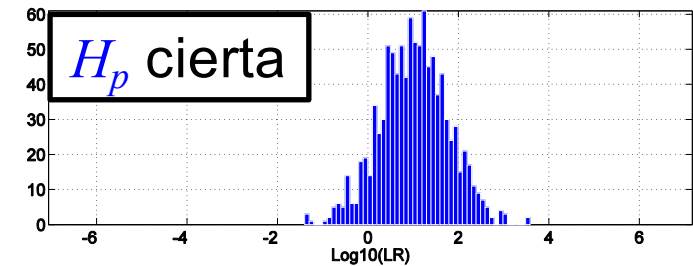
# Medida de Calibración Independiente de Aplicación

$C_{llr}$

- Entropía cruzada (cross-entropy)

- Particularizada en  $P(H_p)=0.5$

- Se parte de un conjunto experimental de valores de LR
- Se promedia una regla logarítmica asumiendo probabilidades a priori no informativas (iguales a 0,5)



- El número resultante se conoce como  $C_{llr}$

$$C_{llr} = \frac{1}{2 \cdot N_{p \text{ i of } H_p}} \sum \log_2 \left( 1 + \frac{1}{LR_i} \right) + \frac{1}{2 \cdot N_{d \text{ j of } H_d}} \sum \log_2 (1 + LR_j)$$

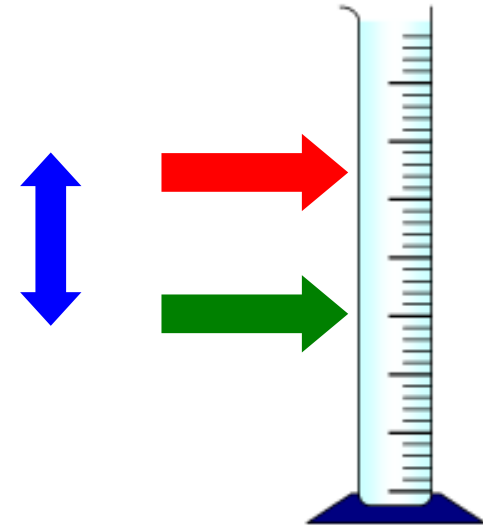


- $C_{llr}$ : medida de la bondad de los LR calculados
  - Valor numérico: cuanto más alto, peor el conjunto experimental de valores de LR
    - Permite ordenar la bondad de los métodos de forma objetiva
- $C_{llr}$  tiene propiedades interesantes
  - Se escapan del objetivo de este curso

Niko Brümmer<sup>a,b,\*</sup>, Johan du Preez<sup>b</sup>  
Application-independent evaluation of speaker detection  
Computer Speech and Language 20 (2006) 230–275

# Descomposición de $C_{llr}$

- Evaluación en dos pasos
- Primer paso: ¿discriminación?
  - ❑  $score$
  - ❑  $minC_{llr}$
- Segundo paso: ¿calibración?
  - ❑  $score$  tras calibración
  - ❑  $calC_{llr}$
- El rendimiento global del sistema será la suma de ambas:  $C_{llr}$ 
  - ❑ Para cualquier coste y prioris
- Descomposición posible gracias al algoritmo Pool Adjacent Violators (PAV)



# Probabilidad de Error

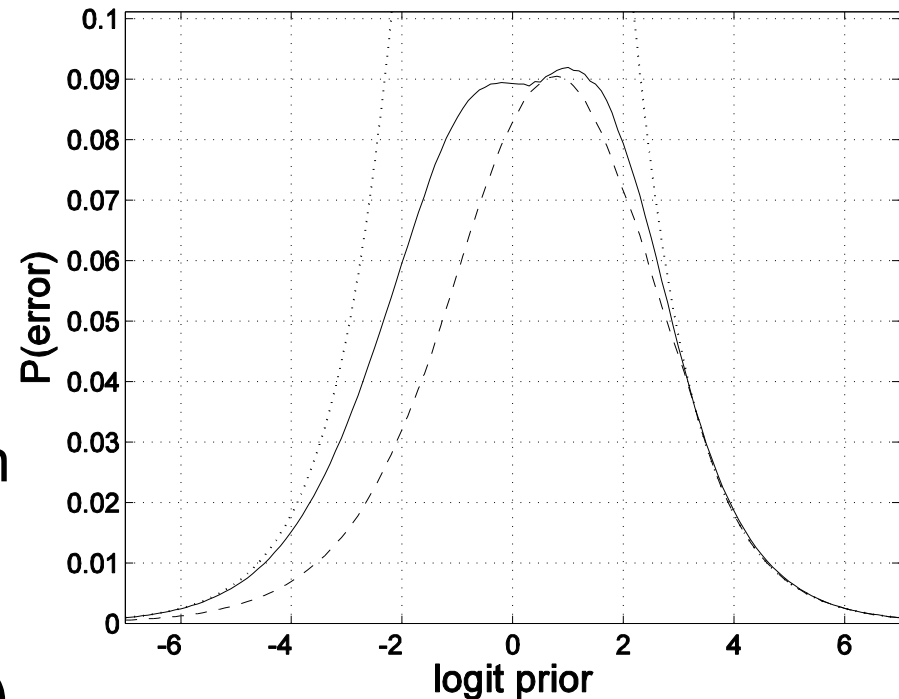
- Podemos representar la probabilidad de error media

$$E\{P_e\} = P(H_p) \cdot P(\text{error} | H_p) + P(H_d) \cdot P(\text{error} | H_p)$$

- Es una medida intuitiva
  - ▣ Si me equivoco más veces en media el sistema es peor
- Depende del umbral de decisión
  - ▣ Depende de la probabilidad priori
  - ▣ Los costes de decisión se asumen iguales a uno
    - Cada error cuenta igual

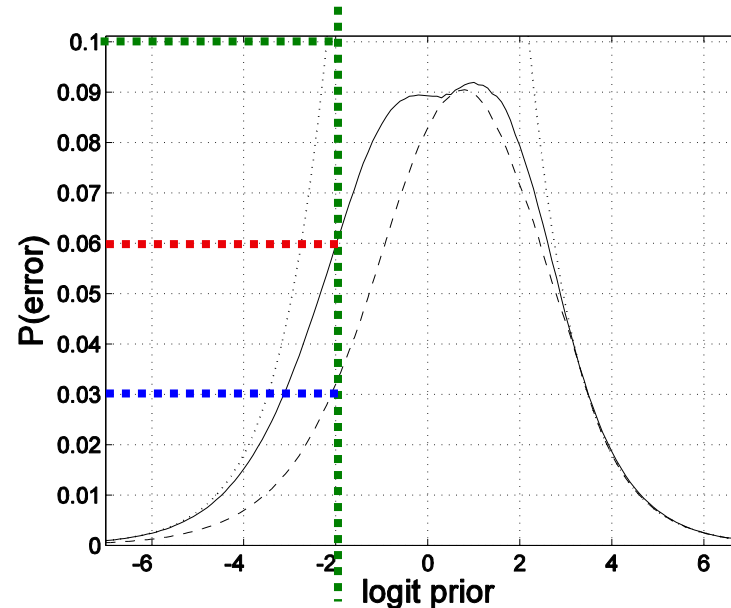
# Medir Discriminación y Calibración

- Esa representación es la curva APE
  - Applied Probability of Error
- Se representan
  - Probabilidad de error de los scores del sistema (sólida)
  - Probabilidad de error de los scores del sistema óptimamente calibrados con PAV (rayada)
  - Probabilidad de error de un sistema score=1 (punteada)
- No se fija el umbral (aplicación)
  - Se representa la probabilidad de error para cualquier umbral



