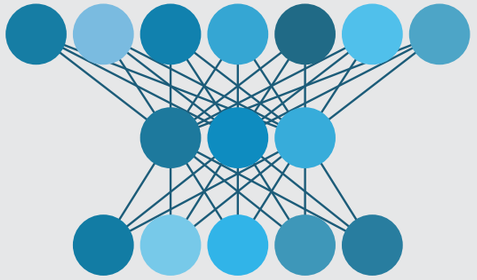
Calibration and validation of likelihood-ratio systems

Geoffrey Stewart Morrison

Forensic Data Science Laboratory
Aston University

Aston University
BIRMINGHAM UK



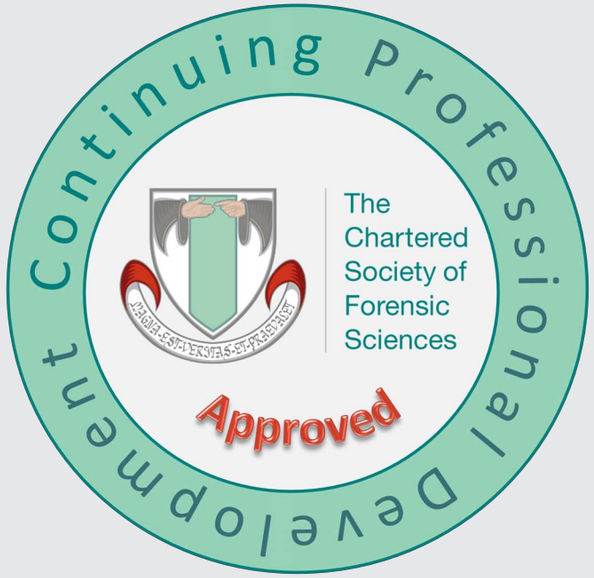
Workshop material

- Slides

https://forensic-data-science.net/#EAFS2025

Additional training

- Concepts of forensic inference and statistics
  - Master's level continuing professional development course
  - Online delivery with weekly interactive sessions
  - Delivered in 22 weeks spread over 6 months
  - ~1 day per week workload
  - Competency assessment



https://www.aston.ac.uk/study/courses/concepts-forensic-inference-and-statistics-standalone-module/

Contents

- Preliminaries
  - Black boxes
  - Logarithms
- Calibration
  - Calibration in weather forecasting
  - Calibration principles
  - Well-calibrated likelihood ratios
  - Calibration models
- Validation
  - Validation protocols
  - Validation metric (log-likelihood-ratio cost, C_{llr})
  - Validation graphic (Tippett plot)
- Calibration revisited
  - bi-Gaussianized calibration

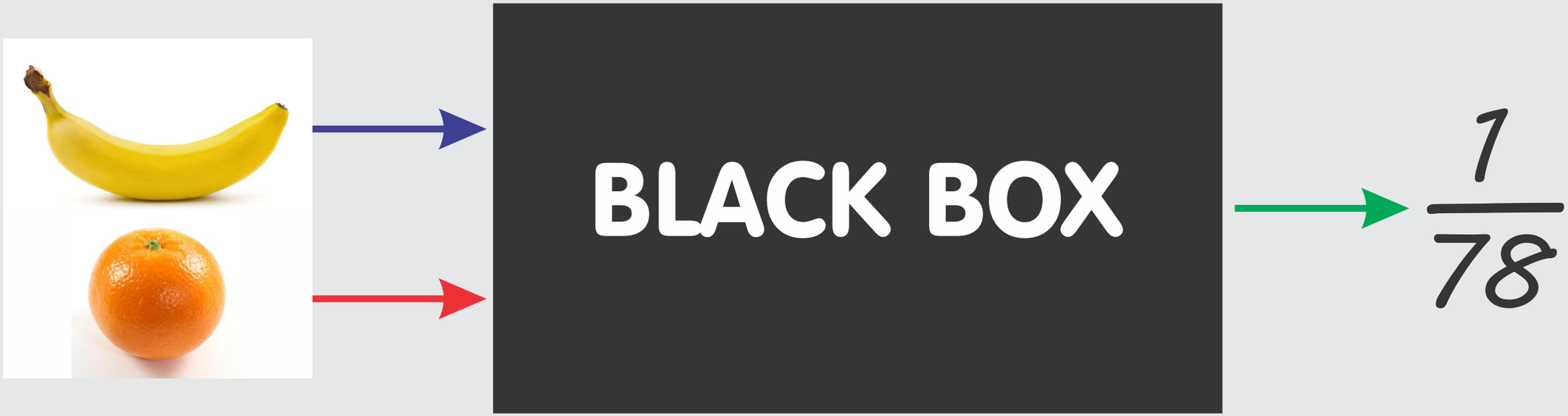
Preliminaries:

Black boxes

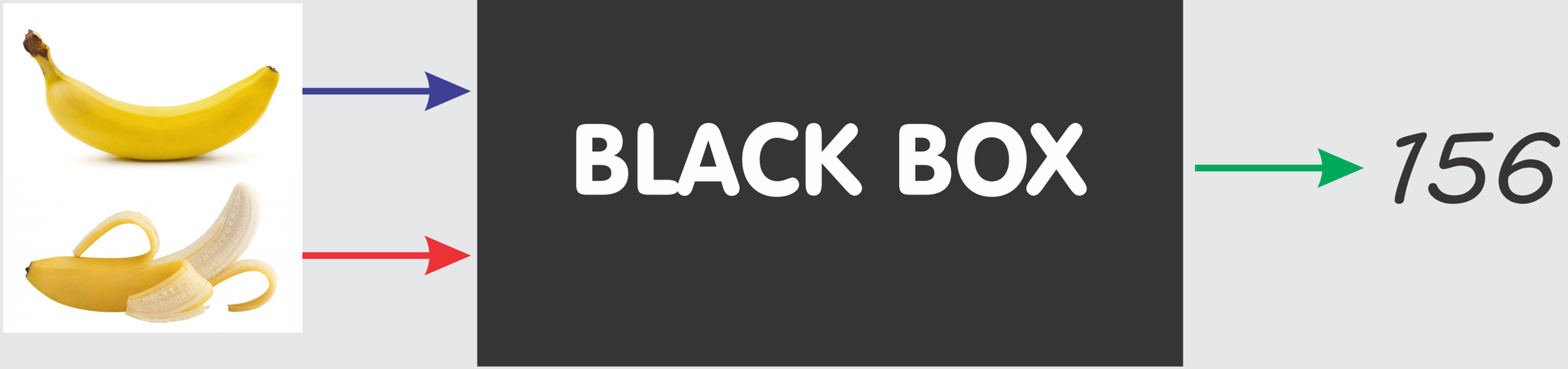
Preliminaries – black boxes

- Both calibration and validation treat forensic-evaluation systems as black boxes:
    - not concerned with what is inside the box
    - only with what the box outputs in response to inputs

