Reliable Support: Measuring Calibration of Likelihood Ratios

Daniel Ramos

ATVS – Biometric Recognition Group

Research Institute of Forensic Science and Security

Universidad Autonoma de Madrid

daniel.ramos@uam.es

http://arantxa.ii.uam.es/~dramos

Performance of Likelihood Ratios

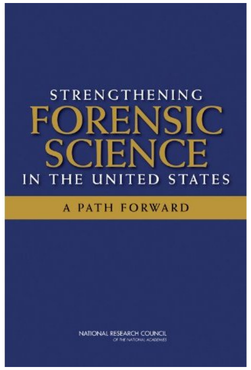
Performance Assessment in Forensic Science

- Scientific assessment of the performance of any processes involved in forensic science
  - Critical, increasing importance since Daubert rules
  - Recent and well-known references claiming/implementing it

Science AAAS

The Coming Paradigm Shift in Forensic Identification Science

Michael J. Saks[1] and Jonathan J. Koehler[2]



PNAS

Accuracy and reliability of forensic latent fingerprint decisions

Bradford T. Ulery[a], R. Austin Hicklin[a], JoAnn Buscaglia[b,1], and Maria Antonia Roberts[c]

THE ADMISSIBILITY OF EXPERT EVIDENCE IN CRIMINAL PROCEEDINGS IN ENGLAND AND WALES

A New Approach to the Determination of Evidentiary Reliability

Performance Assessment in Forensic Science

- Performance of evidence evaluation methods should be measured
- Value of the evidence: Likelihood Ratio
  - Increasingly supported in Europe

Science and Justice

Guest editorial

Expressing evaluative opinions: A position statement

The Board of the European Network of Forensic Science Institutes (ENFSI) also supports this position statement and engages itself to work towards a full implementation within the ENFSI laboratories (ENFSI has 58 member institutes in 33 countries).

Dr. Jan De Kinder, Chairman
Paweł Rybicki, Chairman designate
Tore Olsson, Member
Burhanettin Cihangiroğlu, Member
Dr. Torsten Ahlhorn, Member

ENFSI

- Standards to be defined for all ENFSI laboratories

KEY PROJECTS ENFSI MONOPOLY PROGRAMME 2010

1 The development and implementation of an ENFSI standard for reporting evaluative forensic evidence.

Sheila Willis
(EFE – Ireland)

ENFSI

The Problem

- Performance of analytical methods: quite standardized
  - E.g.: measurement of refractive index of glass, concentration of drugs, etc.

The Problem

- Performance of analytical methods: quite standardized
  - E.g.: measurement of refractive index of glass, concentration of drugs, etc.
- Not the case of evidence evaluation with likelihood ratios
  - Absence of widely accepted, standard procedures
- Recent workshop organized at the NFI
  - Participation of NFI and external experts
  - Different approaches proposed, not a general consensus
  - Results to be made public soon

Netherlands Forensic Institute
Ministry of Security and Justice

The Problem

- Performance of analytical methods: quite standardized
  - E.g.: measurement of refractive index of glass, concentration of drugs, etc.

- Not the case of evidence evaluation with likelihood ratios
  - Absence of widely accepted, standard procedures

- Recent workshop organized at the NFI
  - Participation of NFI and external experts
  - Different approaches proposed, not a general consensus
  - Results to be made public soon

Netherlands Forensic Institute
Ministry of Security and Justice

- Performance of Likelihood Ratios: **what** to measure, **and how**?

Aim of This Talk

1. Present a methodology for measuring the performance of likelihood ratios
   - Solid grounds on Bayesian statistics (probabilistic assessments)

Aim of This Talk

1. Present a methodology for measuring the performance of likelihood ratios
   - Solid grounds on Bayesian statistics (probabilistic assessments)
2. Describe the concept of calibration
   - Measures important properties of the likelihood ratio

Aim of This Talk

1. Present a methodology for measuring the performance of likelihood ratios
   - Solid grounds on Bayesian statistics (probabilistic assessments)
2. Describe the concept of calibration
   - Measures important properties of the likelihood ratio
3. Propose a way of measuring calibration of LR values
   - Empirical Cross-Entropy
   - Free software tool available…

Aim of This Talk

1. Present a methodology for measuring the performance of likelihood ratios
   - Solid grounds on Bayesian statistics (probabilistic assessments)
2. Describe the concept of calibration
   - Measures important properties of the likelihood ratio
3. Propose a way of measuring calibration of LR values
   - Empirical Cross-Entropy
   - Free software tool available…
4. Illustrate it with experimental examples

Performance of Probabilistic Assessments

Performance of Probabilistic Assessments

- There is a wealth of statistics literature on the topic
  - Some examples:

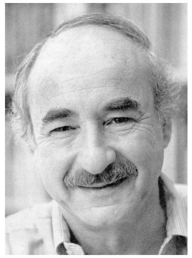
Performance of Probabilistic Assessments

- There is a wealth of statistics literature on the topic
  - Some examples:

The Statistician 32 (1983)

The Comparison and Evaluation of Forecasters†

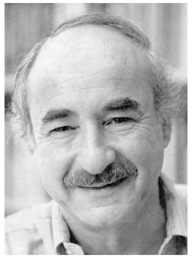
MORRIS H. DeGROOT and STEPHEN E. FIENBERG



Performance of Probabilistic Assessments

- There is a wealth of statistics literature on the topic
  - Some examples:

The Statistician 32 (1983)
The Comparison and Evaluation of Forecasters†
MORRIS H. DeGROOT and STEPHEN E. FIENBERG

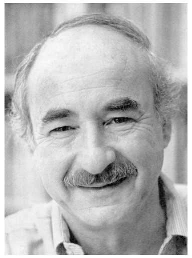


Journal of the American Statistical Association
September 1982, Volume 77, Number 379
The Well-Calibrated Bayesian
A. P. DAWID*

Performance of Probabilistic Assessments

- There is a wealth of statistics literature on the topic
  - Some examples:

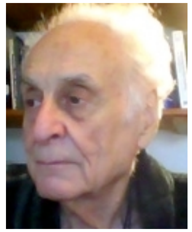
The Statistician 32 (1983)
The Comparison and Evaluation of Forecasters†
MORRIS H. DeGROOT and STEPHEN E. FIENBERG




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

J. R. Statist. Soc. A (1979),
142, *Part* 2, *pp.* 146–180
On the Reconciliation of Probability Assessments
D. V. LINDLEY, A. TVERSKY and R. V. BROWN

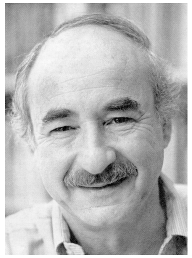




Performance of Probabilistic Assessments

- There is a wealth of statistics literature on the topic
  - Some examples:

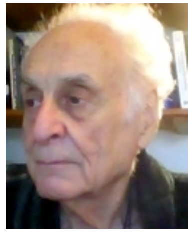
The Statistician 32 (1983)
The Comparison and Evaluation of Forecasters†
MORRIS H. DeGROOT and STEPHEN E. FIENBERG




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

***J. R. Statist. Soc.* A (1979),**
142, *Part* 2, *pp.* 146–180
On the Reconciliation of Probability Assessments
D. V. LINDLEY, A. TVERSKY and R. V. BROWN





- Calibration described as a desirable characteristic

Probabilistic Weather Forecasting

- Performance of these probabilistic assessments
- Classical example: weather forecasting
  - What is the probability of raining tomorrow?

Weather Forecasting Formally

- Variable of interest (event)
  - Rain a given day: θ

- Two possible outcomes (complementary): that given day…
  - It rained: $\theta = \theta_p$ 
  - It did not rain: $\theta = \theta_d$ 

- After next day the outcome of θ will be known (observed)
  - Either $\theta = \theta_p$ (it rained) or $\theta = \theta_d$ (it did not rain)

Weather Forecasting Formally

- What is the probability of raining tomorrow considering the available knowledge of the forecaster today?

$$P\left(\theta_p \middle| K\right)$$

- Given K: All knowledge available to the forecaster
  - May include training and education of the forecaster, data, other forecasts…

Performance of the Forecasts

- Ground-truth: status of θ in past predictions of the forecaster
  - For some predictions, θ_p was true: true-θ_p forecasts (it rained)
  - For some others, θ_d is true: true-θ_d forecasts (it did not rain)

Performance of the Forecasts

- Ground-truth: status of θ in past predictions of the forecaster
  - For some predictions, θ_p was true: true-θ_p forecasts (it rained) 
  - For some others, θ_d is true: true-θ_d forecasts (it did not rain) 

- Desired behavior of a forecast for a given day
  - If θ_p is true one day: it rained that day (ground-truth)
    - $P(\theta_p|K)$ should be **high** (close to 1)

Performance of the Forecasts

- Ground-truth: status of θ in past predictions of the forecaster
  - For some predictions, θ_p was true: true-θ_p forecasts (it rained)
  - For some others, θ_d is true: true-θ_d forecasts (it did not rain)

- Desired behavior of a forecast for a given day
  - If θ_p is true one day: it rained that day (ground-truth)
    - $P(\theta_p|K)$ should be **high** (close to 1)
  - If θ_d is true one day: it did not rain that day (ground-truth)
    - Thus, $P(\theta_p|K)$ should be **low** (close to 0)

Performance of the Forecasts

- Performance metric: *accuracy* of the forecasts
  - Average value of the “deviation” from the ground truth
- We need a measure of “deviation”
- Solution in classical statistical literature
  - Strictly Proper Scoring Rules (SPSR)

© Journal of the American Statistical Association
December 1971, Volume 66, Number 336
Theory and Methods Section

Elicitation of Personal Probabilities and Expectations

LEONARD J. SAVAGE*

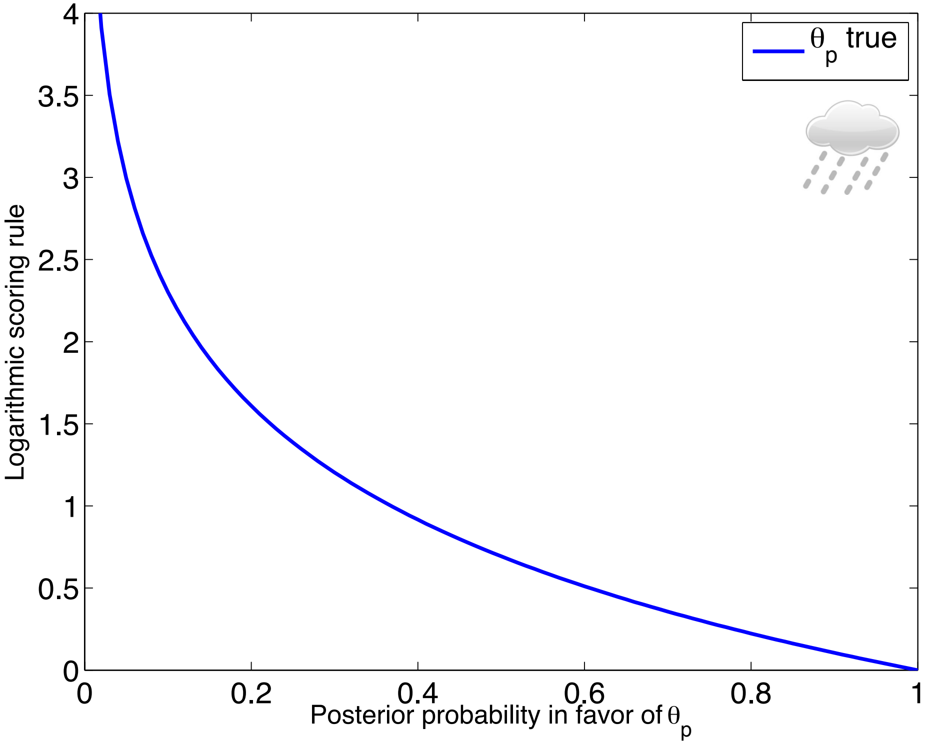
  - They present many desirable properties

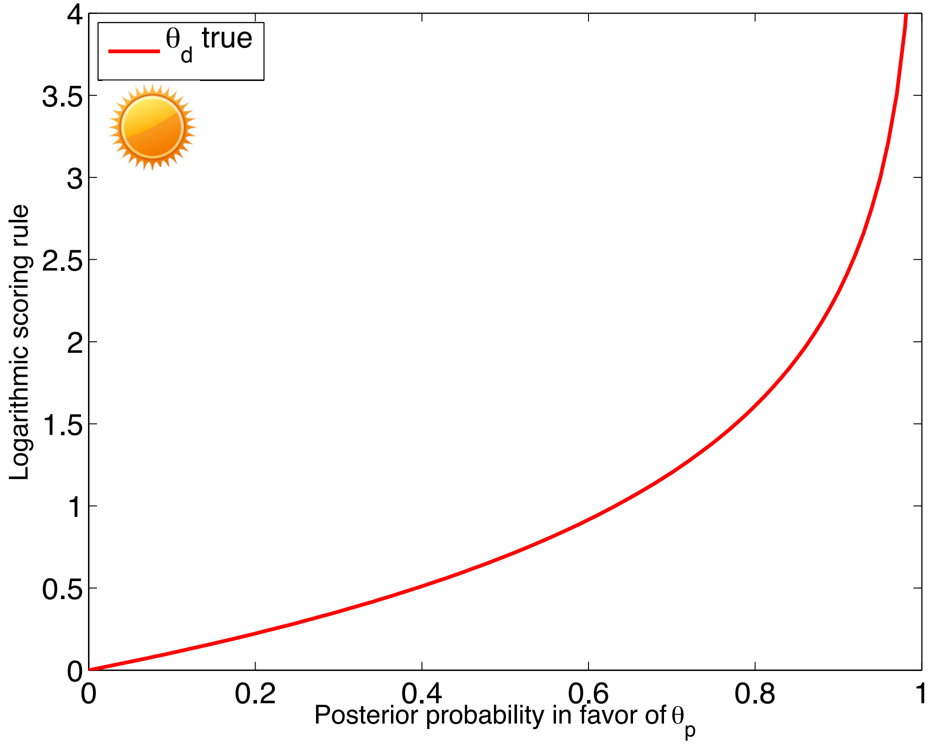
Example: Logarithmic SPSR

- Assigns a penalty to a forecast, given the ground-truth
  - Deviation of the forecast with respect to the ground-truth