# Medir Discriminación y Calibración

## ■ Interpretación de la APE

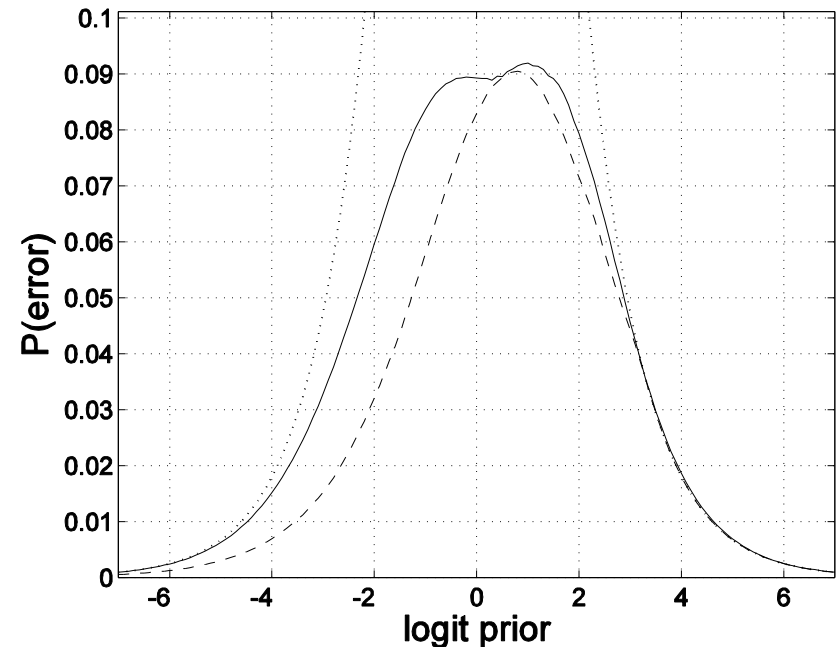


## ■ Para una aplicación dada...

- ❑ Si se usan los scores del sistema para tomar decisiones, me equivoco en media el 6% de las veces
- ❑ Si hubiese calibrado bien, el 3% de las veces
- ❑ Un sistema que no hace nada (score=1), el 10% de las veces

# Relación de la APE y $C_{llr}$

- Se puede demostrar que:
  - El área bajo la curva sólida de la APE es  $C_{llr}$
  - El área sobre la curva rayada de la APE es  $\min C_{llr}$
  - La diferencia entre ambas áreas es  $\text{cal}C_{llr}$
- Esta demostración confirma que minimizar  $C_{llr}$  nos lleva a decisiones mejores
- APE y  $C_{llr}$  suelen presentarse juntos
  - $C_{llr}$  es el valor escalar que resume la APE





# Calibración Extrínseca

# Detección Basada en “Scores”

- Arquitectura básica de los sistemas automáticos de reconocimiento de locutores: basados en *scores*
  - Ampliamente extendida
  - Especialmente en arquitecturas de tipo “caja negra”

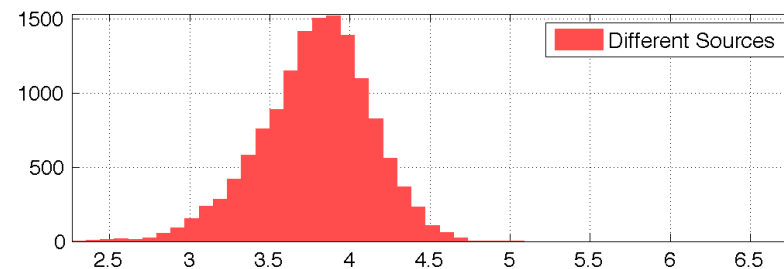
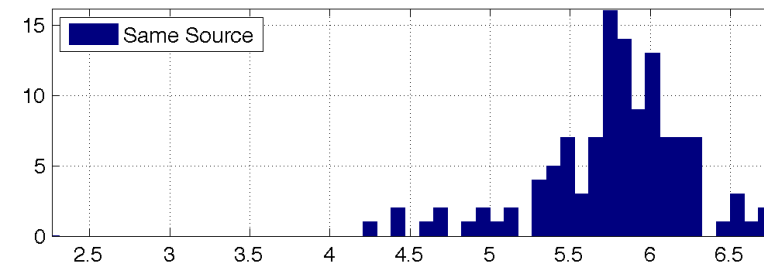
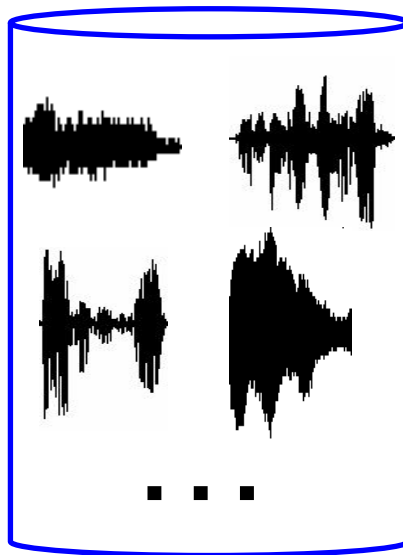


- Score: en general, única salida del sistema
  - No se puede interpretar directamente como un LR...