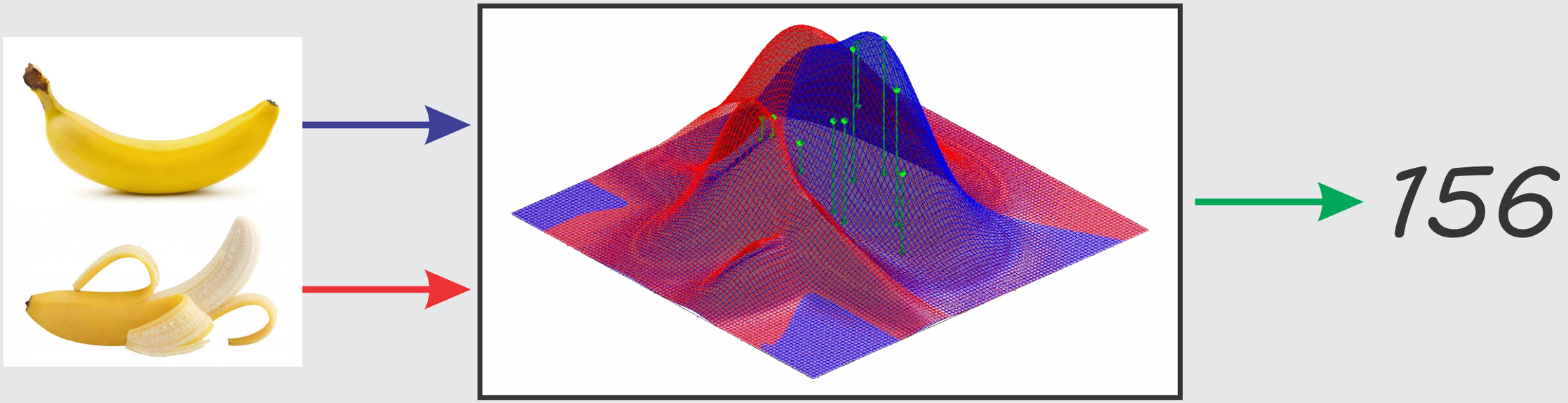
Preliminaries – black boxes



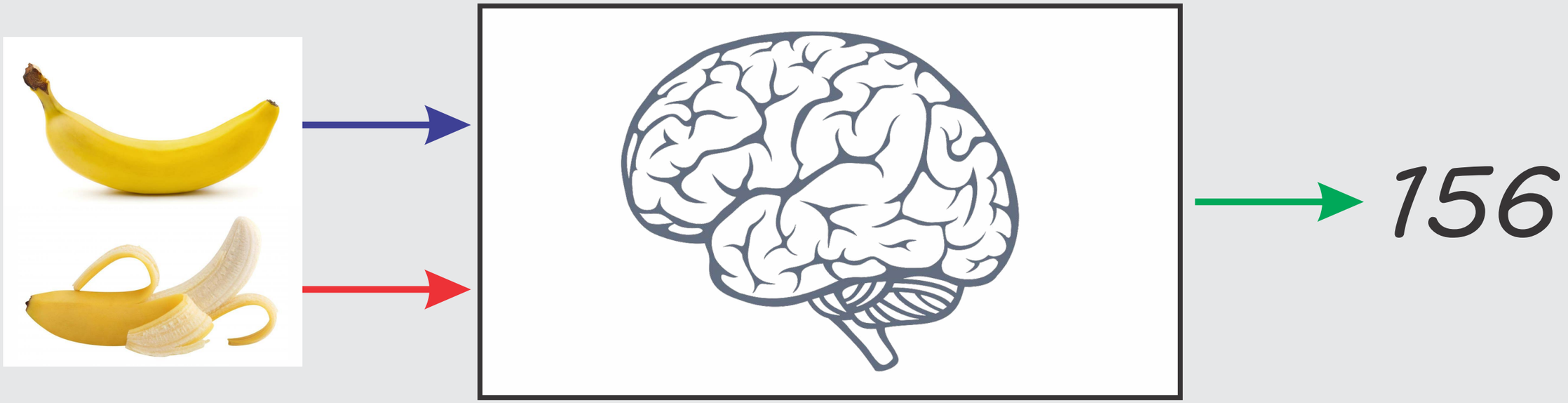
Preliminaries – black boxes



Preliminaries – black boxes



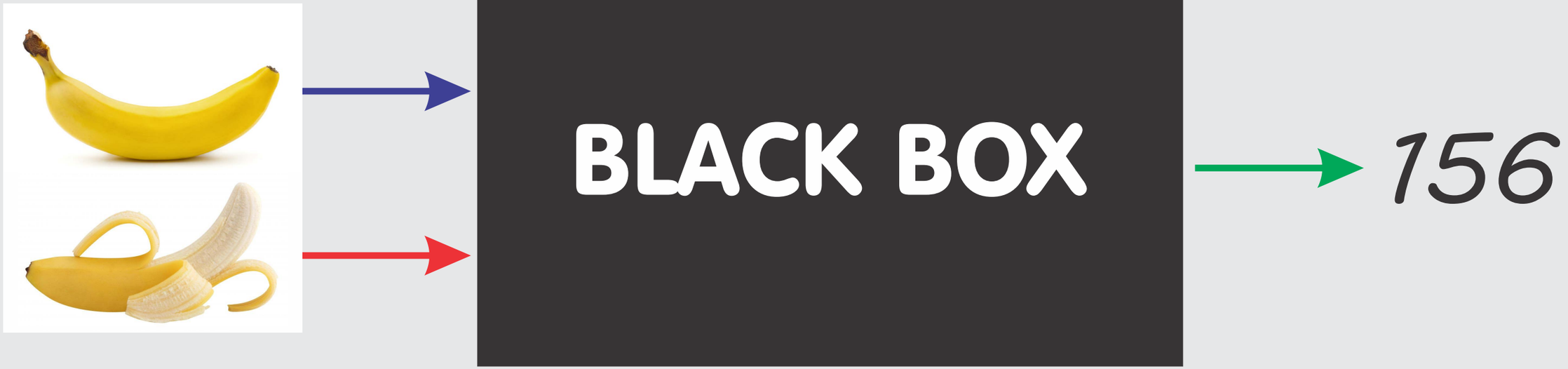
Preliminaries – black boxes



Preliminaries – black boxes

156

Preliminaries – black boxes



Preliminaries:

Logarithms

Preliminaries – logarithms

- Base 10 logarithms

			LR			
1/1000	**1/100**	**1/10**	**1**	**10**	**100**	**1000**
0.001	**0.01**	**0.1**	**1**	**10**	**100**	**1000**
$\mathbf{10^{-3}}$	$\mathbf{10^{-2}}$	$\mathbf{10^{-1}}$	$\mathbf{10^{0}}$	$\mathbf{10^{1}}$	$\mathbf{10^{2}}$	$\mathbf{10^{3}}$
			$\log_{10}(\mathrm{LR})$			
−3	**−2**	**−1**	**0**	**+1**	**+2**	**+3**

Preliminaries – logarithms

- Base 2 logarithms

LR						
1/8	**1/4**	**1/2**	**1**	**2**	**4**	**8**
0.125	**0.25**	**0.5**	**1**	**2**	**4**	**8**
$\mathbf{2^{-3}}$	$\mathbf{2^{-2}}$	$\mathbf{2^{-1}}$	$\mathbf{2^{0}}$	$\mathbf{2^{1}}$	$\mathbf{2^{2}}$	$\mathbf{2^{3}}$
$\log_2(\mathrm{LR})$						
−3	**−2**	**−1**	**0**	**+1**	**+2**	**+3**

Preliminaries – logarithms

- Natural logarithms
  - ln = $\log_e$
  - $e \approx 2.718$ (Euler's number)

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

Preliminaries – logarithms

Likelihood ratios

- in favour of the denominator hypothesis are in the range:

  0 to 1

- in favour of the numerator hypothesis are in the range:

  1 to $+\infty$

Log likelihood ratios

- in favour of the denominator hypothesis are in the range:

  $-\infty$ to 0

- in favour of the numerator hypothesis are in the range:

  0 to $+\infty$

Calibration in weather forecasting

Calibration in weather forecasting

- Weather forecaster predicts:
    - Probability of precipitation for tomorrow is 40%.
- The next day it either rains or it doesn't rain.
- Looking at lots of days for which the weather forecaster's PoP was 40%, on what percentage of those days did it actually rain?