$-\log_2 P(\theta_p|E)$ θ_p is true

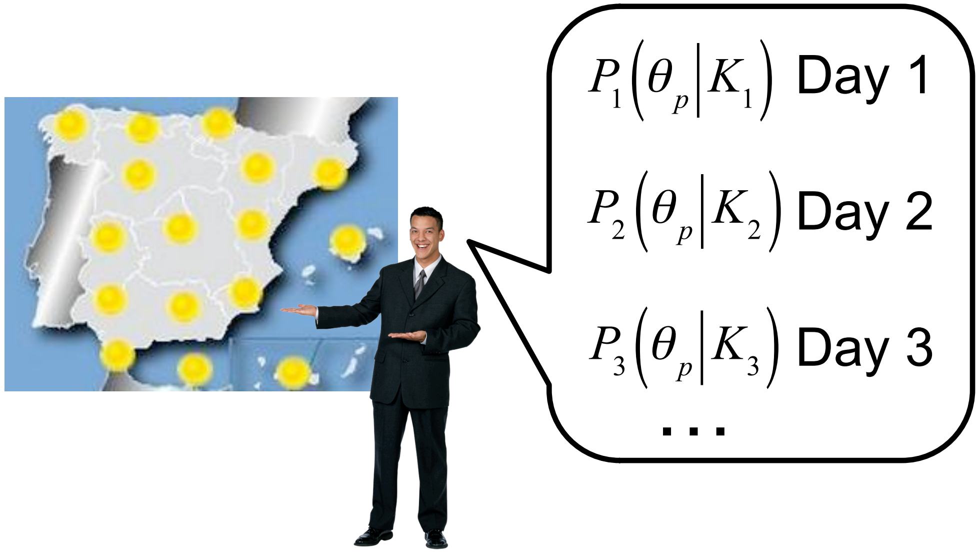
$-\log_2 P(\theta_d|E)$ θ_d is true





Overall Performance: Accuracy

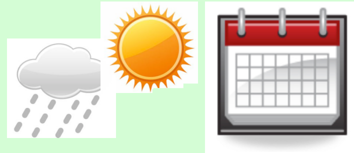
- For each forecast, SPSR means deviation from the ground-truth
- Average of those deviations: accuracy
  - The lower its value, the better
  - Example for logarithmic SPSR



Overall Performance: Accuracy

- For each forecast, SPSR means deviation from the ground-truth
- Average of those deviations: accuracy
  - The lower its value, the better
  - Example for logarithmic SPSR

Ground-truth



$P_1(\theta_p|K_1)$ Day 1

$P_2(\theta_p|K_2)$ Day 2

$P_3(\theta_p|K_3)$ Day 3

...

Overall Performance: Accuracy

- For each forecast, SPSR means deviation from the ground-truth
- Average of those deviations: accuracy
  - The lower its value, the better
  - Example for logarithmic SPSR

Ground-truth

$P_1(\theta_p|K_1)$ Day 1

$P_2(\theta_p|K_2)$ Day 2

$P_3(\theta_p|K_3)$ Day 3

...

Average of SPSR

$$LS = \boxed{-\frac{1}{N_p}\sum_{i\in true-\theta_p}\log_2 P_i(\theta_p|K_i)}\ \boxed{-\frac{1}{N_d}\sum_{j\in true-\theta_d}\log_2 P_j(\theta_d|K_j)}$$

Accuracy

Calibration of Probabilistic Assessments

Calibration

- Given all the forecasts from past days
  - With their corresponding ground-truth
- **Calibration** means
  - Forecasts $P(\theta_p|K)$ (probability of rain) approximate actual proportions of occurrence of θ_p (it rained)

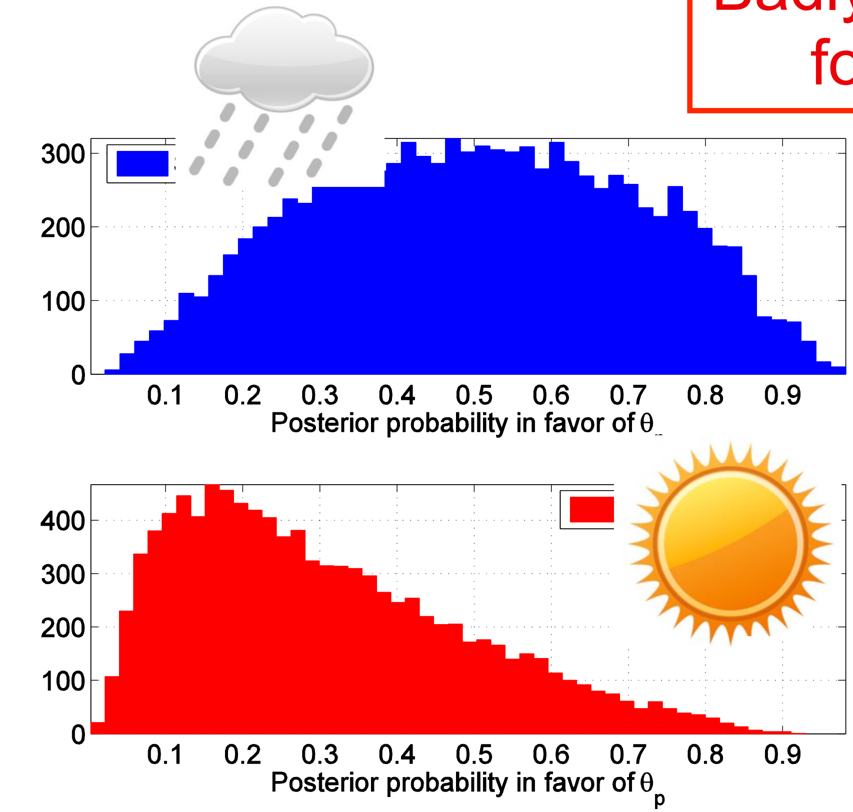
LINDLEY, TVERSKY AND BROWN – *Reconciliation of Probability Assessments*

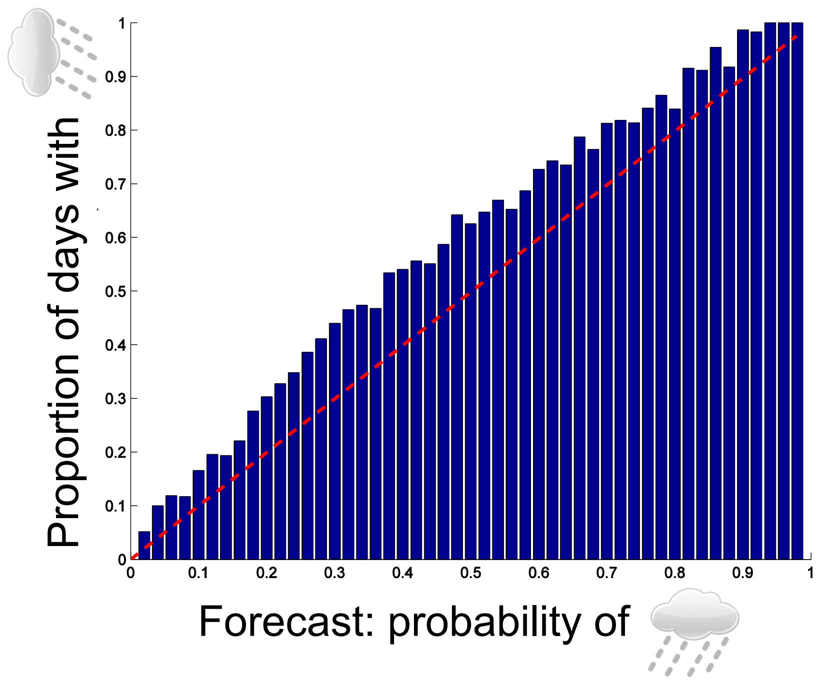
assessments in terms of a semantic criterion that pertains to the meaning of the probability scale. Clearly, there is no way of validating, for example, a meteorologist's single judgement that the probability of rain is 2/3. If the meteorologist is using the scale properly, however, we would expect that 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,

Calibration

- Example: experimental set of past forecasts
  - Separated by the status of the ground-truth

Badly-calibrated forecasts

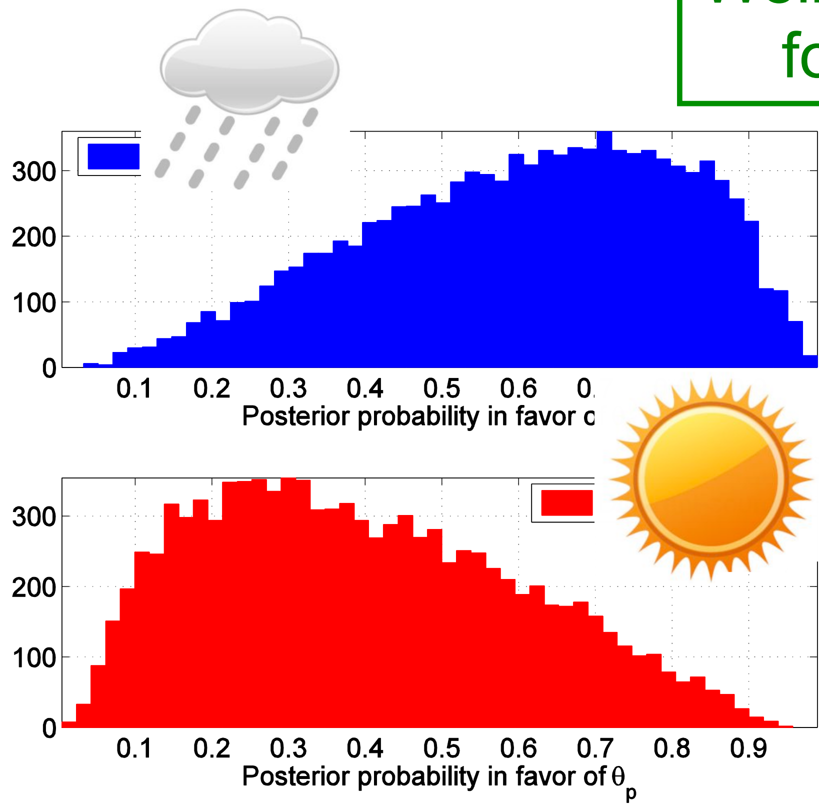


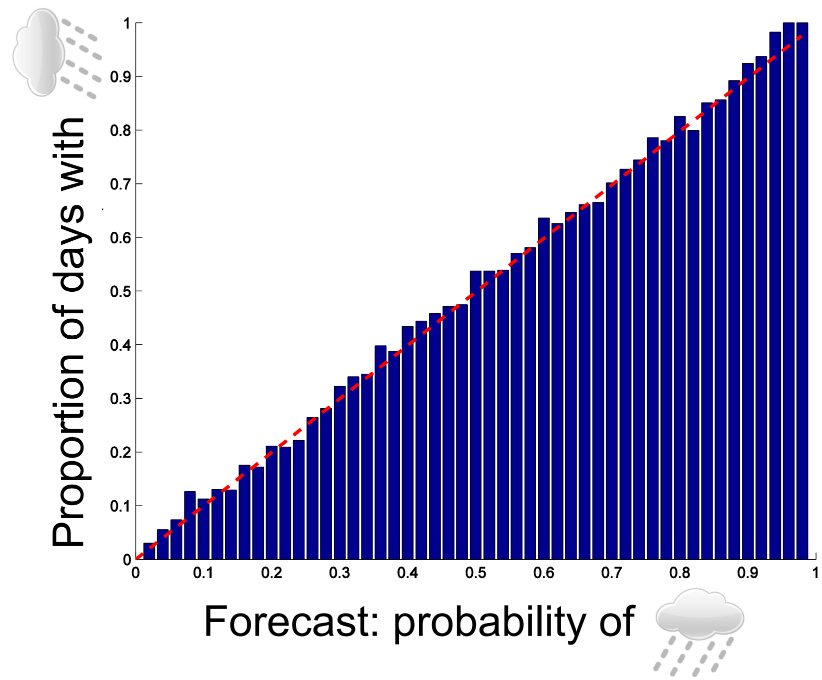


Calibration

- Example: experimental set of past forecasts
  - Separated by the status of the ground-truth