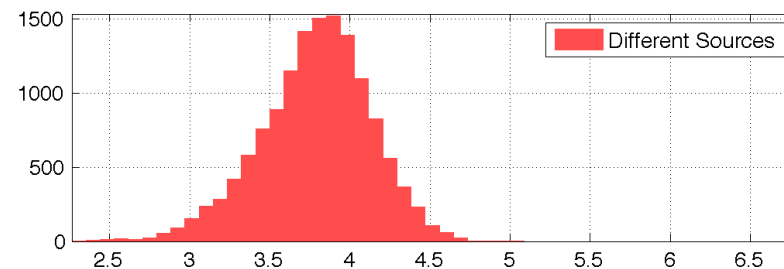
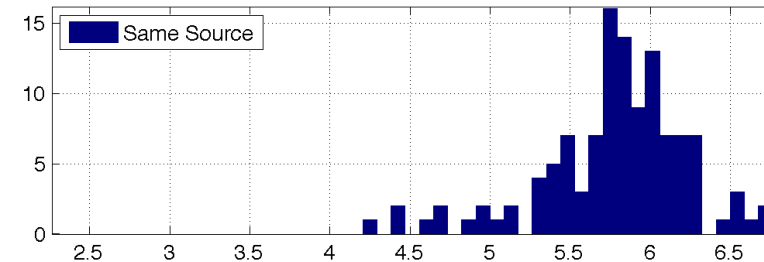
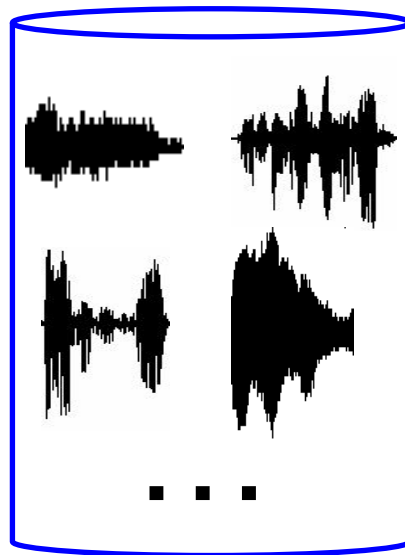
# Calibración Extrínseca: Ejemplo Simplificado

- Cálculo de LR: a partir de una base de datos sabemos que...



# Calibración Extrínseca: Ejemplo Simplificado

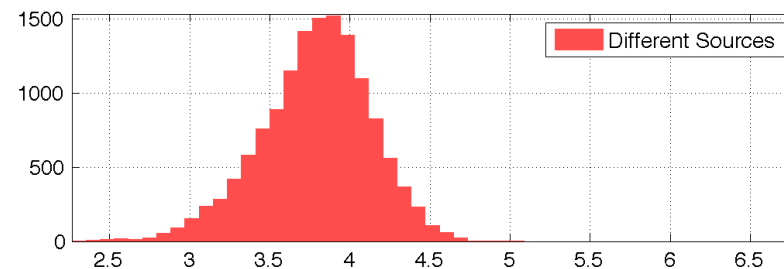
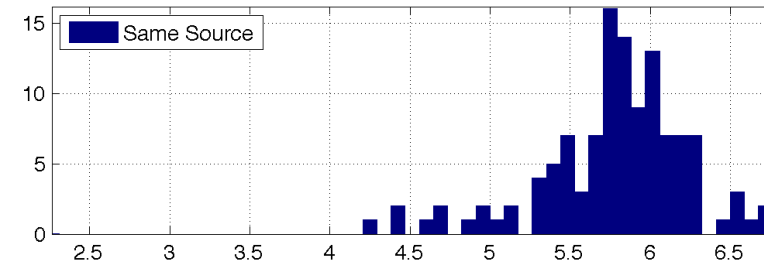
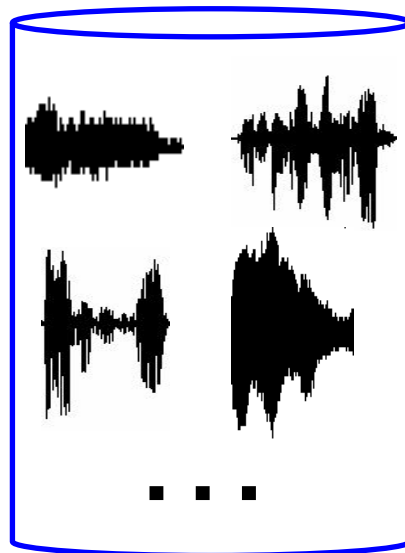
- Cálculo de LR: a partir de una base de datos sabemos que...



- Puntuaciones de comparaciones entre el “mismo individuo”
  - Rango entre 4 y 7

# Calibración Extrínseca: Ejemplo Simplificado

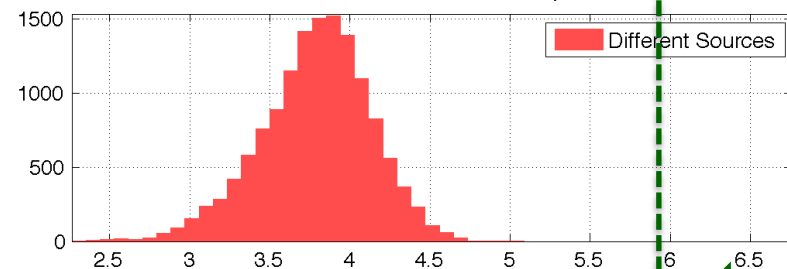
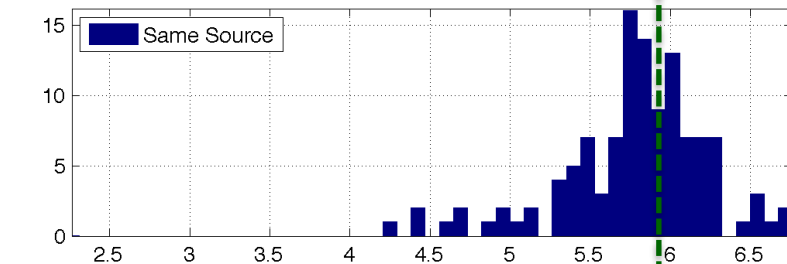
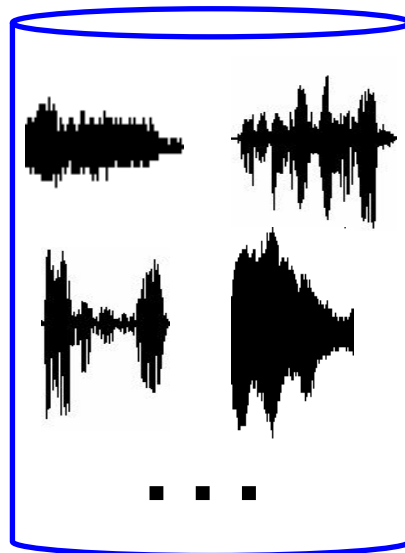
- Cálculo de LR: a partir de una base de datos sabemos que...



- Puntuaciones de comparaciones entre el “mismo individuo”
  - Rango entre 4 y 7
- Puntuaciones de comparaciones entre “individuos diferentes”
  - Rango entre 2 y 5

# Calibración Extrínseca: Ejemplo Simplificado

- Cálculo de LR: a partir de una base de datos sabemos que...