Calibration in weather forecasting

Well calibrated:

- Prediction: 40%
- Actual: 40%

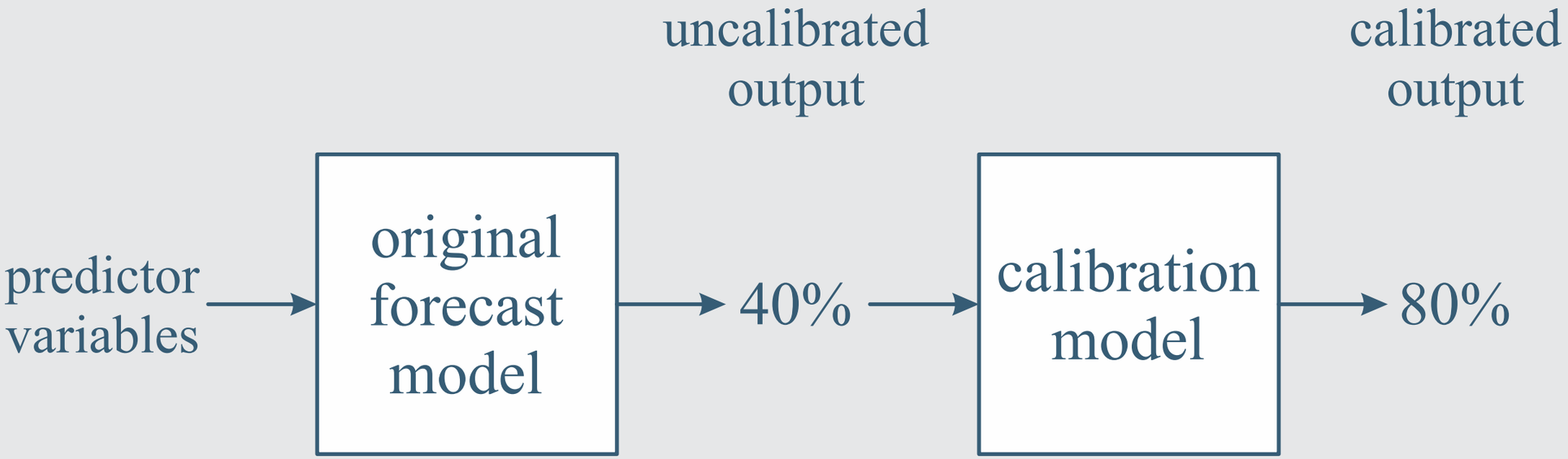
Not well calibrated:

- Prediction: 40%
- Actual: 80%

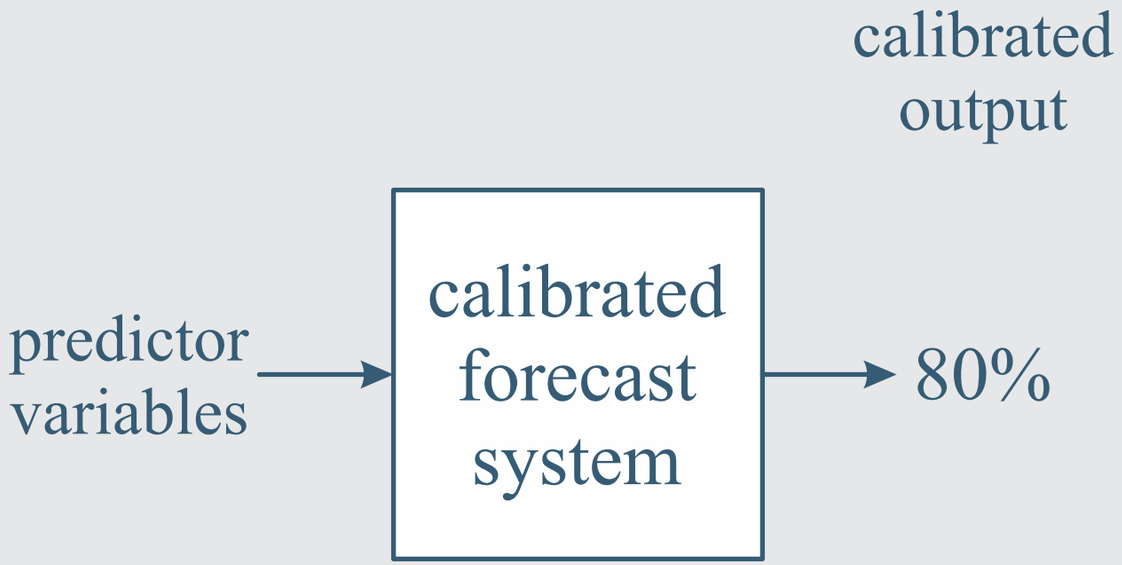
Calibration in weather forecasting

- Solution:
  - Collect data from a large number of past days.
  - For each day collect: **prediction** **actual weather**
  - Use those data to train a calibration model.
  - Use the model to calibrate future predictions.

Calibration in weather forecasting



Calibration in weather forecasting



Calibration principles

Calibration principles

- If:
    - a model is a parsimonious parametric model
    - there is a large amount of training data relative to the number of parameter values to be estimated
    - the data are representative of the relevant population
    - the assumptions of the model are not violated by the population distributions
- Then the output of the model will be well calibrated

Calibration principles

- In forensic science:
    - Models often fit complex distributions to high-dimensional data
    - The amount of case-relevant training data is often small relative to the number of parameter values to be estimated
    - The assumptions of the models may be violated
    - Therefore:
        - The outputs of the models are often not well calibrated