Well-calibrated forecasts

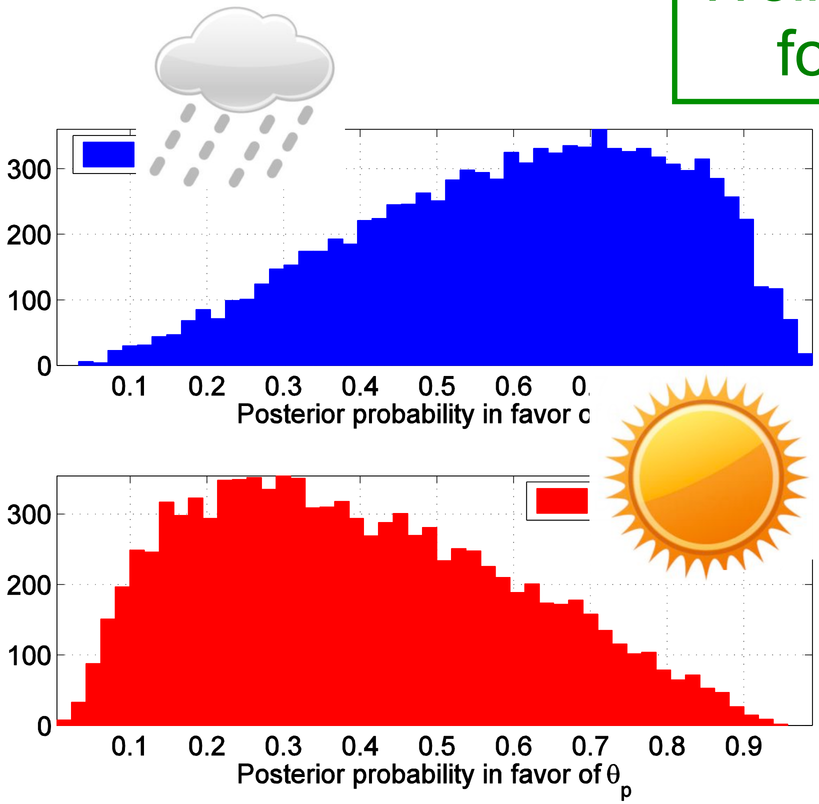


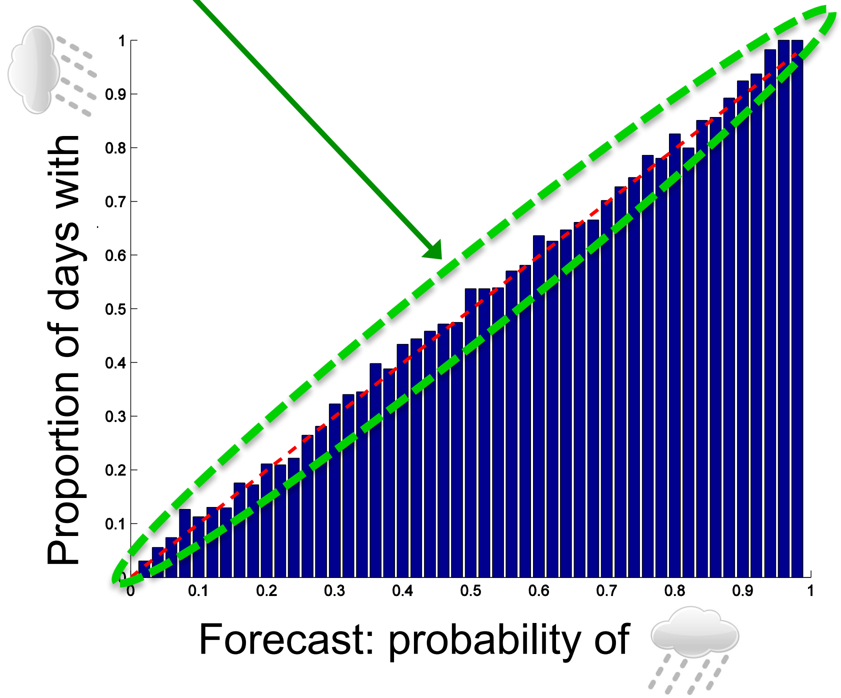


Calibration

- Example: experimental set of past forecasts
  - Separated by the status of the ground-truth

Well-calibrated forecasts

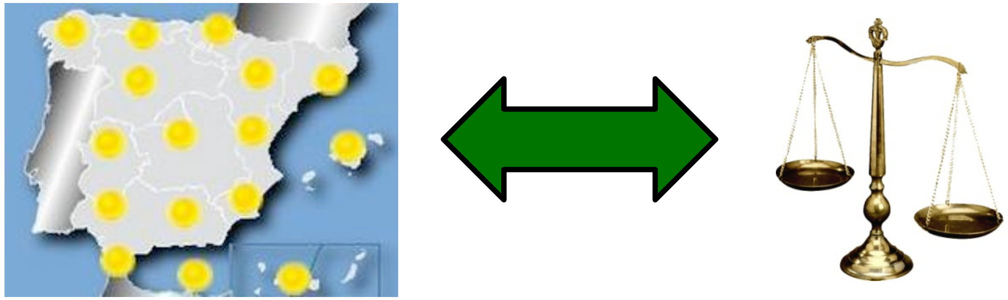




Calibration in Forensic Science

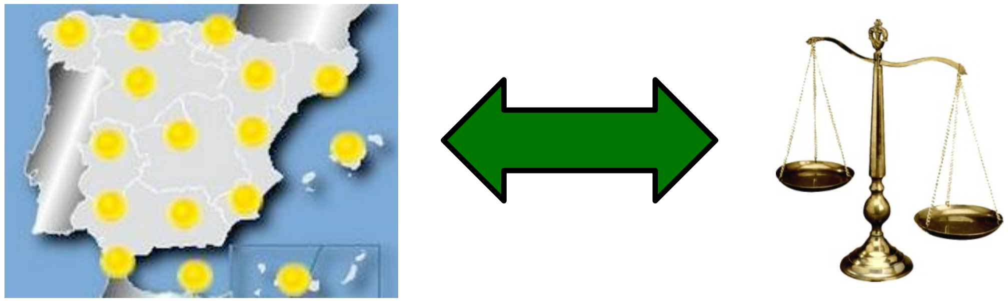
Calibration in Forensic Science?

- Well-calibrated probabilistic weather forecasts have many nice properties, studied during decades
- Can we use this performance framework for evidence evaluation in forensic science?



Calibration in Forensic Science?

- Well-calibrated probabilistic weather forecasts have many nice properties, studied during decades
- Can we use this performance framework for evidence evaluation in forensic science?

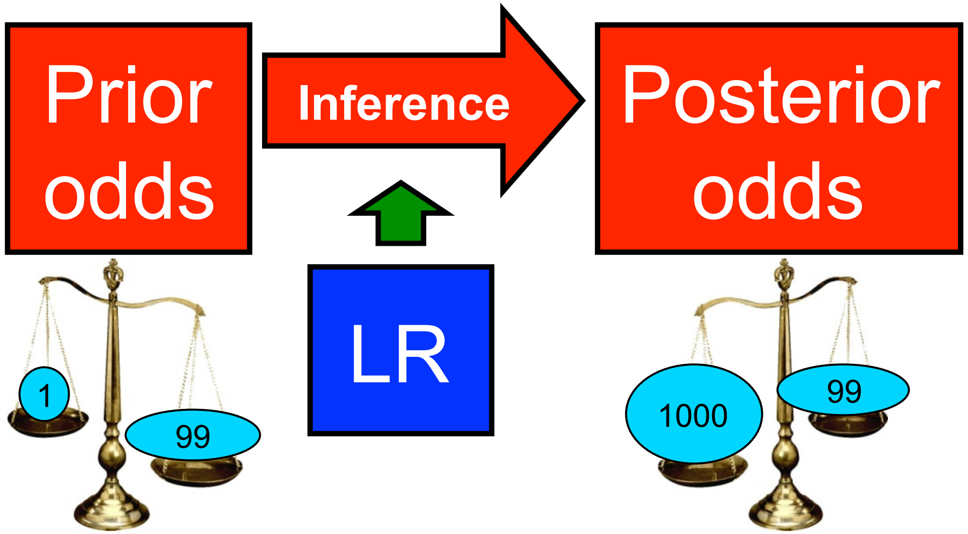


- Not straightforward…
- At all…

ATVS
UAM UNIVERSIDAD AUTONOMA DE MADRID

Inference in Forensic Science

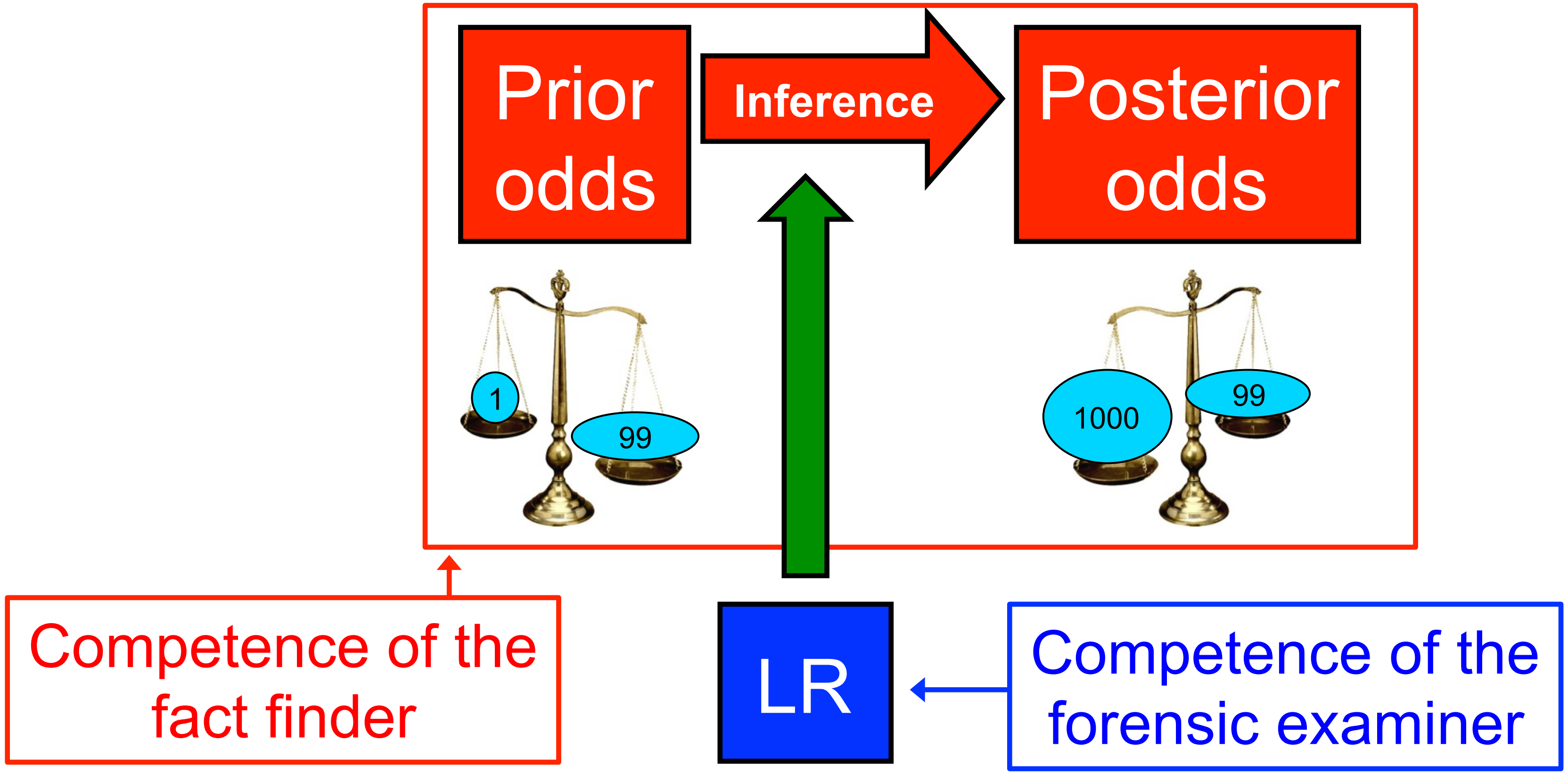
- Likelihood ratio: value of the evidence



$$\underbrace{\frac{P(\theta_p|I)}{P(\theta_d|I)}}_{\text{Prior odds}} \times \underbrace{\frac{P(E|\theta_p,I)}{P(E|\theta_d,I)}}_{LR} = \underbrace{\frac{P(\theta_p|E,I)}{P(\theta_d|E,I)}}_{\text{Posterior odds}}$$

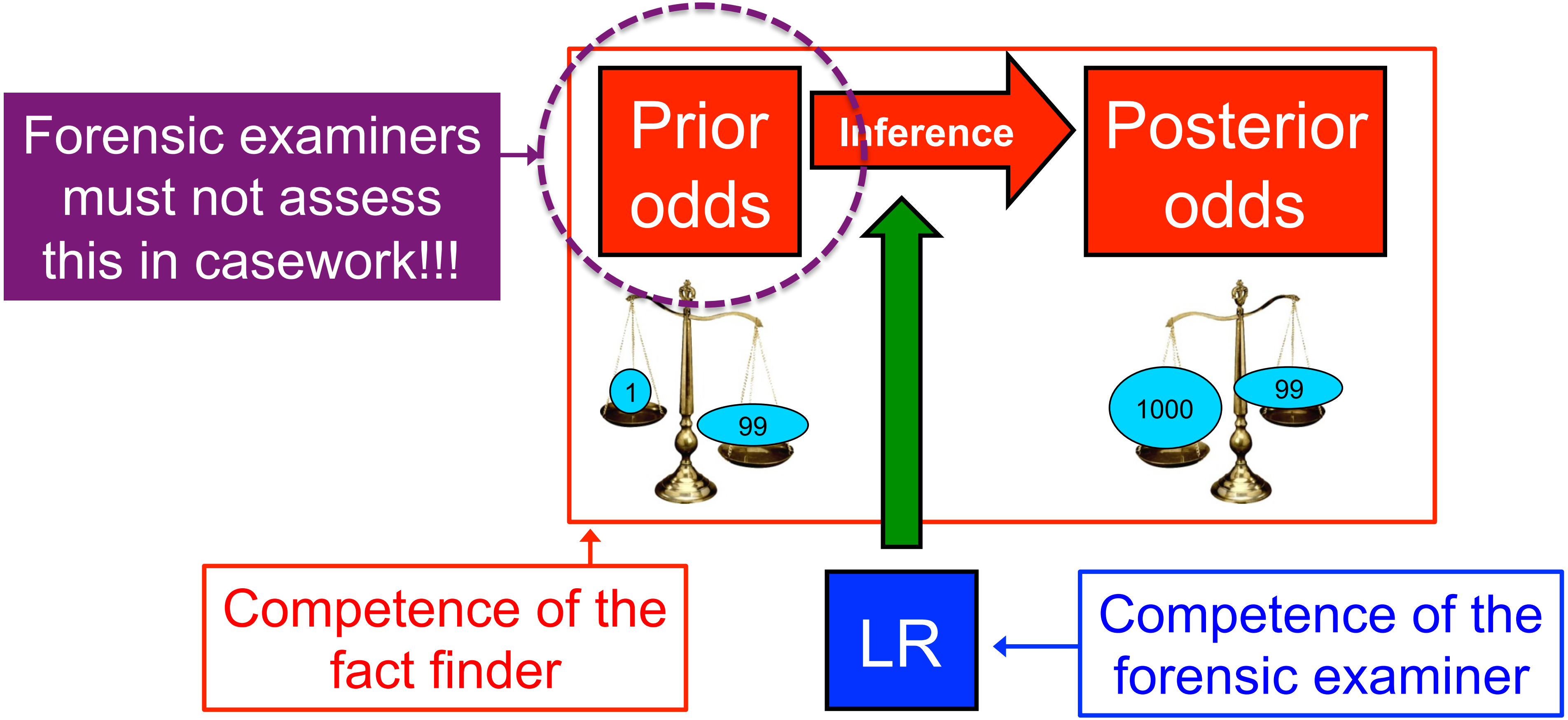
Inference in Forensic Science

- Role of the forensic examiner: LR
- Role of the fact finder: prior and posterior odds