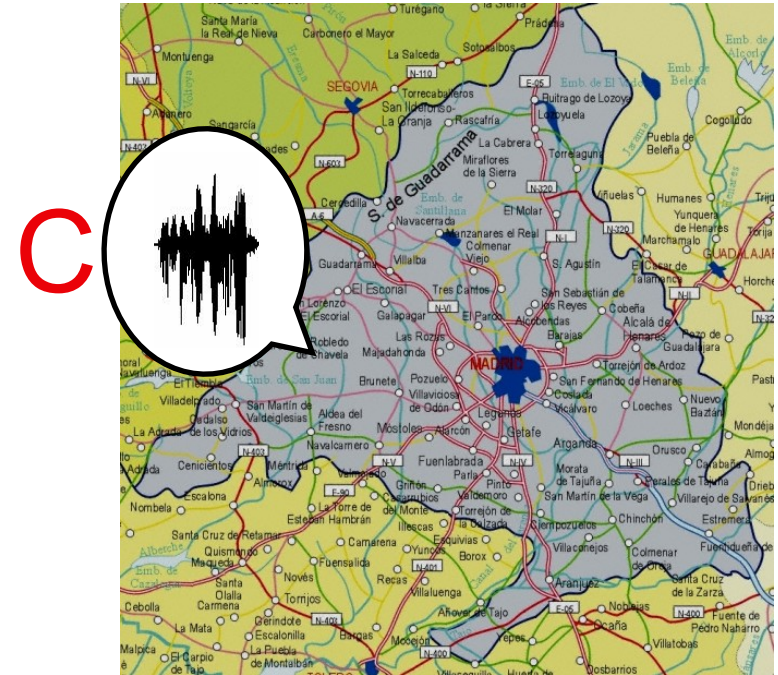


- Puntuaciones de comparaciones entre el “mismo individuo”
  - Rango entre 4 y 7
- Puntuaciones de comparaciones entre “individuos diferentes”
  - Rango entre 2 y 5
- Puntuación = 5,96: fuerte apoyo a que sea del “mismo individuo”

Ejemplo ilustrativo:  
caso simulado  
(y muy simplificado)

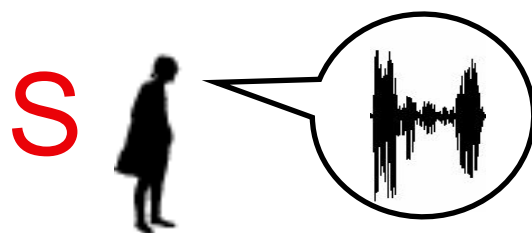
# Caso simulado

- Grabaciones incriminatorias tomadas en la Comunidad Autónoma de Madrid
  - Población: potenciales criminales
    - Hablantes de Madrid con características similares al hablante de la toma dubitada
      - Idioma
      - Acento
      - ...
    - Sistema: GSM grabado en cinta magnetofónica
- Las investigaciones policiales llevan a la detención de un sospechoso



## Caso simulado

- Se realizan grabaciones del sospechoso (voz indubitada)



- En principio, la abundancia y control sobre las grabaciones suele ser mayor que en la toma dubitada
  - Pero posiblemente en condiciones muy diferentes a la toma dubitada
- Puede haber incluso pinchazos no incriminatorios de los cuales el sospechoso reconoce la autoría
  - Condiciones similares a la toma dubitada
- El juez le pide al perito:
  - Que evalúe la evidencia
  - Que le informe de la precisión de las técnicas utilizadas

# Cálculo del LR (Calibración Extrínseca)

- Paso 1: el sistema automático calcula un score
  - Sin valor por sí mismo
    - ¿10 con respecto a qué?
  - En general, no interpretable
    - A priori, no conocemos su rango de variación





# Cálculo del LR (Calibración Extrínseca)

- Paso 1: el sistema automático calcula un score

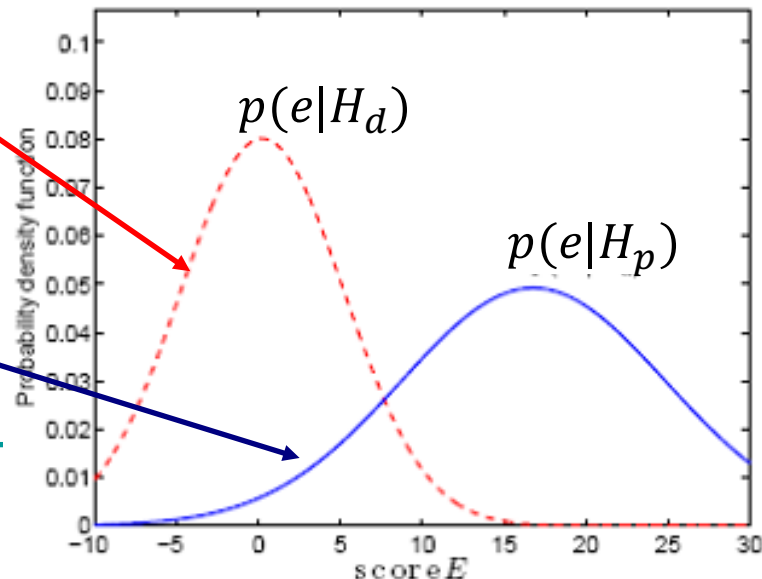
- Sin valor por sí mismo
  - ¿10 con respecto a qué?
- En general, no interpretable
  - A priori, no conocemos su rango de variación



- Paso 2: cálculo del LR
  - En este ejemplo usamos modelado gaussiano

Intervariabilidad  
(población)

Intravariabilidad  
(sospechoso)



< audias >

# Cálculo del LR (Calibración Extrínseca)

- Paso 1: el sistema automático calcula un score

- Sin valor por sí mismo
  - ¿10 con respecto a qué?
- En general, no interpretable
  - A priori, no conocemos su rango de variación

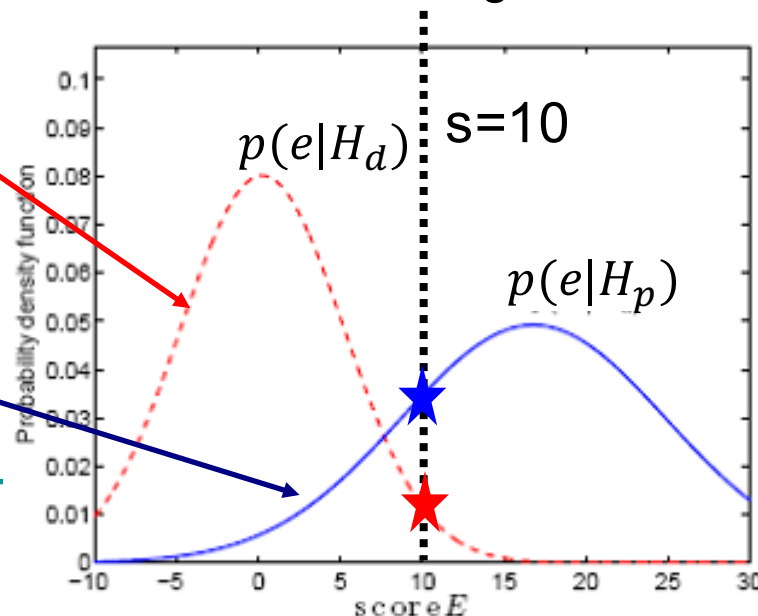


- Paso 2: cálculo del LR
  - En este ejemplo usamos modelado gaussiano

Intervariabilidad  
(población)