Calibration principles

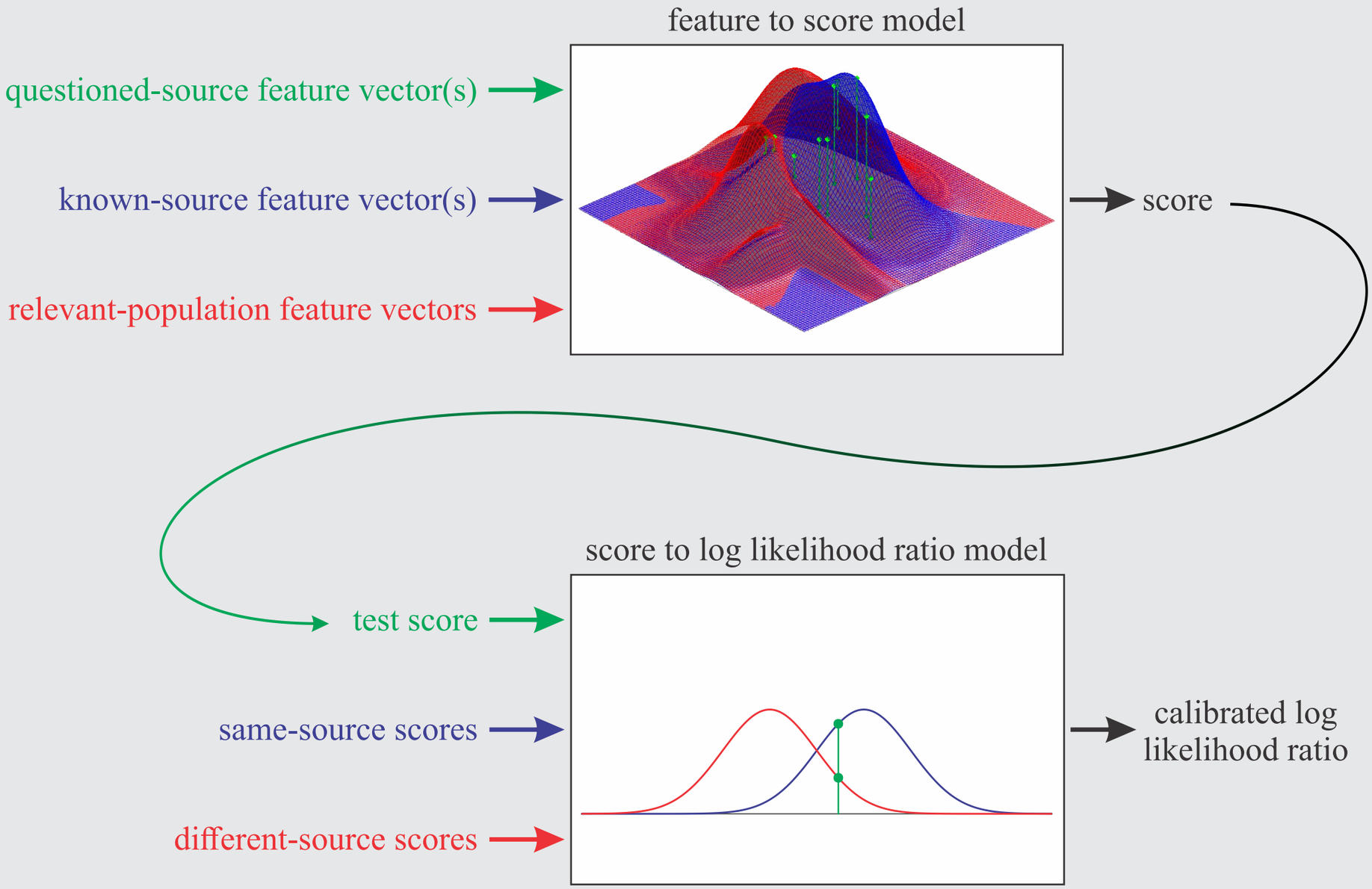
- Solution:
    - Treat the output of the first (complex) model as an uncalibrated log likelihood ratio (a score)
    - Use a parsimonious model to convert the score to a calibrated log likelihood ratio

Vocabulary:

“score” = “uncalibrated log likelihood ratio”

“score” ≠ “similarity score”

Calibration principles



Calibration principles

- Take data that:
    - represent the relevant population in the case
    - reflect the conditions of the questioned-source and known-source items in the case
- Construct same-source pairs and different-source pairs
- Use the first model to calculate a score for each pair
- Use the resulting same-source scores and different-source scores to train the calibration model

Calibration principles

- The scores are unidimensional
- The calibration model is parsimonious
- There is a large amount of data relative to the number of parameter values to be estimated
- Therefore:
    - The output of the calibration model is well calibrated

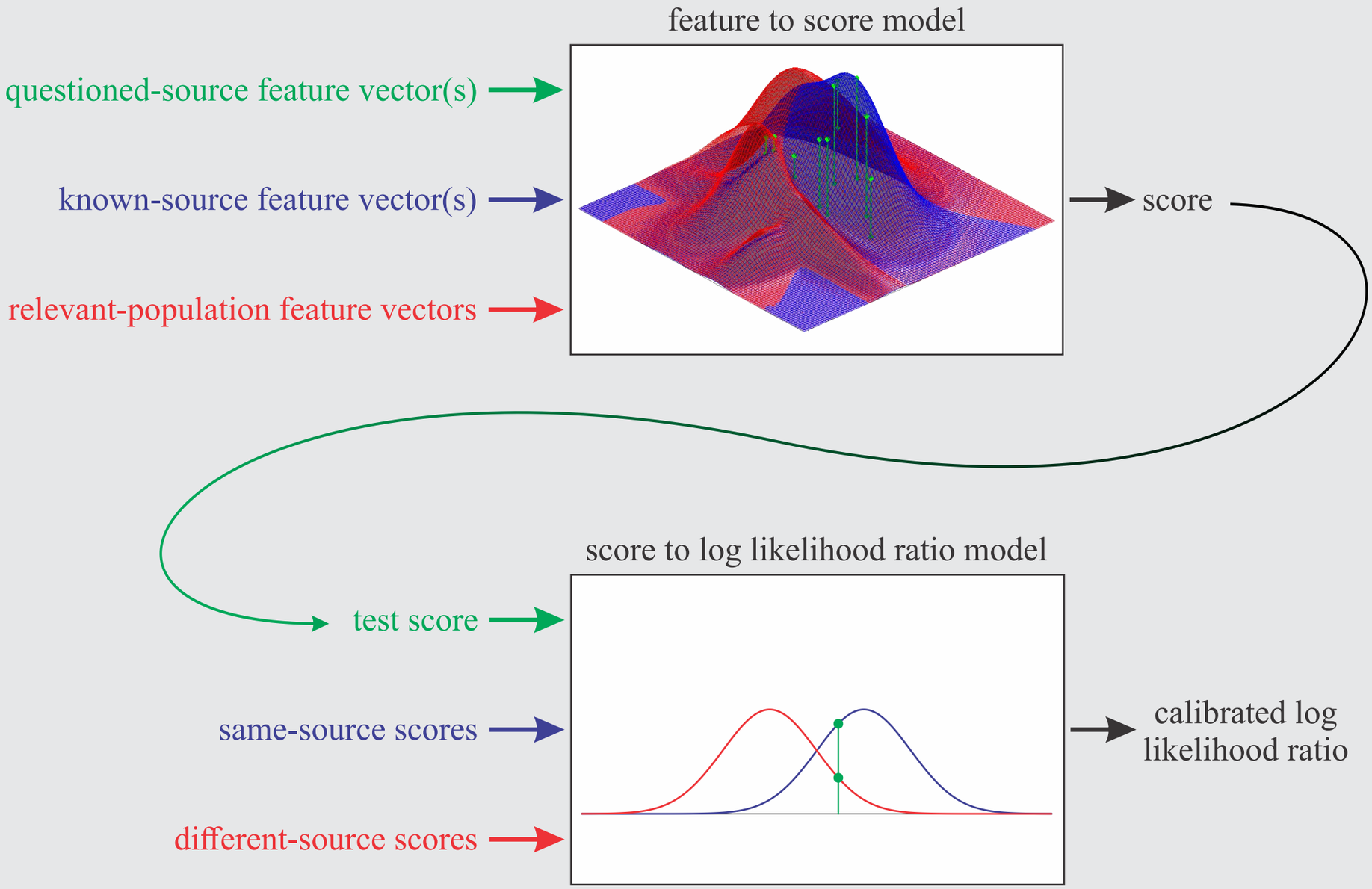
Calibration principles

- Important condition:
    - The data used for training the calibration model must:
        - represent the relevant population in the case
            - including there being enough data
        - reflect the conditions of the questioned-source and known-source items in the case
            - including any mismatches in conditions
    - If not, the system will be miscalibrated