Inference in Forensic Science

- Role of the forensic examiner: LR
- Role of the fact finder: prior and posterior odds



Probabilistic Assessment in Forensic Science

- Probabilistic assessment in weather forecasting

$$P\left(\theta_p \middle| K\right)$$

$\boxed{\text{“Probability of } \theta_p \text{ given } K\text{”}}$

- θ_p: rain
- θ_d: not rain
- K: available knowledge

Probabilistic Assessment in Forensic Science

- Probabilistic assessment in weather forecasting

$$P\left(\theta_p \middle| K\right)$$

“Probability of θ_p given K”

- θ_p: rain
- θ_d: not rain
- K: available knowledge

- Equivalent in forensic science: posterior probability

$$P\left(\theta_p \middle| E, I\right)$$

“Probability of θ_p given E,I”

- θ_p: prosecutor hypothesis
- θ_d: defense hypothesis
- E, I: available knowledge (evidence + other information)

Probabilistic Assessment in Forensic Science

- Probabilistic assessment in weather forecasting

$$P\left(\theta_p \middle| K\right)$$

"Probability of θ_p given K"

- θ_p: rain
- θ_d: not rain
- K: available knowledge

- Equivalent in forensic science: posterior probability

$$P\left(\theta_p \middle| E, I\right)$$

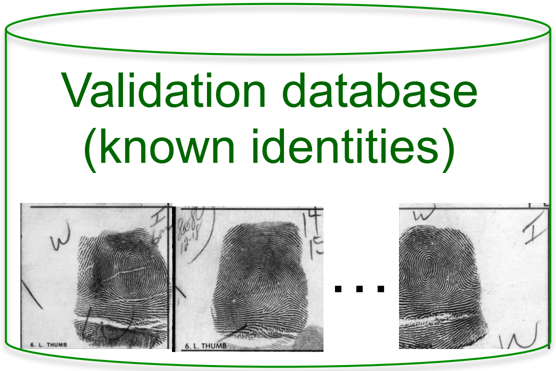
"Probability of θ_p given E,I"

- θ_p: prosecutor hypothesis
- θ_d: defense hypothesis
- E, I: available knowledge (evidence + other information)

- But the forensic examiner must not assess the prior!
  - Therefore, she or he cannot use the posterior!
  - How to measure calibration then?

Calibration in Forensic Science: Solution

- Step 1: set-up a validation experiment
  - Compute LR values
  - Using a validation database
  - This is done <u>for validation</u>, <u>not for casework</u>

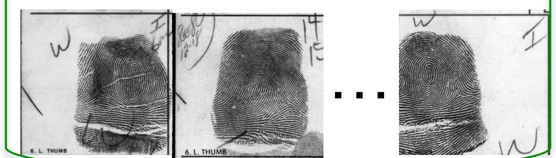


Calibration in Forensic Science: Solution

- Step 1: set-up a validation experiment
  - Compute LR values
  - Using a validation database
  - This is done for validation, not for casework

Ground-truth: status of θ in a comparison
- Either θ_p is true
- Or θ_d is true

Validation database (known identities)

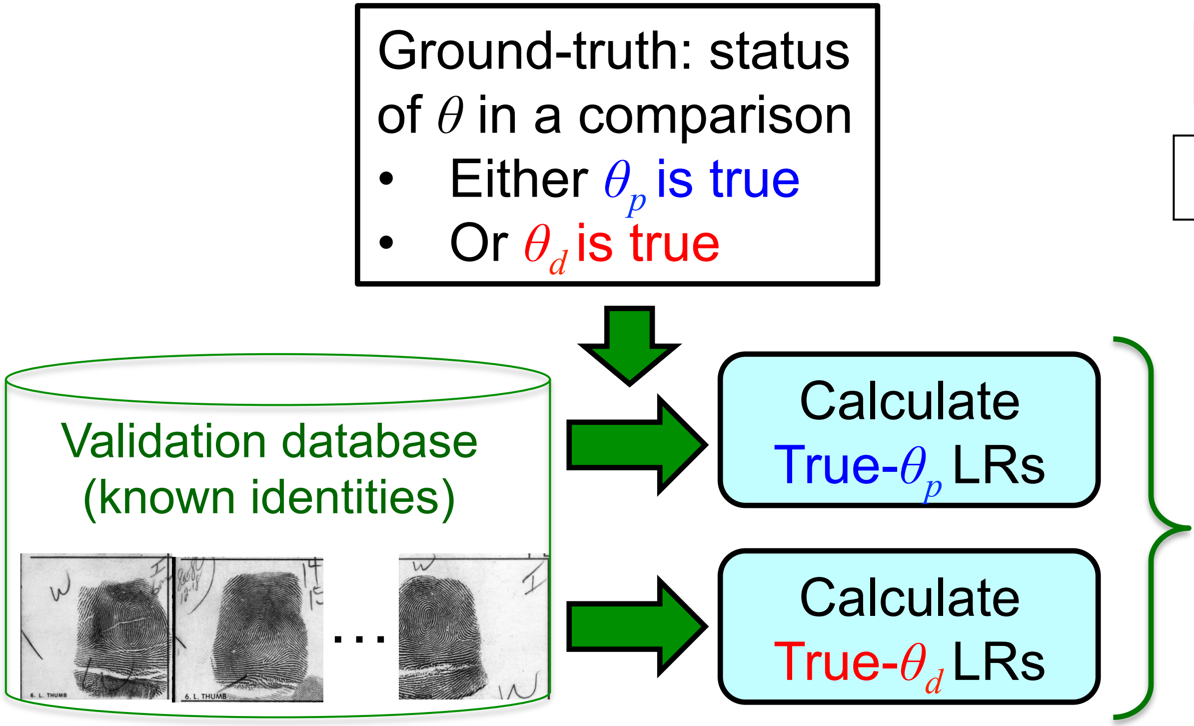


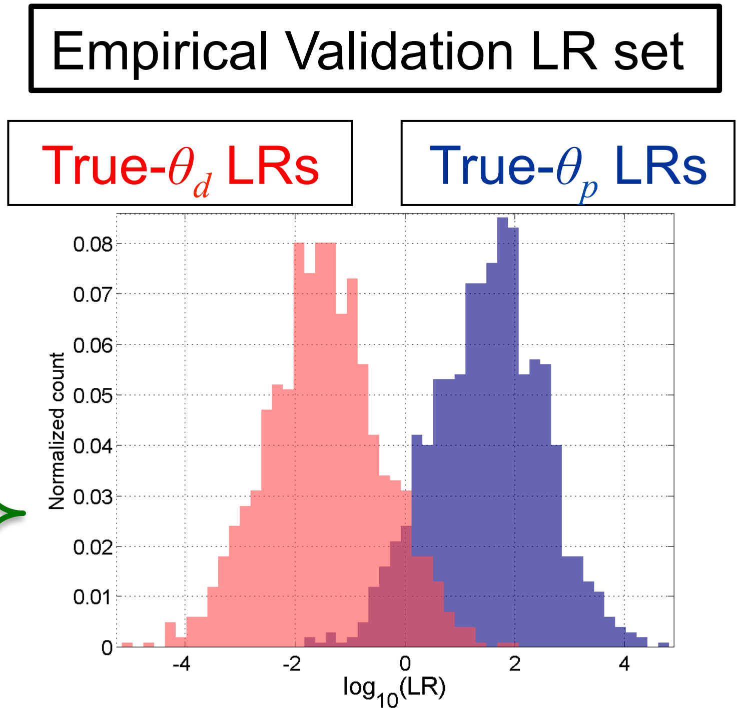
Calculate True-θ_p LRs

Calculate True-θ_d LRs

Calibration in Forensic Science: Solution

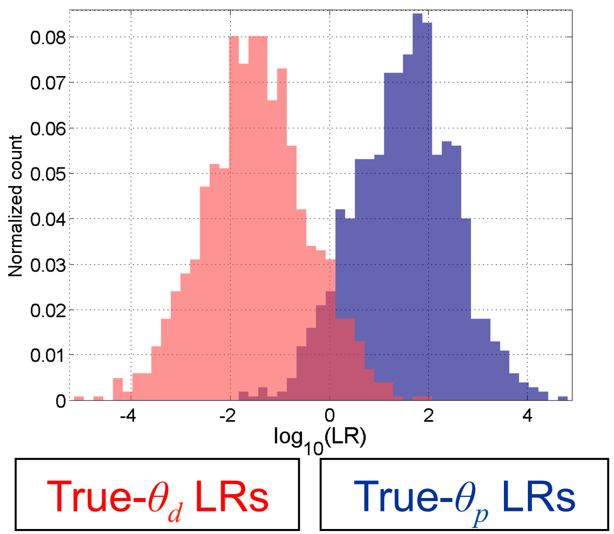
- Step 1: set-up a validation experiment
  - Compute LR values
  - Using a validation database
  - This is done <u>for validation</u>, <u>not for casework</u>

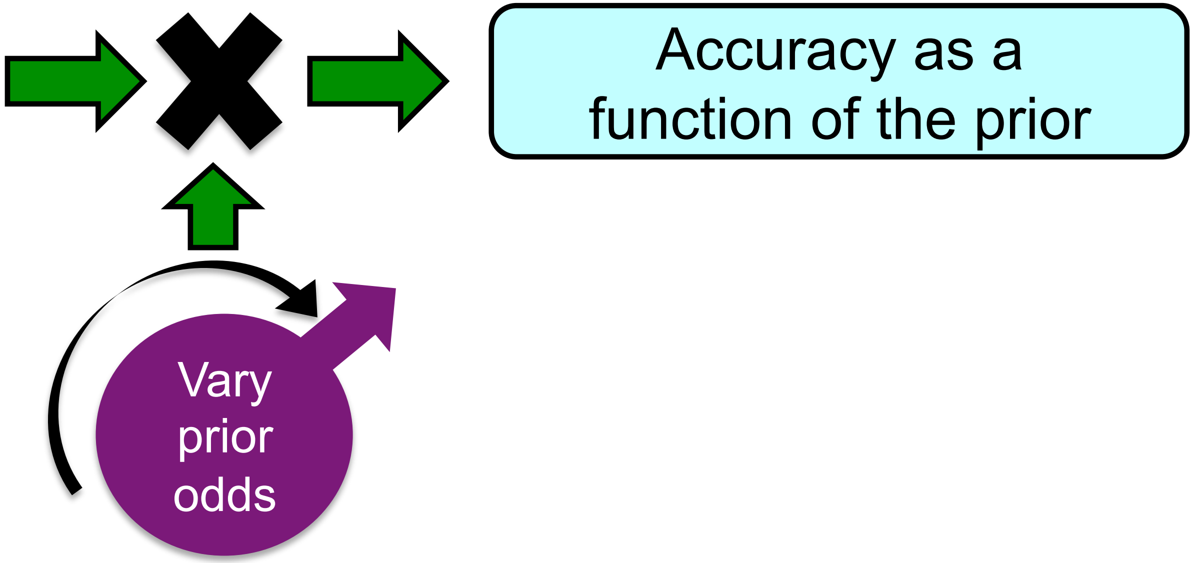




Calibration in Forensic Science: Solution

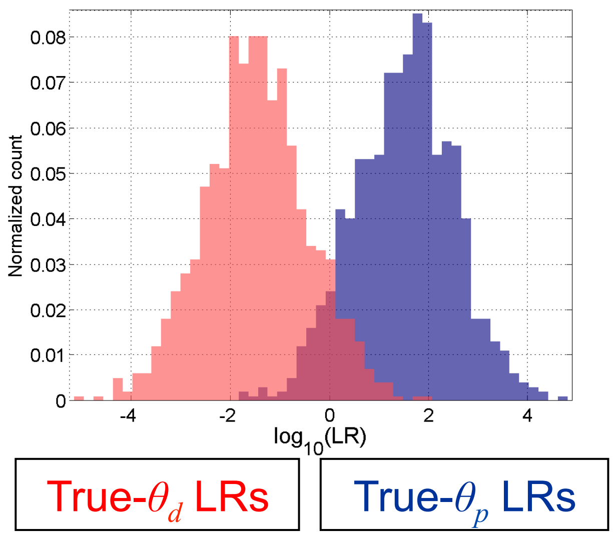
- Step 2: consider the prior as an unknown parameter
  - Do not assess its value in any case!
  - But vary it over a wide range within the experiment
    - In casework, however, you will just compute the LR!
- Compute and represent accuracy (average of SPSR) for all the priors in that range