Intravariabilidad  
(sospechoso)

< audias >



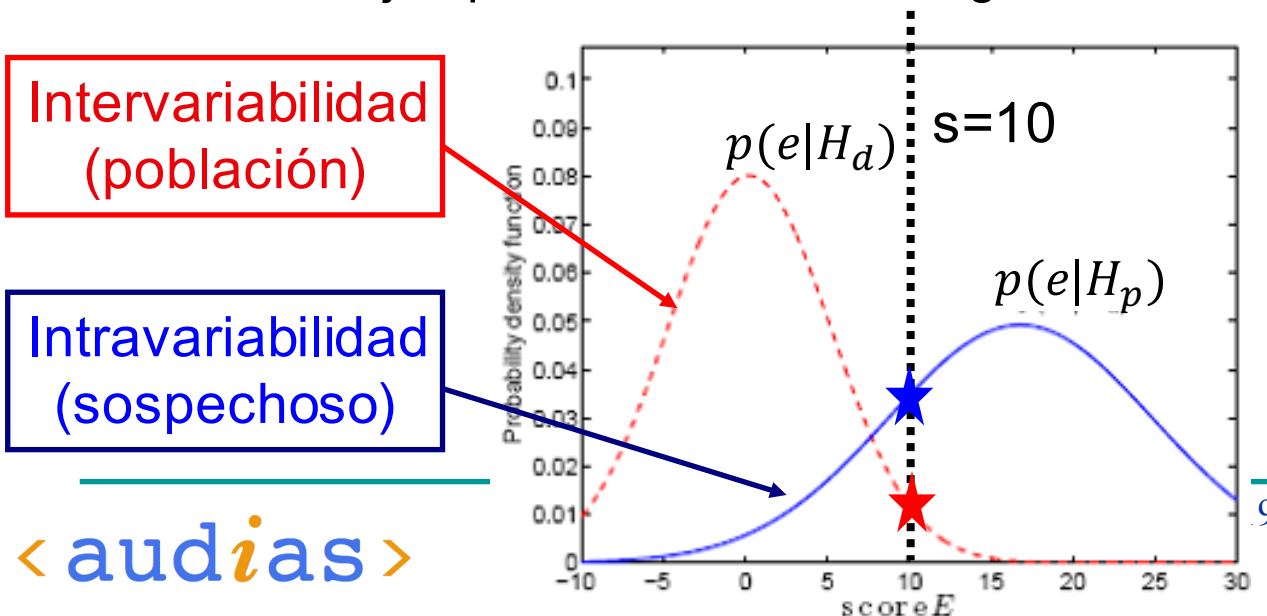
# Cálculo del LR (Calibración Extrínseca)

- Paso 1: el sistema automático calcula un score

- Sin valor por sí mismo
  - ¿10 con respecto a qué?
- En general, no interpretable
  - A priori, no conocemos su rango de variación



- Paso 2: cálculo del LR
  - En este ejemplo usamos modelado gaussiano

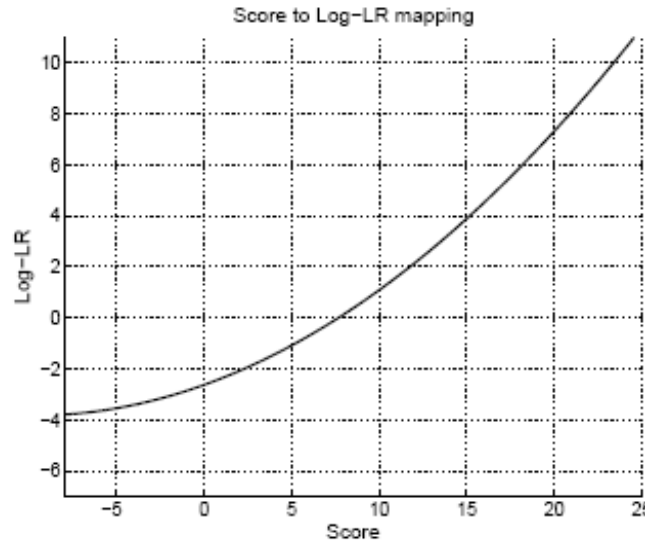
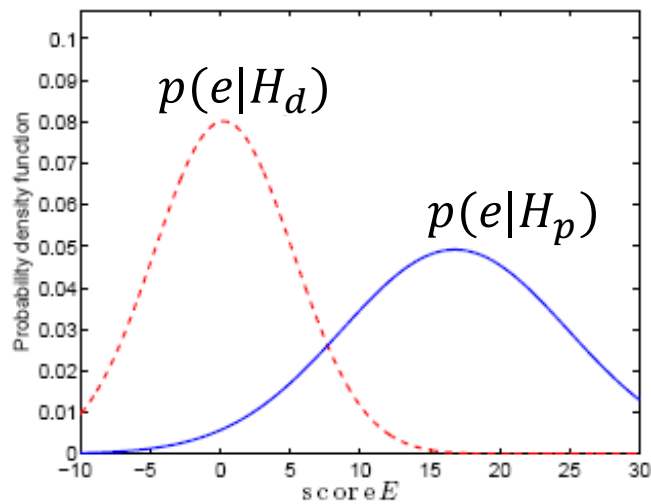