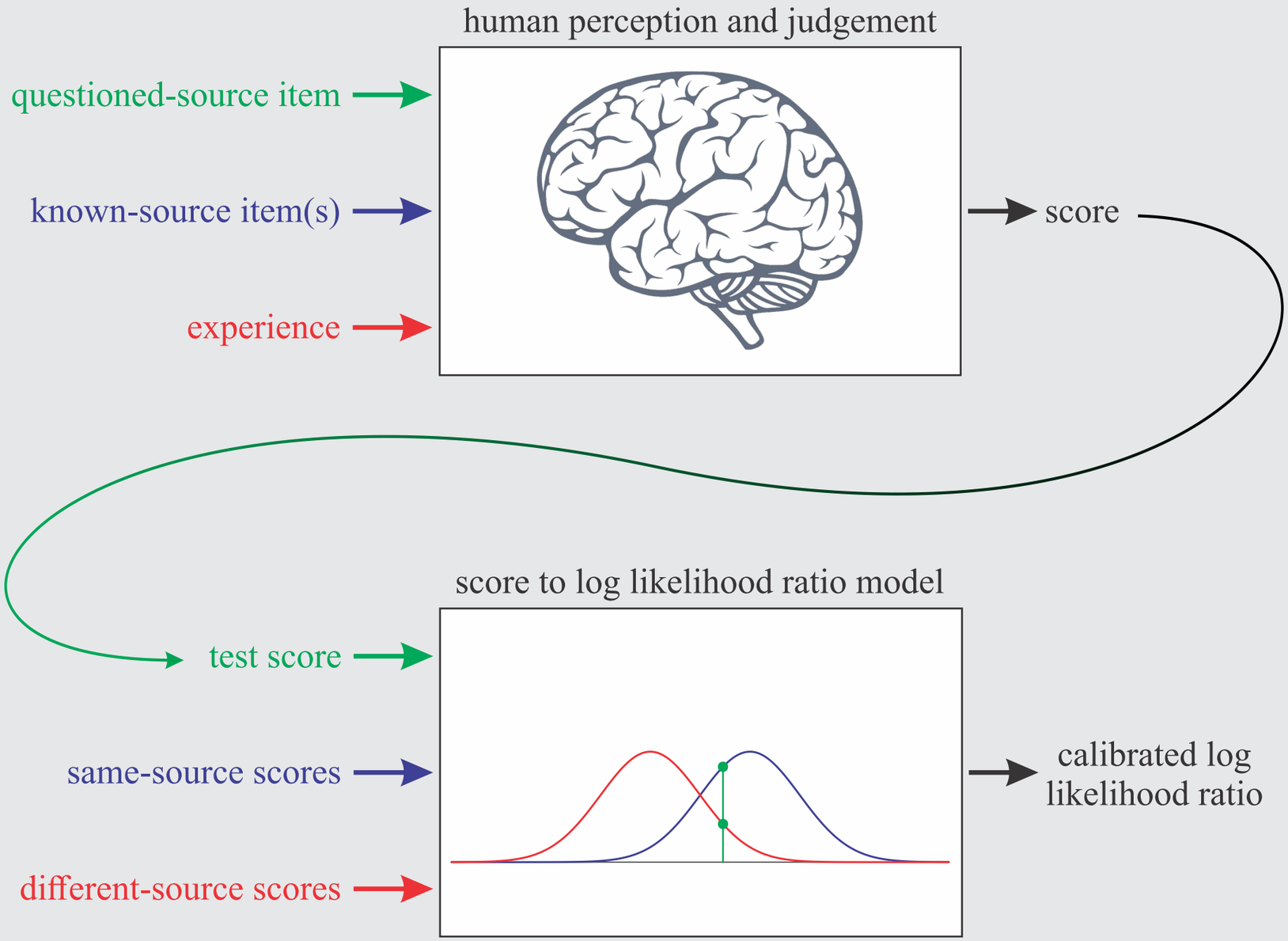
Calibration principles

- Important condition:
  - The first model must output scores which are **uncalibrated log likelihood ratios**. They must take account of both:
    - the **similarity** between the questioned-source and the known-source items
    - their **typicality** with respect to the relevant population
  - Similarity-only scores cannot be used

Calibration principles



Calibration principles



Well-calibrated likelihood ratios

Well-calibrated scales

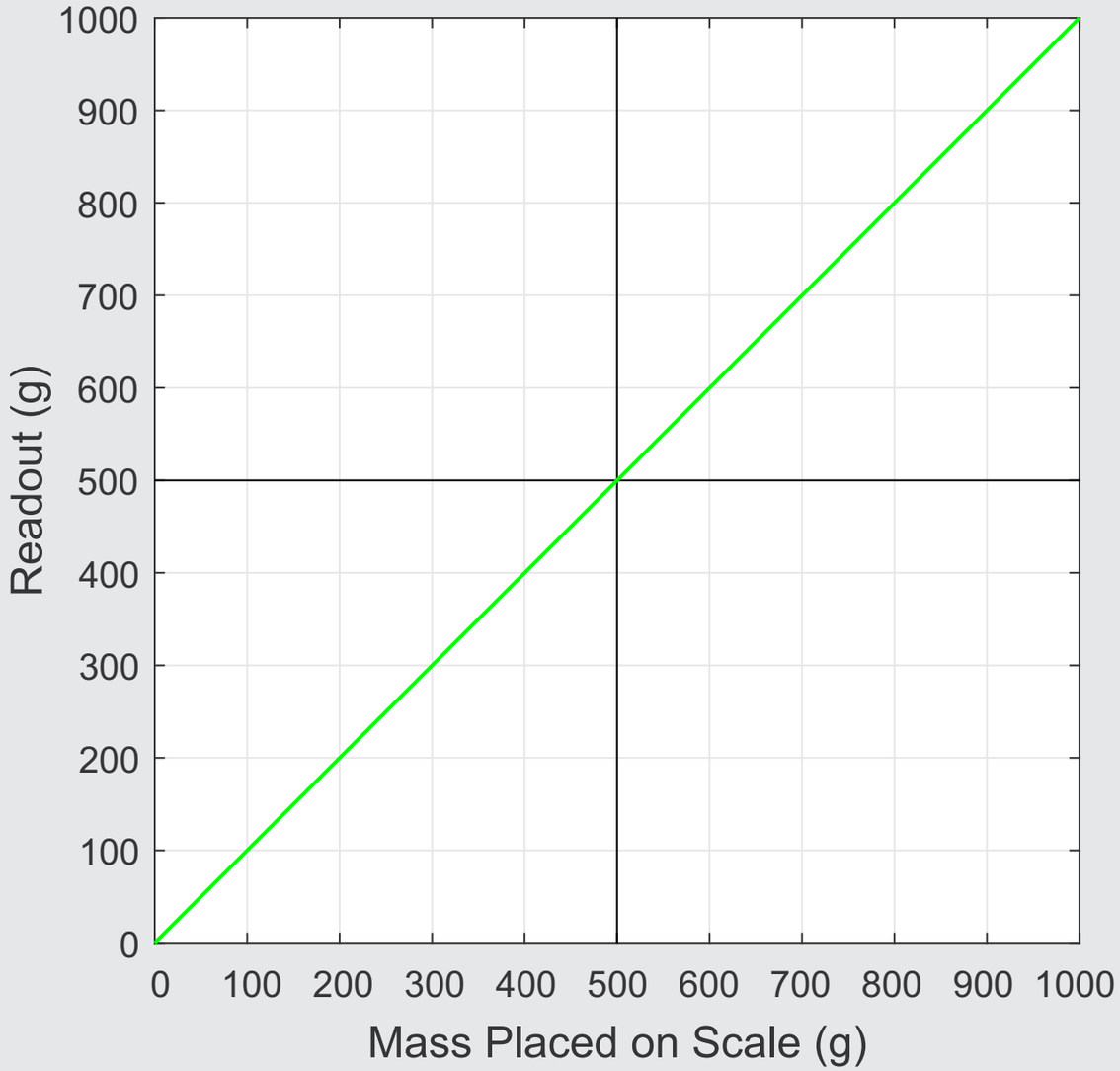
- What is a well-calibrated set of scales?