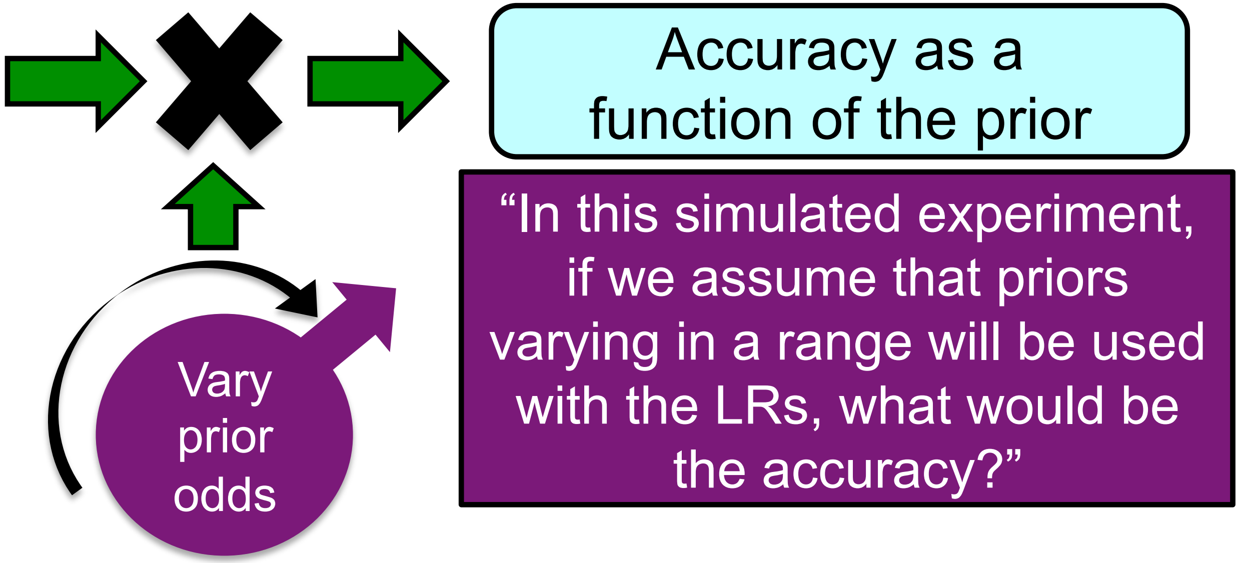




Calibration in Forensic Science: Solution

- Step 2: consider the prior as an unknown parameter
  - Do not assess its value in any case!
  - But vary it over a wide range within the experiment
    - In casework, however, you will just compute the LR!
- Compute and represent accuracy (average of SPSR) for all the priors in that range

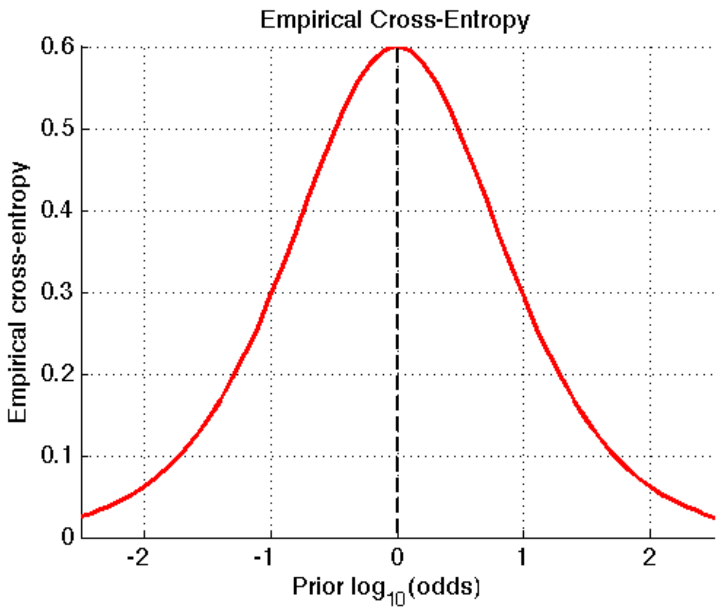




Accuracy of LRs: Empirical Cross-Entropy

- Proposed choice of SPSR: logarithmic SPSR
  - It can be argued that it has nice properties
- Accuracy: Empirical Cross-Entropy

$$ECE = \boxed{-\frac{P(\theta_p|I)}{N_p}\sum_{i\in true-\theta_p}\log_2 P_i(\theta_p|E_i,I)}$$
$$\boxed{-\frac{P(\theta_d|I)}{N_d}\sum_{j\in true-\theta_d}\log_2 P_j(\theta_p|E_j,I)}$$

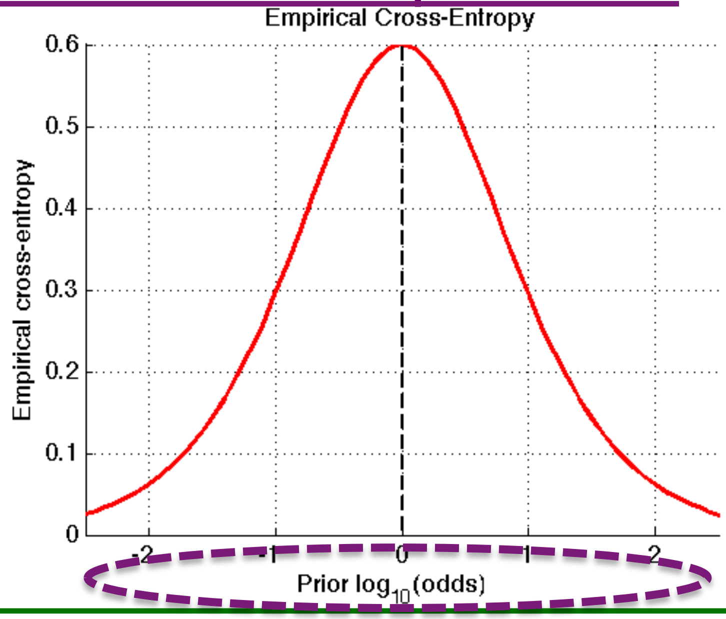


D. Ramos, J. Gonzalez-Rodriguez, G. Zadora and C. Aitken. “Information-theoretical Assessment of the Performance of Likelihood Ratios”. Journal of Forensic Sciences (under minor revision)

Accuracy of LRs: Empirical Cross-Entropy

- Proposed choice of SPSR: logarithmic SPSR
  - It can be argued that it has nice properties
- Accuracy: Empirical Cross-Entropy
  - We only vary prior odds <u>in the validation experiment</u>
  - <u>In casework only the LR will be reported (as usual)</u>

$$ECE = \boxed{-\frac{P(\theta_p|I)}{N_p}\sum_{i\in true-\theta_p}\log_2 P_i(\theta_p|E_i,I)}\ \boxed{-\frac{P(\theta_d|I)}{N_d}\sum_{j\in true-\theta_d}\log_2 P_j(\theta_p|E_j,I)}$$

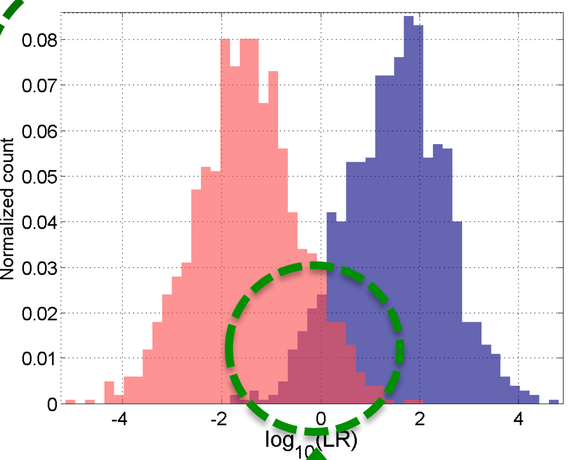


D. Ramos, J. Gonzalez-Rodriguez, G. Zadora and C. Aitken. "Information-theoretical Assessment of the Performance of Likelihood Ratios". Journal of Forensic Sciences (under minor revision)

Accuracy=Discrimination + Calibration

- In order to explicitly measuring calibration
- Accuracy can be decomposed into
  - Discriminating power of the LR set
    - Ability to distinguish between true-θ_p and true-θ_d cases
  - Calibration of the LR set

Accuracy of LRs (ECE) = { Discriminating Power of LRs + Calibration of LRs }



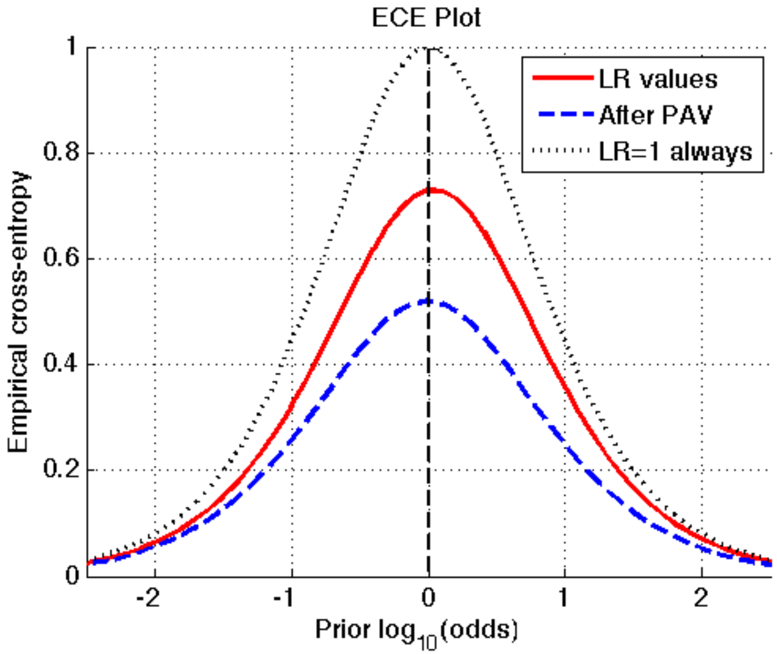
Degree of overlap (roughly speaking)

Discrimination + Calibration: ECE Plot

- Decomposition into discrimination and calibration

Niko Brümmer [a,b,*], Johan du Preez [b]
Application-independent evaluation of speaker detection
Computer Speech and Language 20 (2006) 230–275

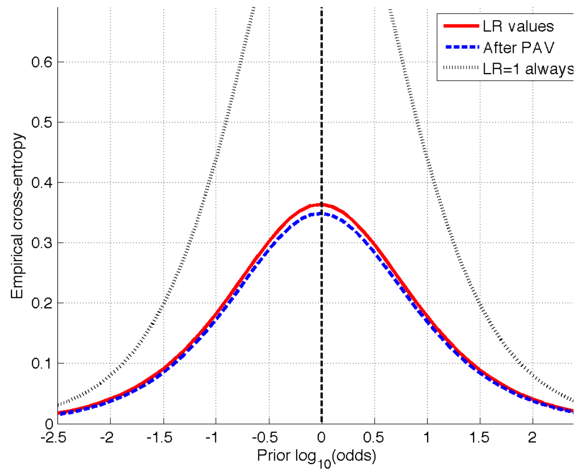
- Allows explicitly and quantitatively measuring calibration
  - Red curve: ECE
    - Accuracy of LRs
  - Blue curve:
    - Discrimination of LRs
  - Red minus Blue
    - **Calibration of LRs**

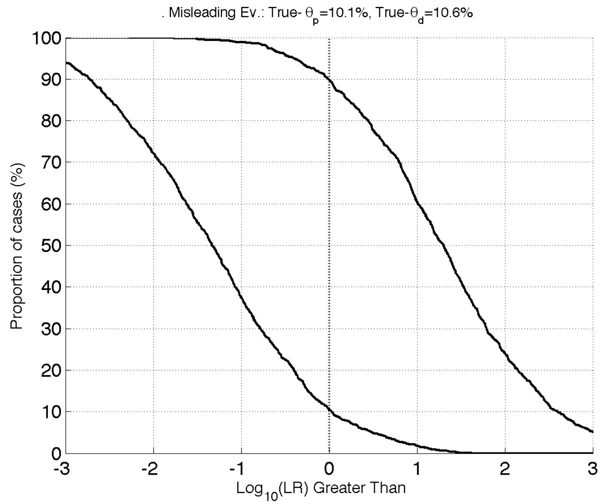


A Nice Property of Well-Calibrated Likelihood Ratios (There are More…)

Calibration and the Weight of the Evidence

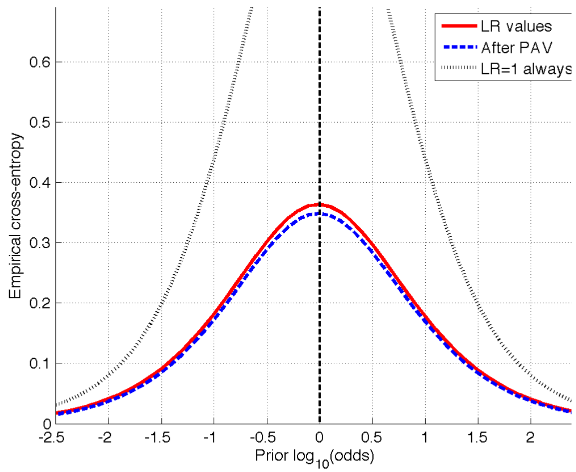
- Example: well-calibrated validation sets of LRs
  - Blue and red curves in ECE are pretty close