$$LR = \frac{0,35}{0,15} = 2,33$$

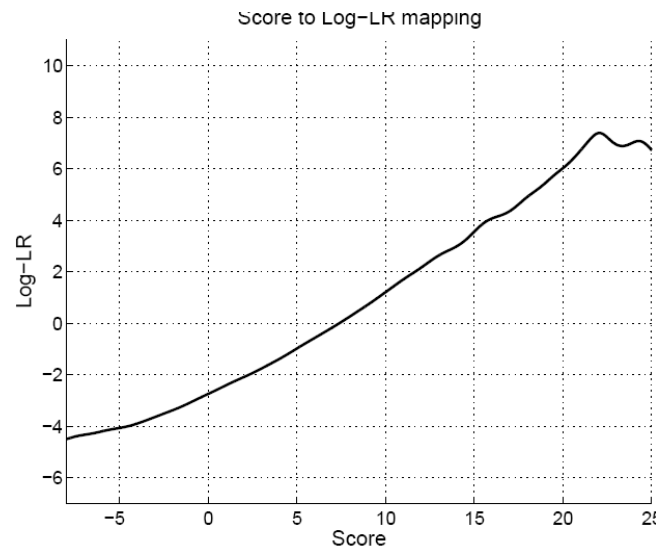
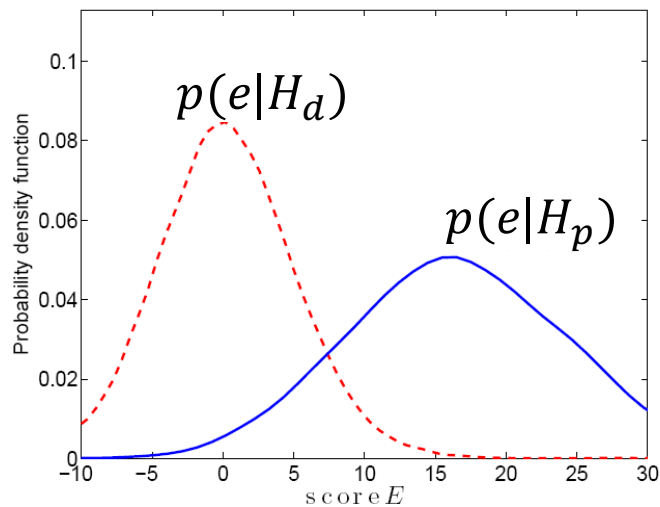
Apoyo 2,33 a 1  
a la hipótesis  $\theta_p$   
("misma fuente")

# Calibración Extrínseca

## ■ Técnicas generativas



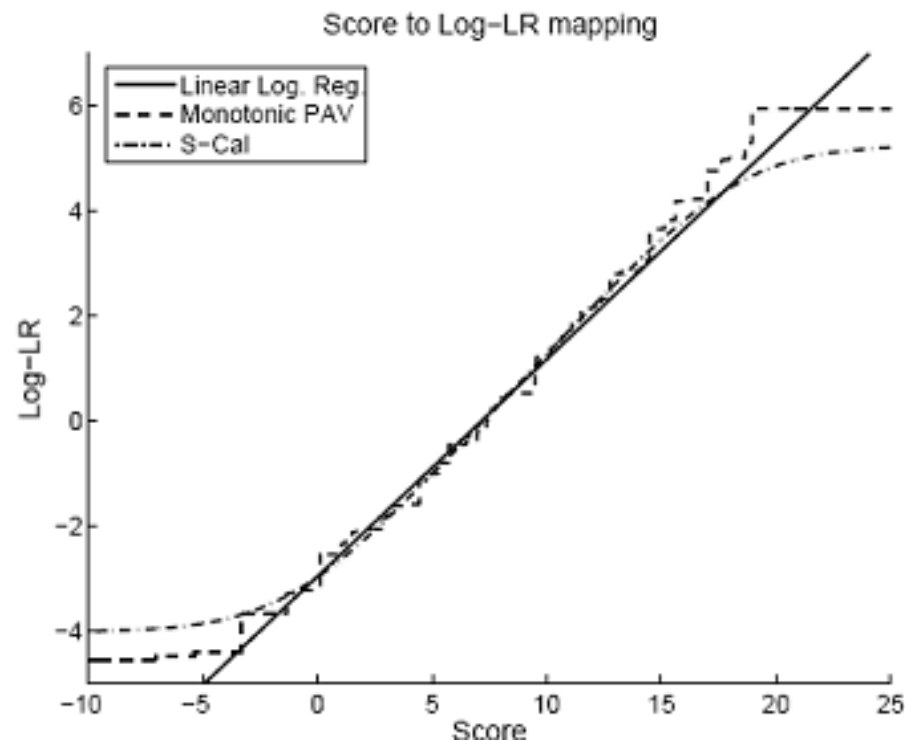
Modelado gaussiano



Kernel Density (KDF)