- A set of scales for which:

    - The mass stated in the readout is the same as the mass placed on the scale

Well-calibrated scales

- Calibration is the process of adjusting the set of scales so that its output is well calibrated.

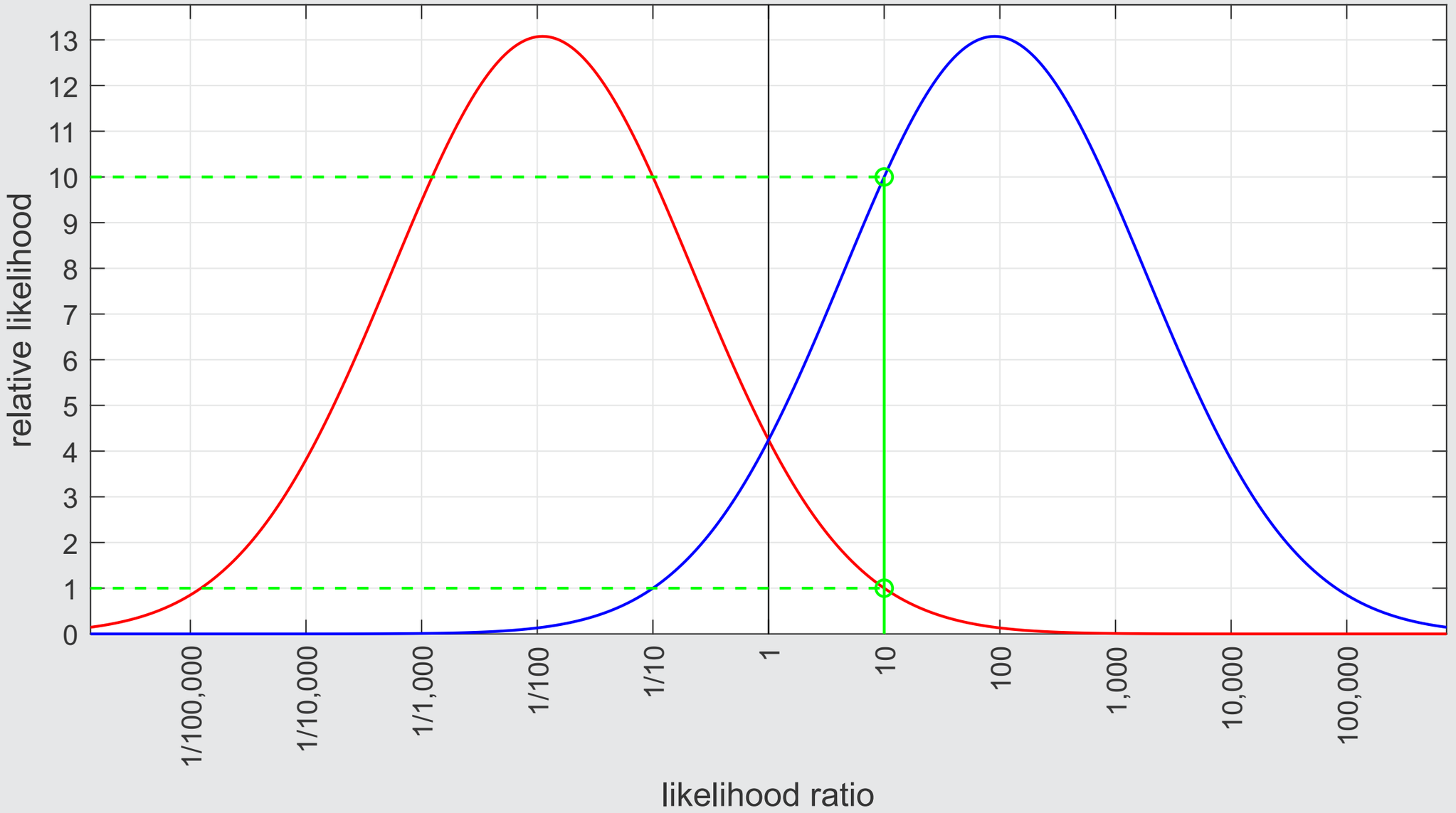


Well-calibrated likelihood ratios

- What is a well-calibrated likelihood-ratio system?
    - The likelihood ratio of the likelihood ratio is the likelihood ratio

$$LR = \frac{f(LR \mid H_s)}{f(LR \mid H_d)}$$

Well-calibrated likelihood ratios

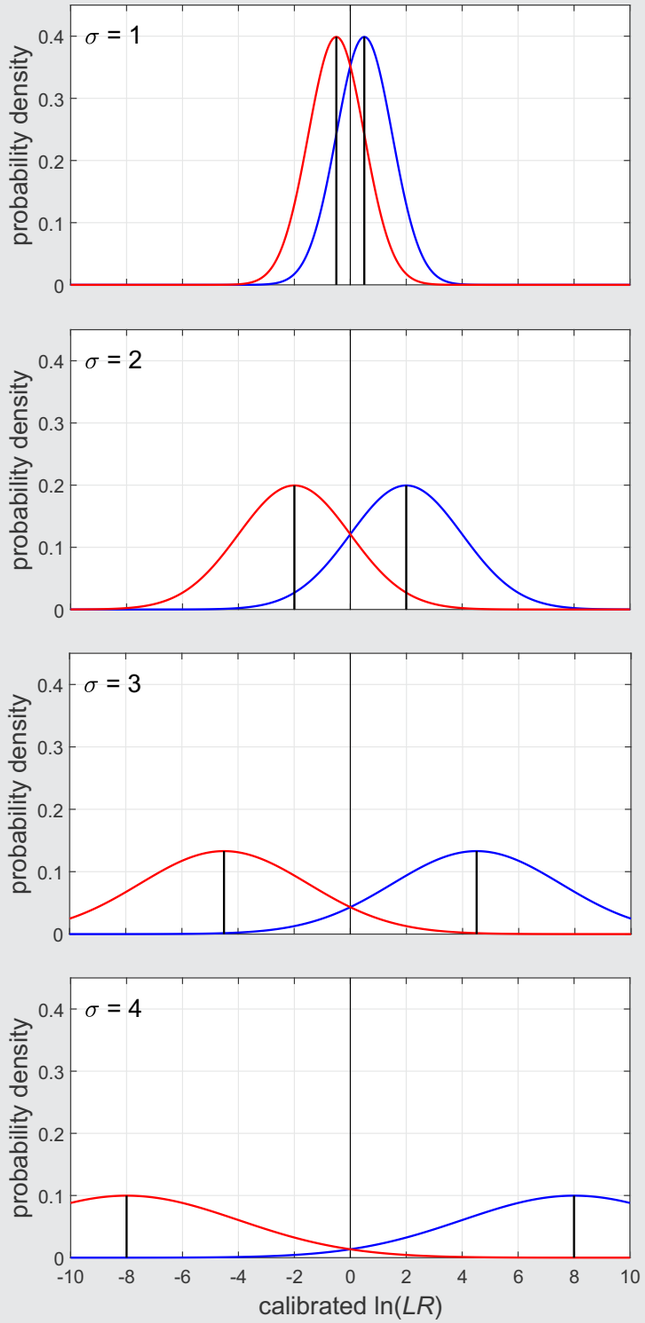


Well-calibrated likelihood ratios

- Perfectly calibrated ln(LR) distributions

- Both same-source and different-source distributions are Gaussian, and they have the same variance

$$\mu_{\mathrm{d}} = -\frac{\sigma^2}{2} \qquad \mu_{\mathrm{s}} = +\frac{\sigma^2}{2}$$