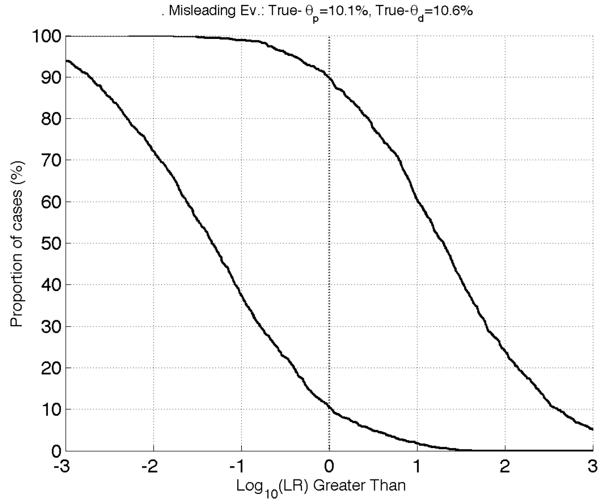


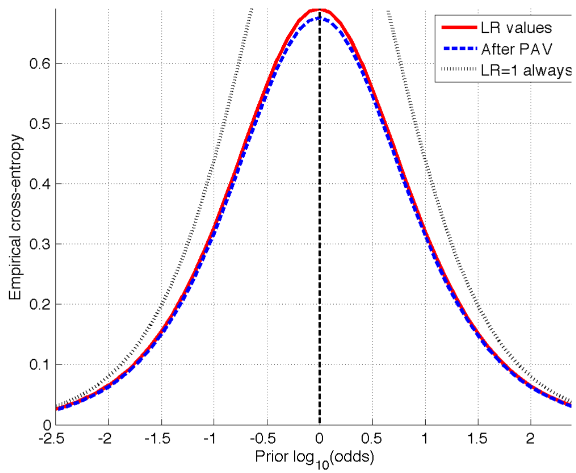


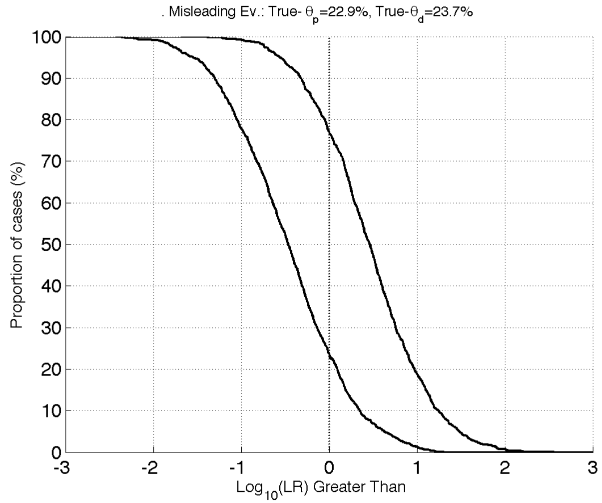
Calibration and the Weight of the Evidence

- Example: well-calibrated validation sets of LRs
  - Blue and red curves in ECE are pretty close







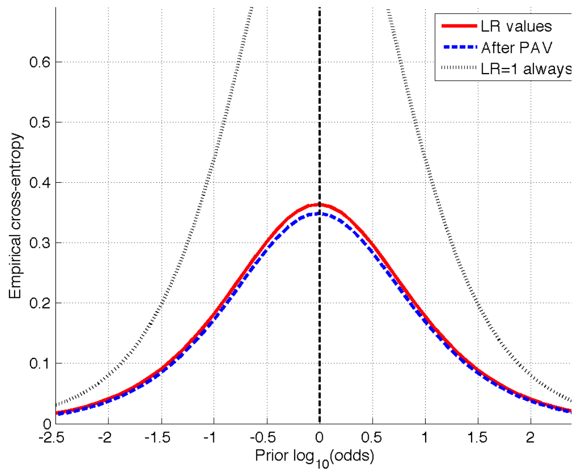


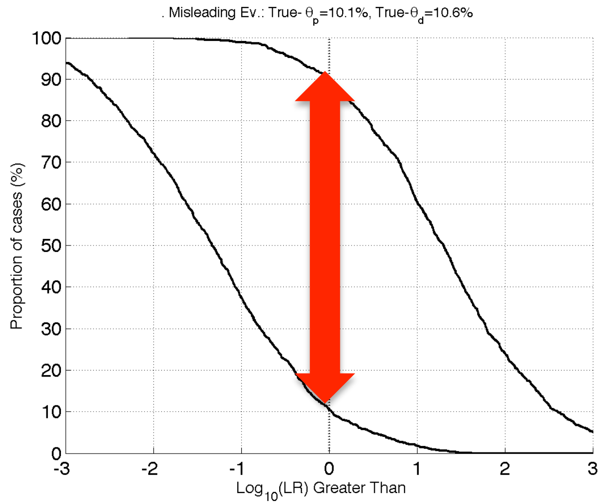
Calibration and the Weight of the Evidence

- Example: well-calibrated validation sets of LRs
  - Blue and red curves in ECE are pretty close

Better discrimination (related to lower rates of misleading evidence)

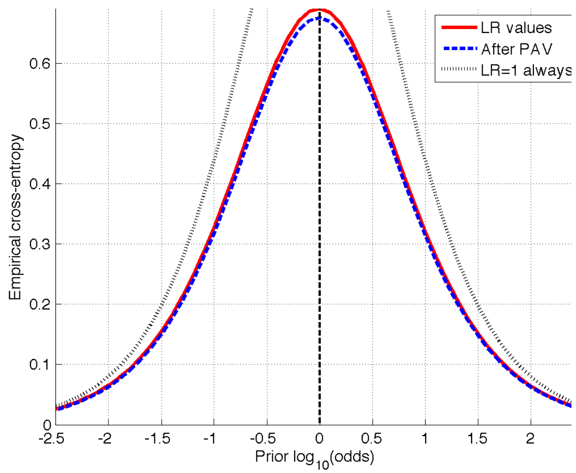
Higher max. |log(LR)|

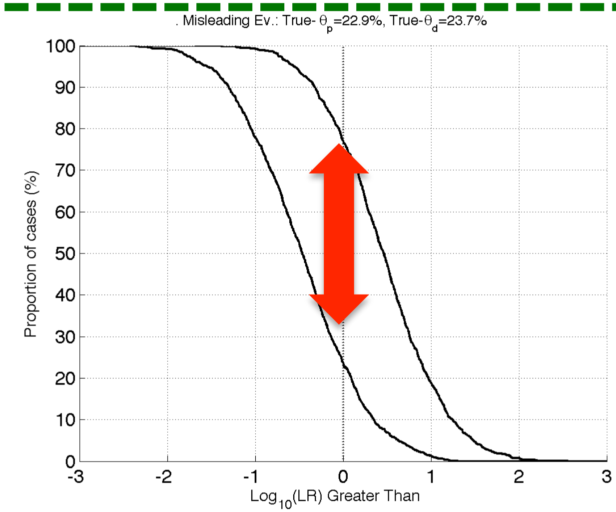




Worse discrimination (related to higher rates of misleading evidence)

Lower max. |log(LR)|





Calibration and the Weight of the Evidence

- The better the discriminating power of a method
- The stronger the LRs if they are well-calibrated
- And vice-versa

Calibration and the Weight of the Evidence

- The better the discriminating power of a method
- The stronger the LRs if they are well-calibrated
- And vice-versa

A reliable behavior indeed

  - If calibration is good, only methods with high discriminating power will be yielding strong LR values
  - Examples:
    - DNA: generally very discriminating, high LRs
    - Speech: generally not so discriminating, lower LRs

Calibration and the Weight of the Evidence

- The better the discriminating power of a method
- The stronger the LRs if they are well-calibrated
- And vice-versa

A reliable behavior indeed

  - If calibration is good, only methods with high discriminating power will be yielding strong LR values
  - Examples:
    - DNA: generally very discriminating, high LRs
    - Speech: generally not so discriminating, lower LRs
- Calibration has been dubbed "reliability"
  - Because of this and other properties

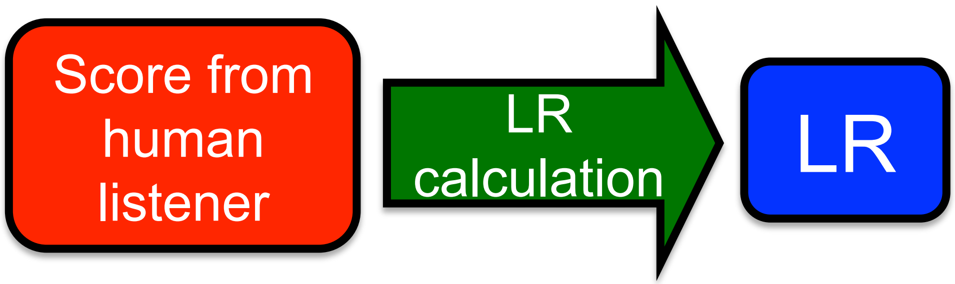
The Statistician 32 (1983)
The Comparison and Evaluation of Forecasters†
MORRIS H. DeGROOT and STEPHEN E. FIENBERG

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

Experimental Examples

Speaker Recognition: Human Lay Listeners

- Context: NIST Human-Assisted Speaker Recognition Evaluation 2010 (HASR)
- Objective: assess the value of the evidence of human lay listeners with LR
  - Scores (support) in a discrete scale: [-3,-2,-1,0,1,2,3]
  - LR calculation with those scores

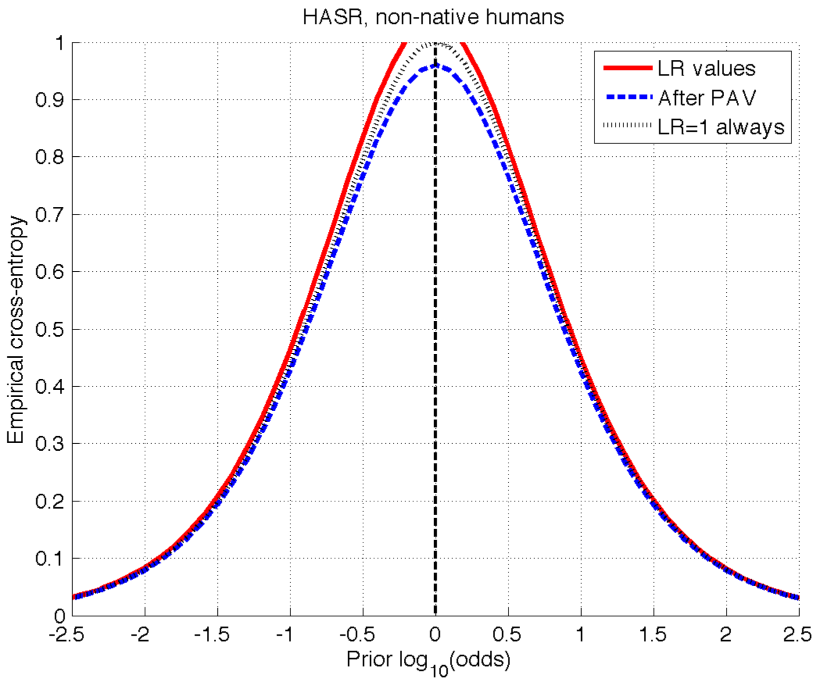


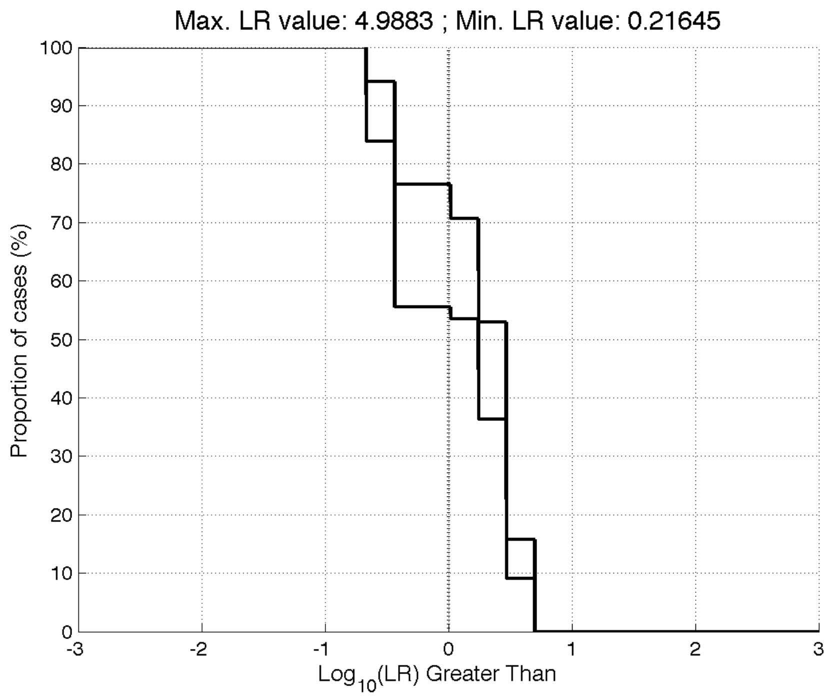
D. Ramos, J. Franco-Pedroso and J. Gonzalez-Rodriguez. "Calibration and weight of the evidence by human listeners. The ATVS-UAM submission to NIST HUMAN-aided speaker recognition 2010." Proc. of ICASSP 2011

- Development data from past NIST Evaluations
  - Human listeners gave scores for all those speech files

Speaker Recognition: Human Lay Listeners

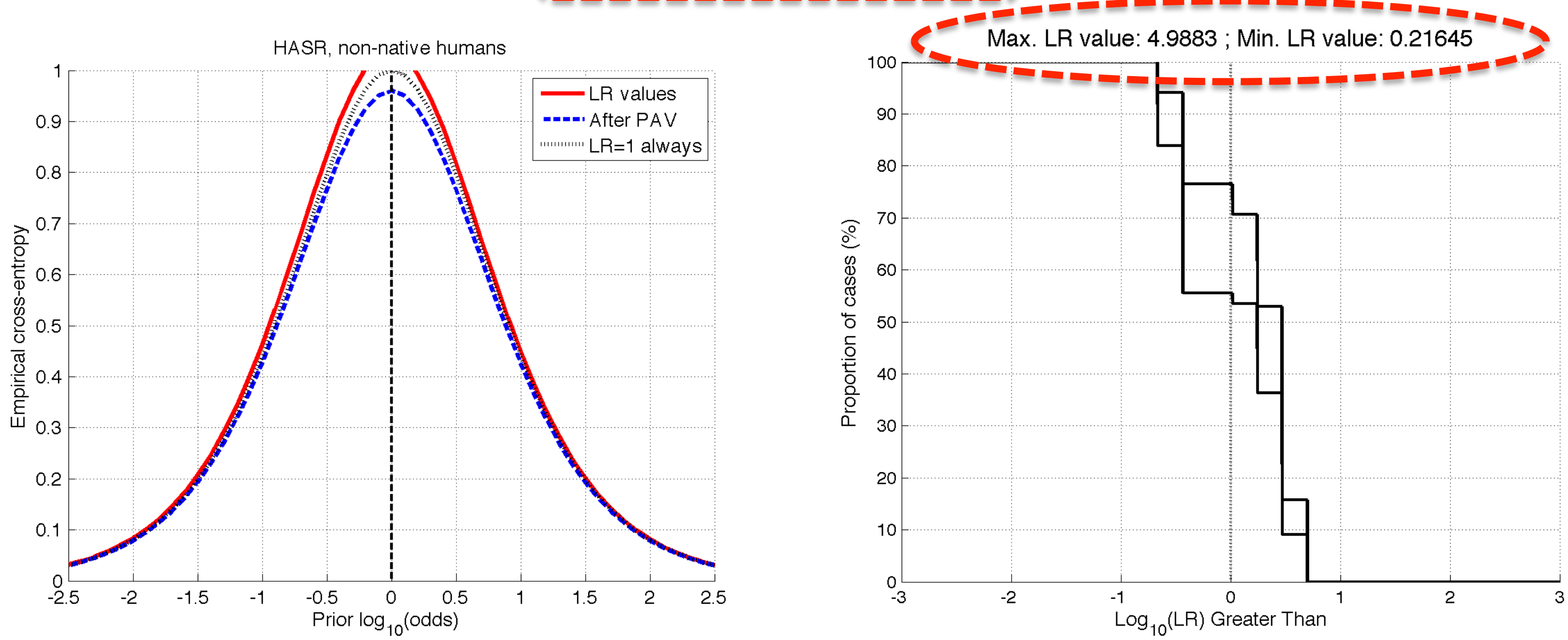
- LR from human listeners: very bad discriminating power
  - But very good calibration…
- Reliable behavior expected: “Very low discriminating power…”
  - “Therefore, only very weak support is given”