# Calibración Extrínseca

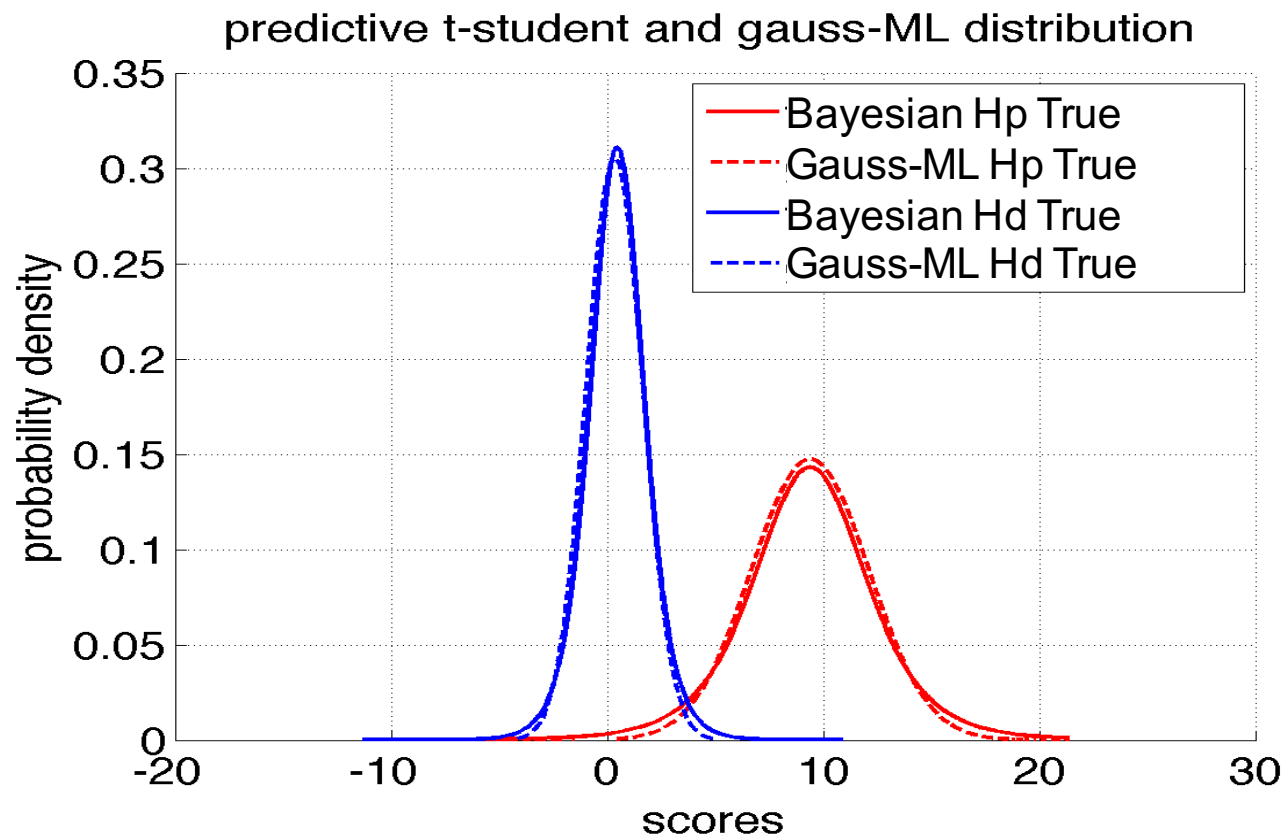
## ■ Otras técnicas discriminativas



- ❑ Regresión logística (lineal)
- ❑ S-cal
- ❑ Pool Adjacent Violators (PAV)

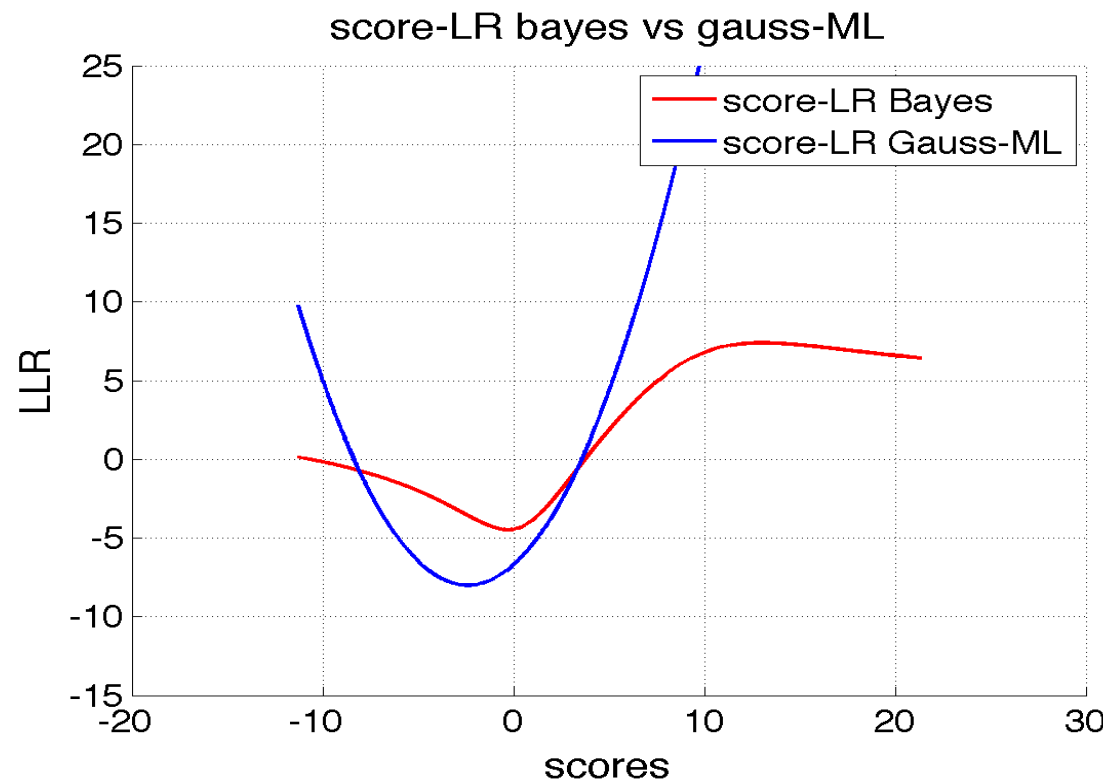
# *Maximum Likelihood* vs. Bayesiano

- Métodos bayesianos más adecuados con pocos datos



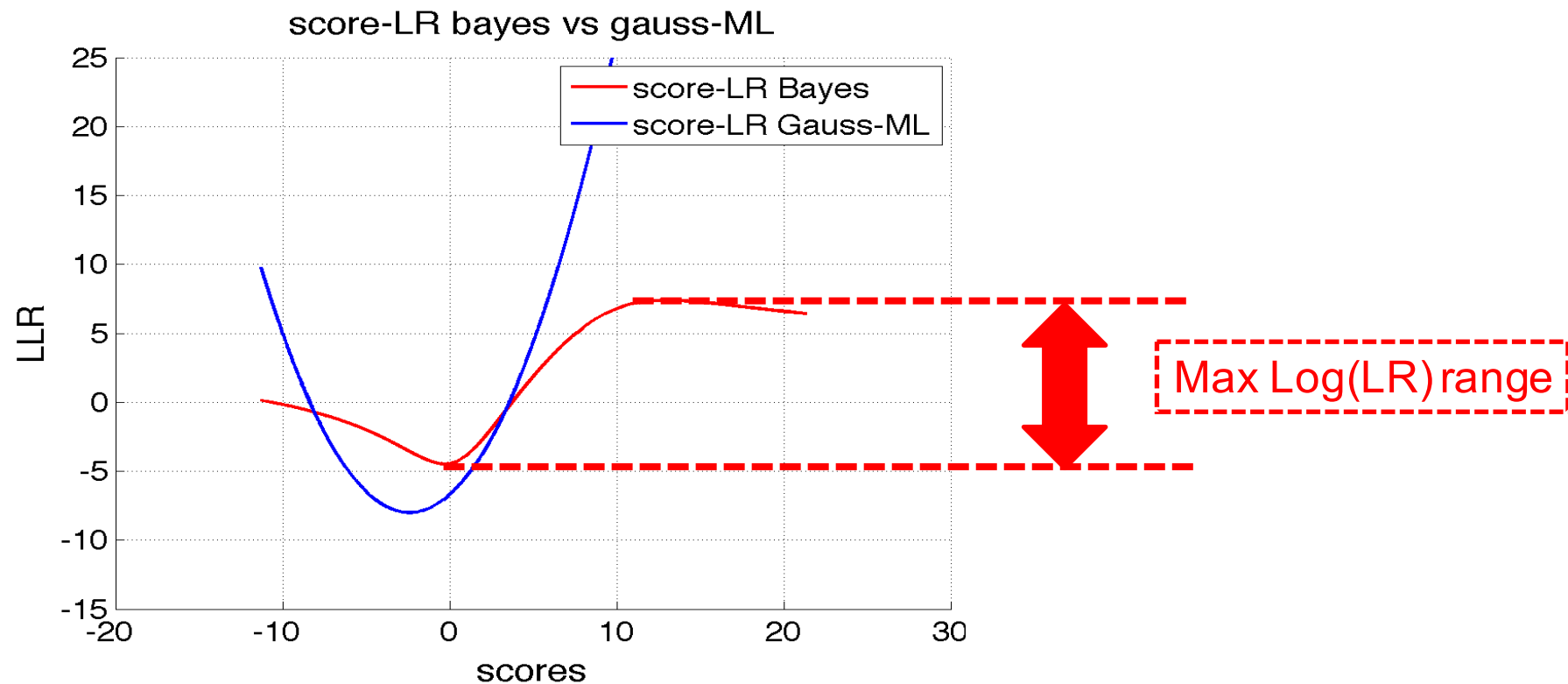
# *Maximum Likelihood* vs. Bayesiano

- Modelo Bayesiano limita la fuerza del LR
  - Tiene en cuenta la falta de datos (incertidumbre)