Calibration models

Calibration models

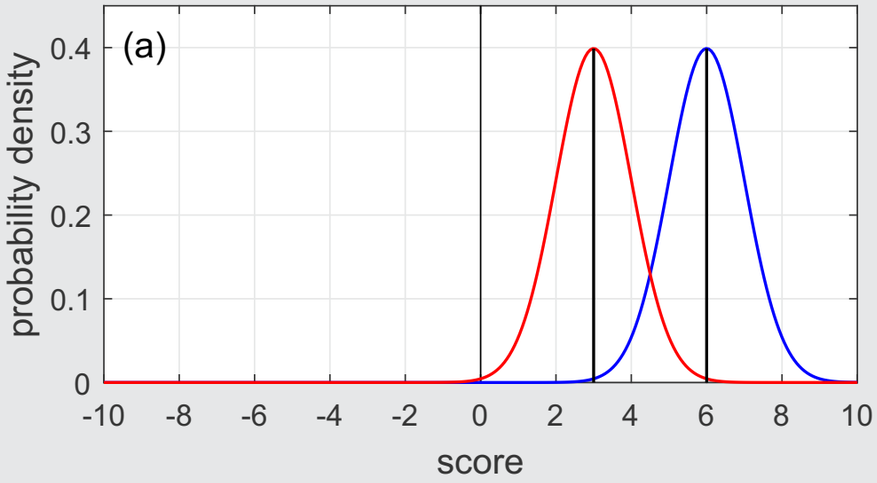
(a)

Uncalibrated scores

$\mu_d = 3$

$\mu_s = 6$

$\sigma = 1$



Calibration models

(a)

Uncalibrated scores

$\mu_d = 3$

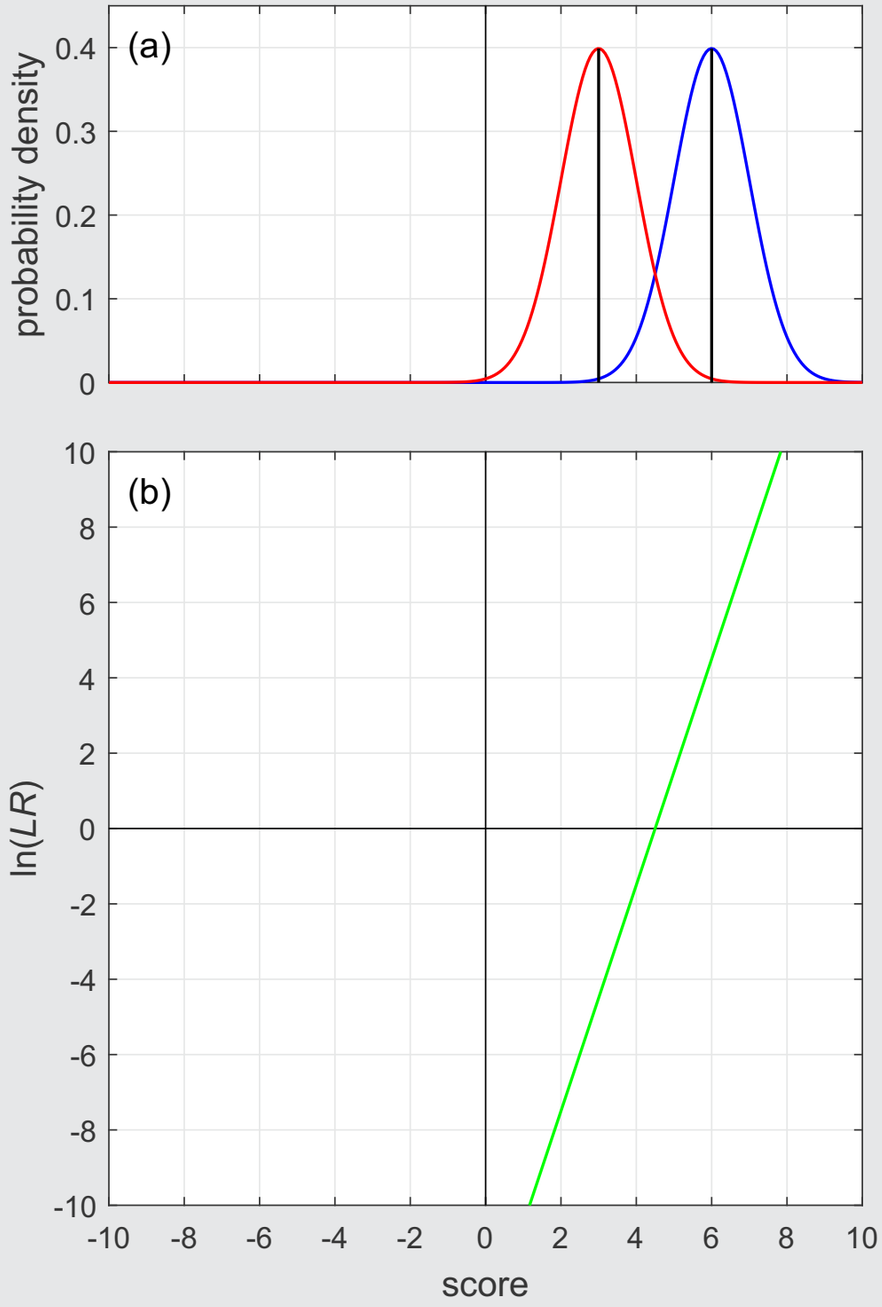
$\mu_s = 6$

$\sigma = 1$

(b)

Score to ln(LR)