Speaker Recognition: Human Lay Listeners

- LR from human listeners: very bad discriminating power
  - But very good calibration…
- Reliable behavior expected: “Very low discriminating power…”
  - “Therefore, only very weak support is given”



Forensic Automatic Speaker Recognition

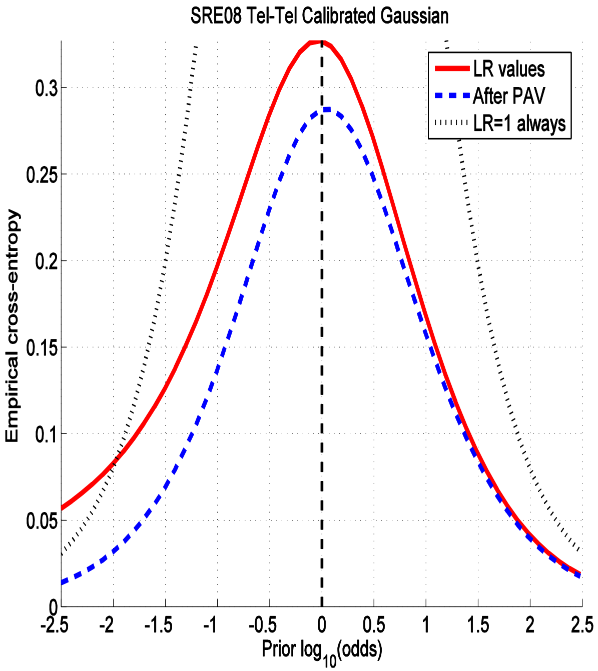
- Computation of LR values from scores by an automatic speaker recognition system

J. Gonzalez Rodriguez. P. Rose, D. Ramos, D. T. Toledano and Javier Ortega-Garcia. "Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition". IEEE Trans. On Speech, Audio and Language Processing, 15(7), 2007.

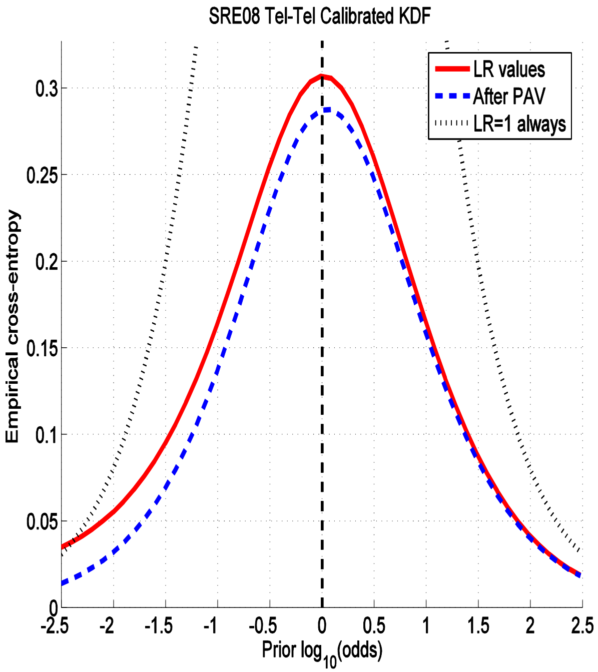
- Database and protocol: NIST Speaker Recognition Evaluation (SRE) 2008
  - Telephone-only subset
  - Hundreds of speakers, hundreds of thousands of comparisons
- Comparison of different LR computation methods
  - Gaussian modelling
  - Kernel density functions (KDF)
  - Logistic regression

NIST SRE 2008, telephone-only data

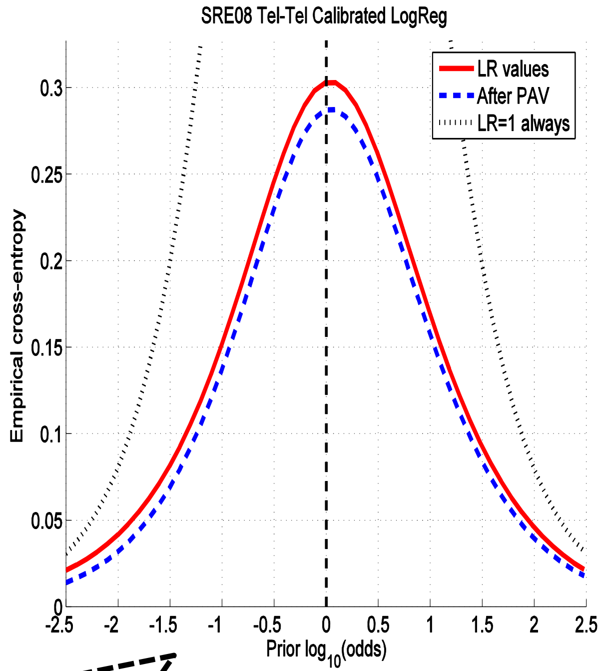
Gaussian



KDF



Logistic regression



Logistic regression better accuracy (ECE, red)
and better calibration (Red – Blue)

Forensic glass analysis

- Database of SEM-EDX profiles
- Collected by the Institute of Forensic Research, Krakow, Pl
  - 7 variables (Log of Na, Si, Ca, Al, K, Fe and Mg normalized to O)
- Performance degradation due to population selection

Background: 

Samples:  

Experiment ID: pw

Samples: 

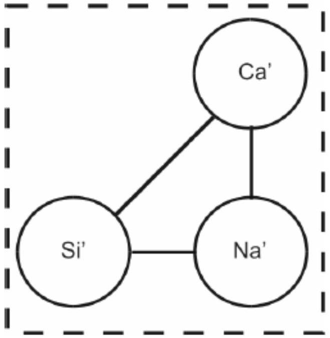
Experiment ID: pp

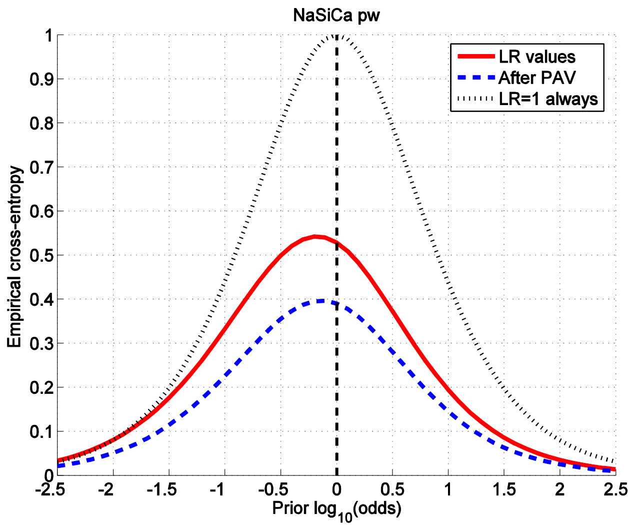
G. Zadora and D. Ramos, "Evaluation of glass samples for forensic purposes - An application of likelihood ratios and an information-theoretical approach.", *Chemometrics and Intelligent Laboratory Systems* 102(2), 2010.

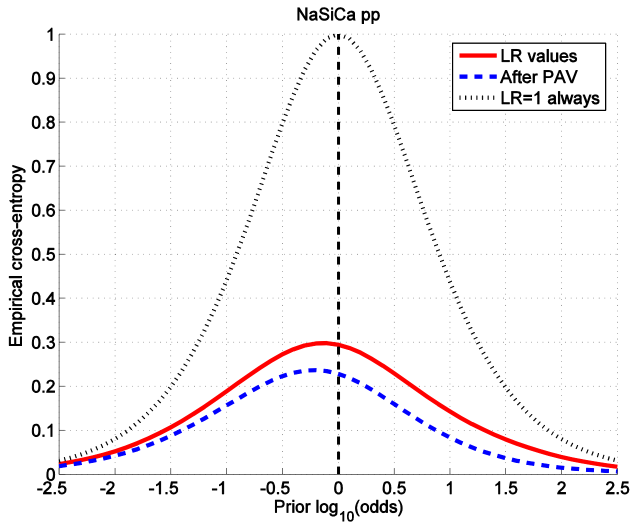
Forensic glass analysis

- Multivariate LR model

Selection of a population database of different type degrades calibration







Conclusions

Conclusions

- With the LR increasingly adopted…
  - We need to measure performance of LR methods
  - But… What to measure and how?

Conclusions

- With the LR increasingly adopted…
  - We need to measure performance of LR methods
  - But… What to measure and how?
- We have proposed a framework
  - Based on solid grounds of Bayesian statistics
  - Accuracy as a measure of goodness (SPSR)
  - Importance of Calibration
    - Well-calibrated LRs present desirable properties

Conclusions

- With the LR increasingly adopted…
  - We need to measure performance of LR methods
  - But… What to measure and how?
- We have proposed a framework
  - Based on solid grounds of Bayesian statistics
  - Accuracy as a measure of goodness (SPSR)
  - Importance of Calibration
    - Well-calibrated LRs present desirable properties
- Measuring calibration: Empirical Cross-Entropy
  - Can be applied to any LR-based forensic discipline
    - Shown in different experimental examples

Calibration: Free Matlab^{TM} Software

- ECE plots (Daniel Ramos)
  - http://arantxa.ii.uam.es/~dramos/software.html
- FoCal and BOSARIS toolkits (Niko Brümmer)
  - Tools for assessment
  - Tools for calibration
  - tinyurl.com/nbrummer

Reliable Support: Measuring Calibration of Likelihood Ratios

Daniel Ramos
ATVS – Biometric Recognition Group
Research Institute of Forensic Science and Security
Universidad Autonoma de Madrid
daniel.ramos@uam.es
http://arantxa.ii.uam.es/~dramos

Additional Slides

Accuracy of LRs: Empirical Cross-Entropy

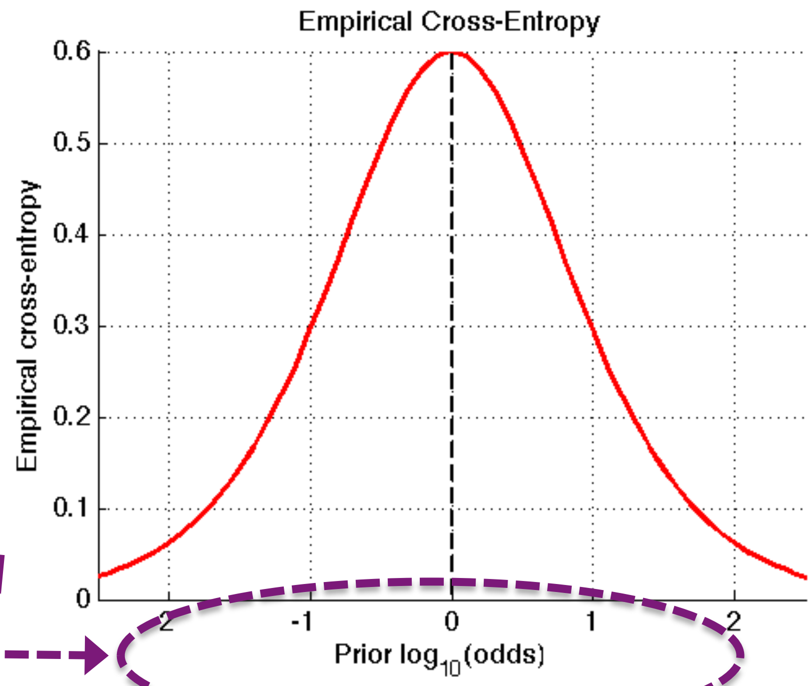
- **The prior is not assessed!**
  - We vary it in a wide range
  - Compute ECE for values in that range
  - Represent in a plot

$$ECE = -\frac{P(\theta_p|I)}{N_p}\sum_{i\in true-\theta_p}\log_2 P_i(\theta_p|E_i,I)$$
$$-\frac{P(\theta_d|I)}{N_d}\sum_{j\in true-\theta_d}\log_2 P_j(\theta_p|E_j,I)$$

$$P(\theta_p|I) = \frac{O(\theta_p)}{1+O(\theta_p)}$$

$$P(\theta_p|E,I) = \frac{LR\times O(\theta_p)}{1+LR\times O(\theta_p)}$$

$O(\theta_p)$



Calibration and Other Measures

- Relationship to the expectation of LRs in the validation set
  - $\mathrm{E}[LR]$ for true-θ_d values tends to be 1
  - $\mathrm{E}[1/LR]$ for true-θ_p values tends to be 1
- Empirical version in the validation set of LR values

$$1 = E\left[LR\right]\Big|_{true-\theta_d} \approx \frac{1}{N_d} \sum_{i \in true-\theta_d} LR_i$$