# *Maximum Likelihood* vs. Bayesiano

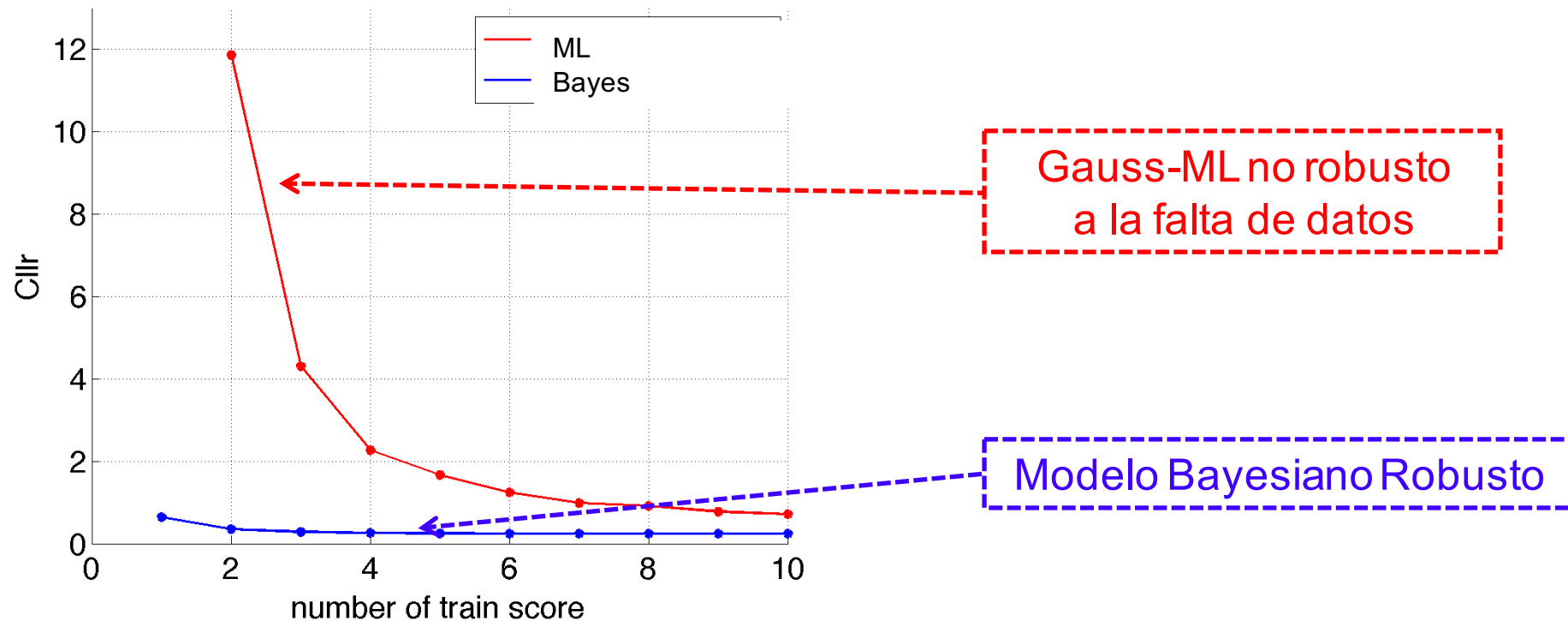
- Modelo Bayesiano limita la fuerza del LR
  - Tiene en cuenta la falta de datos (incertidumbre)





# Maximum Likelihood vs. Bayesiano

- Cllr: cuanto menor, mejor



# Calibración Extrínseca: Referencias

- Robustez frente a falta de datos

D. Ramos-Castro, J. Gonzalez-Rodriguez, A. Montero-Asenjo and J. Ortega-Garcia, "Suspect-adapted MAP estimation of within-source distributions in generative likelihood ratio estimation", IEEE Odyssey 2006.

- Calibración de scores (extrínseca) no supervisada

Niko Brummer and Daniel Garcia-Romero, "Generative Modelling for Unsupervised Score Calibration", ICASSP 2014.

- Calibración extrínseca bayesiana

Niko Brummer and Albert Swart, 'Bayesian calibration for forensic evidence reporting', Interspeech 2014.

D. Ramos et al., 'Bayesian strategies for Likelihood Ratio computation in forensic voice comparison with automatic systems.', Subsidia 2017.

- Análisis de la distribución de scores calibrados

David van Leeuwen, Niko Brummer, "The distribution of calibrated likelihood-ratios in speaker recognition", Interspeech 2013.

- NNs para calibración de LRs

W. Campbell et al., "Estimating and Evaluating Confidence for Forensic Speaker Recognition", ICASSP 2005.

# Calibración Extrínseca Multiclase

# Calibración Extrínseca: Referencias

- Calibración de DNNs

|  |
|--|
| <b>On Calibration of Modern Neural Networks</b>  |
| Chuan Guo <sup>*1</sup> Geoff Pleiss <sup>*1</sup> Yu Sun <sup>*1</sup> Kilian Q. Weinberger <sup>1</sup>  |
| <i>Proceedings of the 34<sup>th</sup> International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).</i> |

- Extrínseca

- Se toman las salidas de la DNN
- Se transforman utilizando un algoritmo sencillo
- Mejor opción: *Temperature Scaling*



# Clasificadores Probabilísticos en Aprendizaje Automático

## Calibración y Rendimiento Independiente de Aplicación

**Daniel Ramos Castro**

Contribuciones de Segio Álvarez Balanya (Estudiante de Máster UAM)

[daniel.ramos@uam.es](mailto:daniel.ramos@uam.es)

Audias – Audio, Data Intelligence and Speech  
Universidad Autónoma de Madrid

<http://audias.ii.uam.es>

< audias >

Audio, Data Intelligence and Speech

UAM