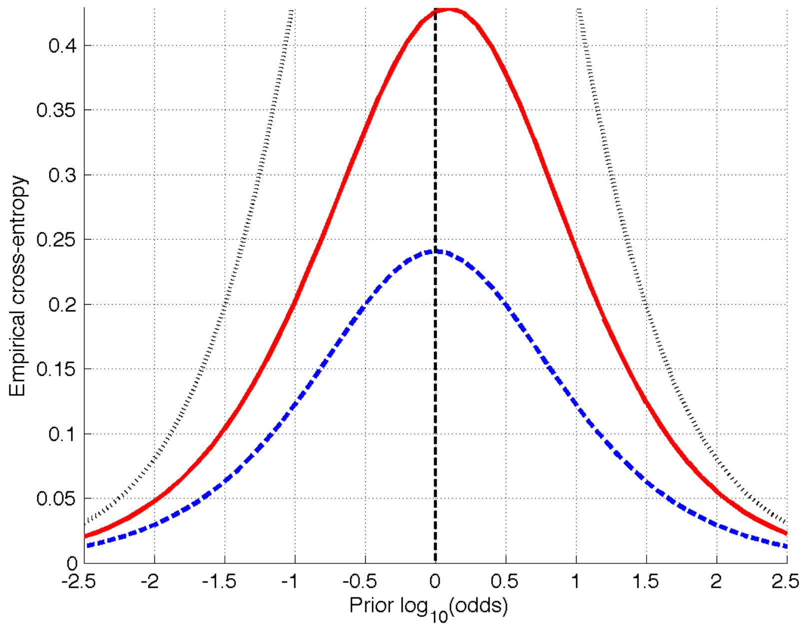
$$1 = E\left[\frac{1}{LR}\right]\Bigg|_{true-\theta_p} \approx \frac{1}{N_p} \sum_{j \in true-\theta_p} \frac{1}{LR_j}$$

- If the LRs are well-calibrated, this criterion tends to follow
  - Again, can be proof in some cases

Calibration and Other Measures

- Example with synthetic data:
  - Calibration improves the empirical expectation criteria

Badly-calibrated
E[LR]=0.60 for true-θ_d
E[1/LR]=0.22 for true-θ_p

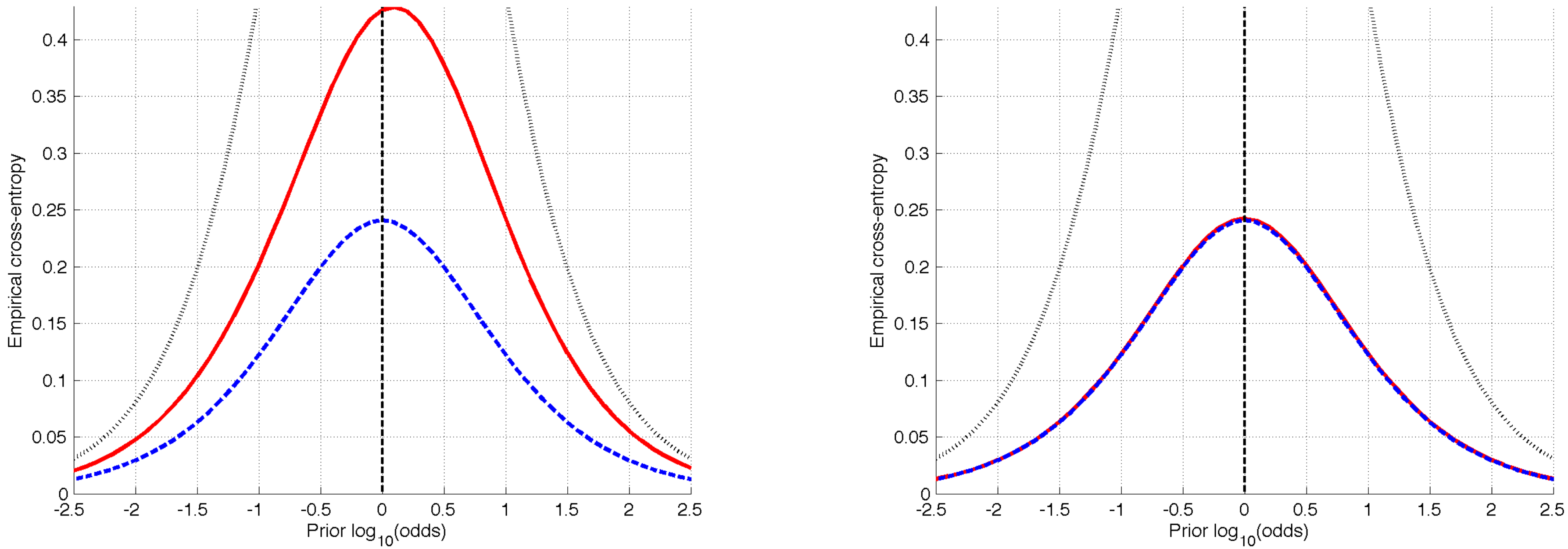


Calibration and Other Measures

- Example with synthetic data:
  - Calibration improves the empirical expectation criteria

Badly-calibrated
E[LR]=0.60 for true-θ_d
E[$1/LR$]=0.22 for true-θ_p

Well-calibrated
E[LR]=0.82 for true-θ_d
E[$1/LR$]=1.16 for true-θ_p



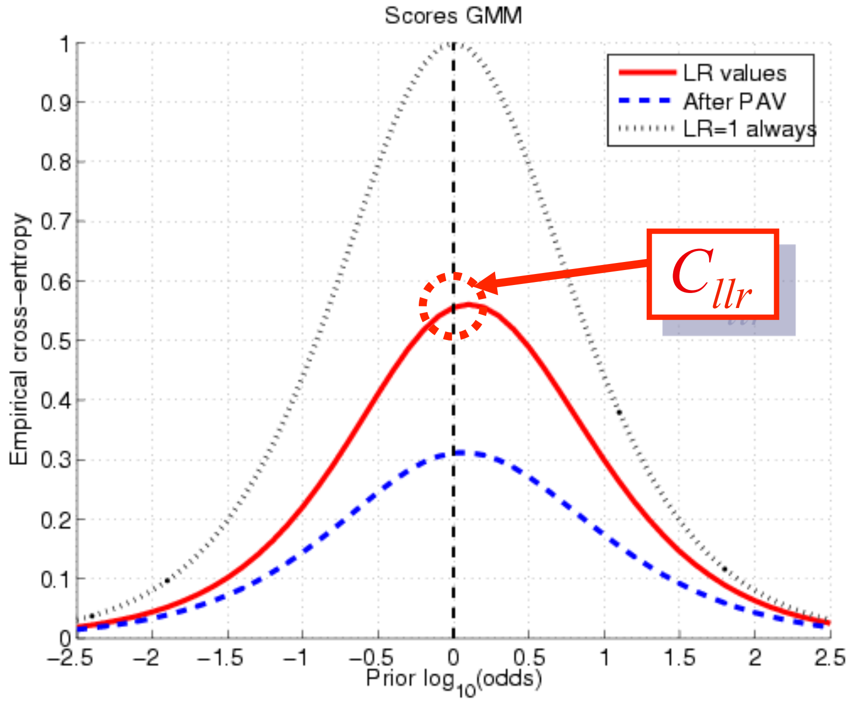
Empirical Cross-Entropy and C_{llr}

- C_{llr} also proposed as accuracy of a likelihood ratio set
  - Important decision-theoretical properties

Niko Brümmer [a,b,*], Johan du Preez [b]
Application-independent evaluation of speaker detection
Computer Speech and Language 20 (2006) 230–275

- C_{llr} and ECE are closely related

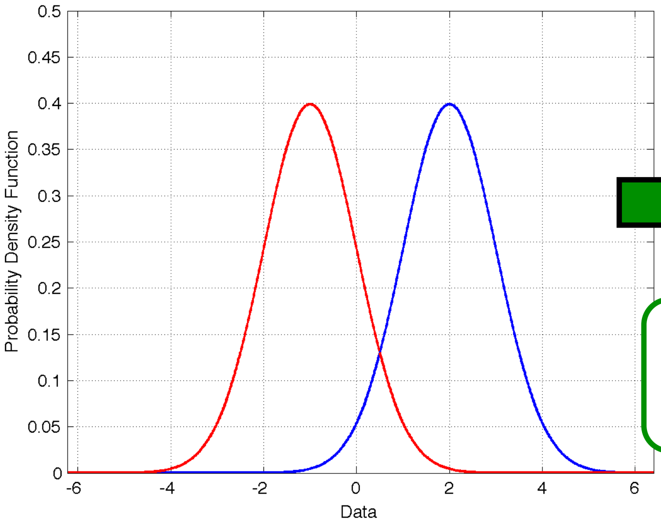
$$C_{llr} = ECE\Big|_{P(\theta_p|I)=0.5}$$



Calibration and Generative Models

- Example: known generative model of the data
- If the generating model is used for computing LR…
  - The resulting LR set will be well-calibrated!

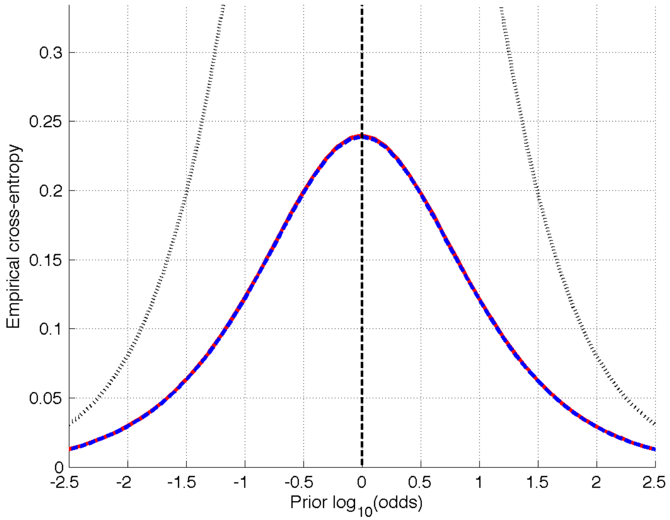
Solid: model generating data
Dashed: model for the LR



Calculate LRs

Using known generating model

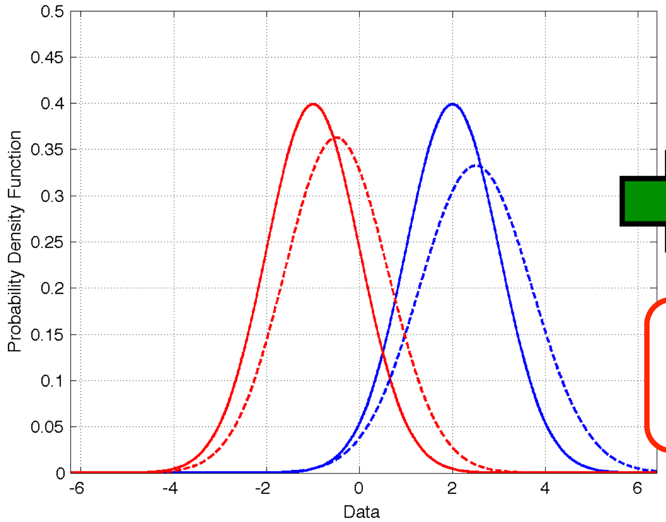
Well-calibrated LRs



Calibration and Generative Models

- Example: known generative model of the data
- If we use a different model…
  - Lack of calibration: warns about bad models!

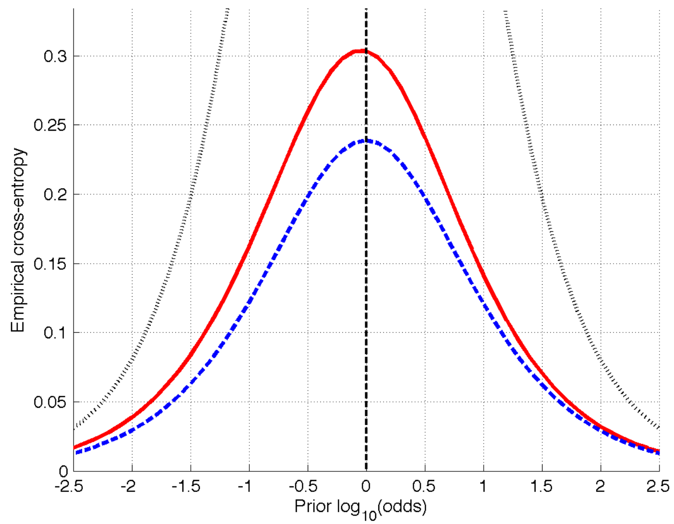
Solid: model generating data
Dashed: model for the LR



Calculate LRs

Using other model

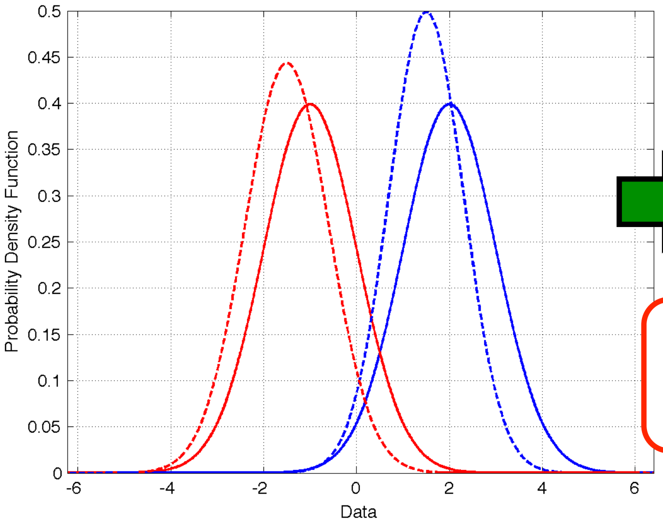
Badly-calibrated LRs



Calibration and Generative Models

- Example: known generative model of the data
- If we use a different model…
  - Lack of calibration: warns about bad models!

Solid: model generating data
Dashed: model for the LR



Calculate LRs

Using other model

Badly-calibrated LRs