mapping function



Calibration models

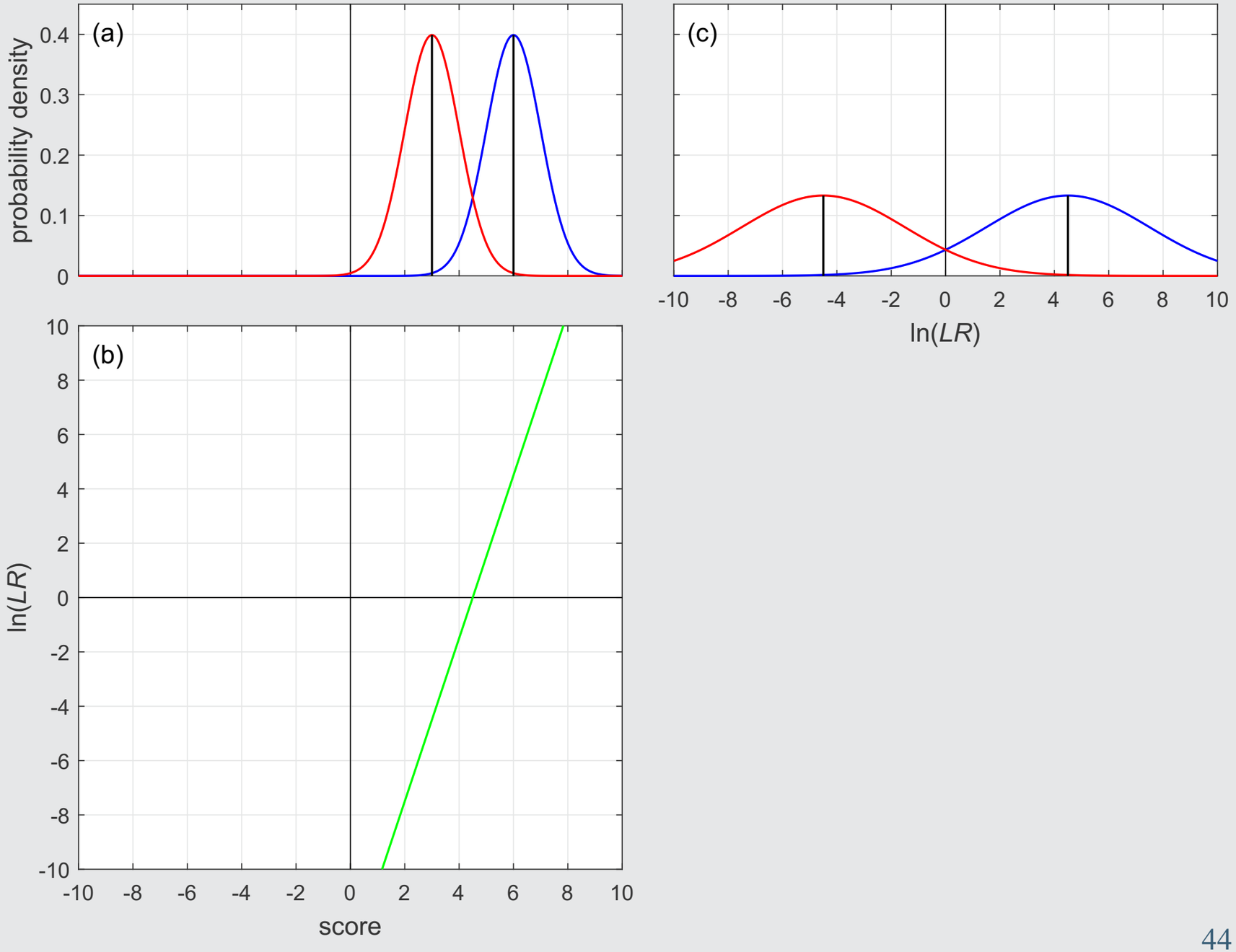
(c)

Calibrated ln(LR)

$\mu_d = -4.5$

$\mu_s = +4.5$

$\sigma = 3$



Calibration models

(c)

Calibrated ln(LR)

$\mu_d = -4.5$

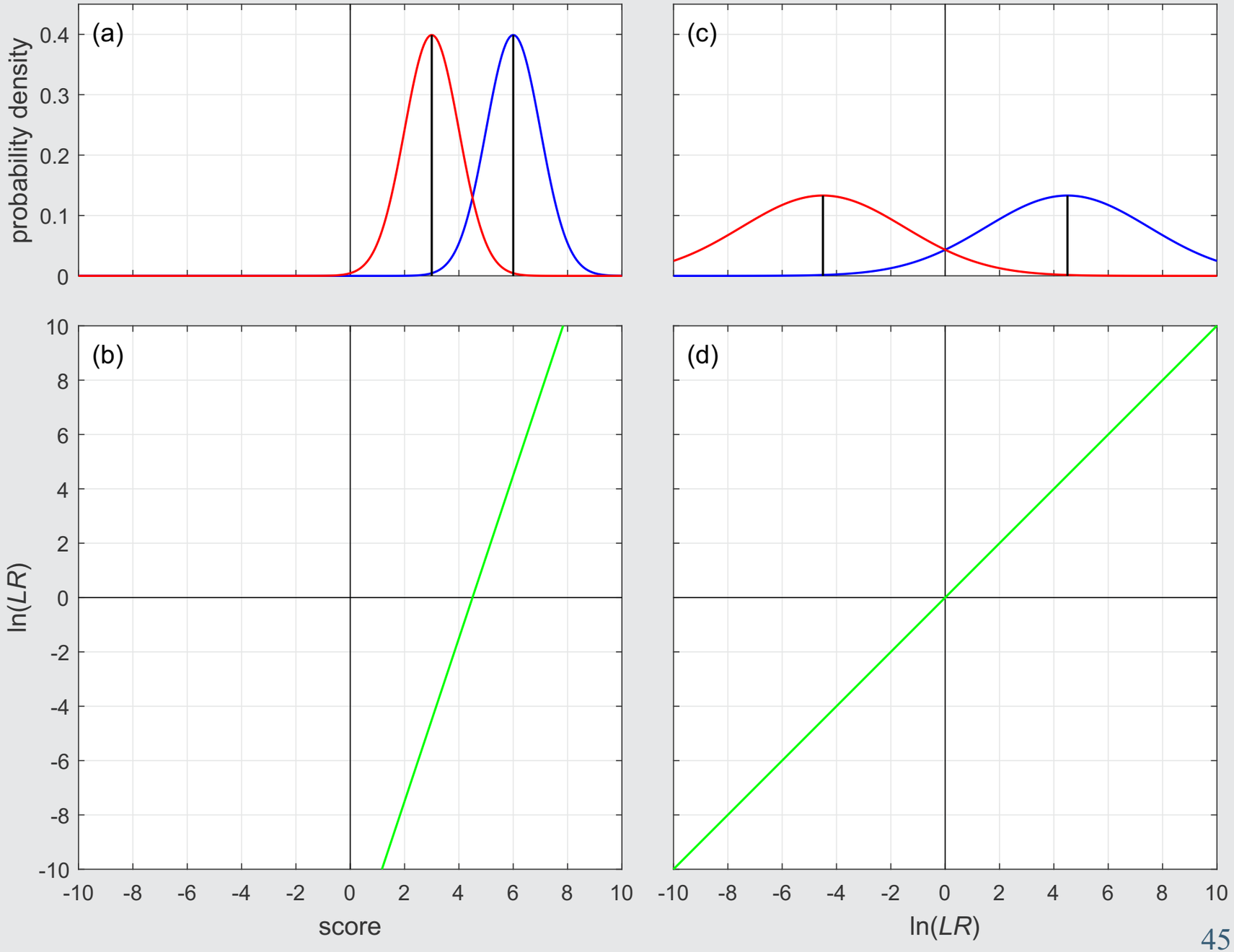
$\mu_s = +4.5$

$\sigma = 3$

(d)

ln(LR) to ln(LR)

mapping function



Calibration models

- Score [x] to ln(LR) [y] mapping function:

$$y = a + bx$$

$$a = -b\frac{\mu_s + \mu_d}{2} \qquad b = \frac{\mu_s - \mu_d}{\sigma^2}$$

- Where μ_s, μ_d, σ are the statistics for the scores

Calibration models

- Score [x] to ln(LR) [y] mapping function:

$$y = a + bx$$

- In practice, **logistic regression** is commonly used to calculate a and b

- It is more robust to violations of the assumptions of Gaussian distributions with the same variance

Validation protocols

Validation protocols

- Take data that:
    - represent the relevant population in the case
    - reflect the conditions of the questioned-source and known-source items in the case
- Construct same-source pairs and different-source pairs
- Use the calibrated forensic-evaluation system to calculate a likelihood ratio for each pair
- Assess how good each output is given knowledge of whether the corresponding input was a same-source pair or a difference-source pair

Validation protocols

- Important condition:
    - The data used for training the calibration model must:
        - represent the relevant population in the case
            - including there being enough data
        - reflect the conditions of the questioned-source and known-source items in the case
            - including any mismatches in conditions
    - If not, the results will not be indicative of how well the forensic-evaluation system works in the context of the case

Validation protocols

- If you have suitable data for calibration, you also have suitable data for validation, and vice versa:
    - Cross-validation:
        - leave-one-source out (for same-source comparisons)
        - leave-two-sources out (for different-source comparisons)

Validation metric

log-likelihood-ratio cost (C_{llr})

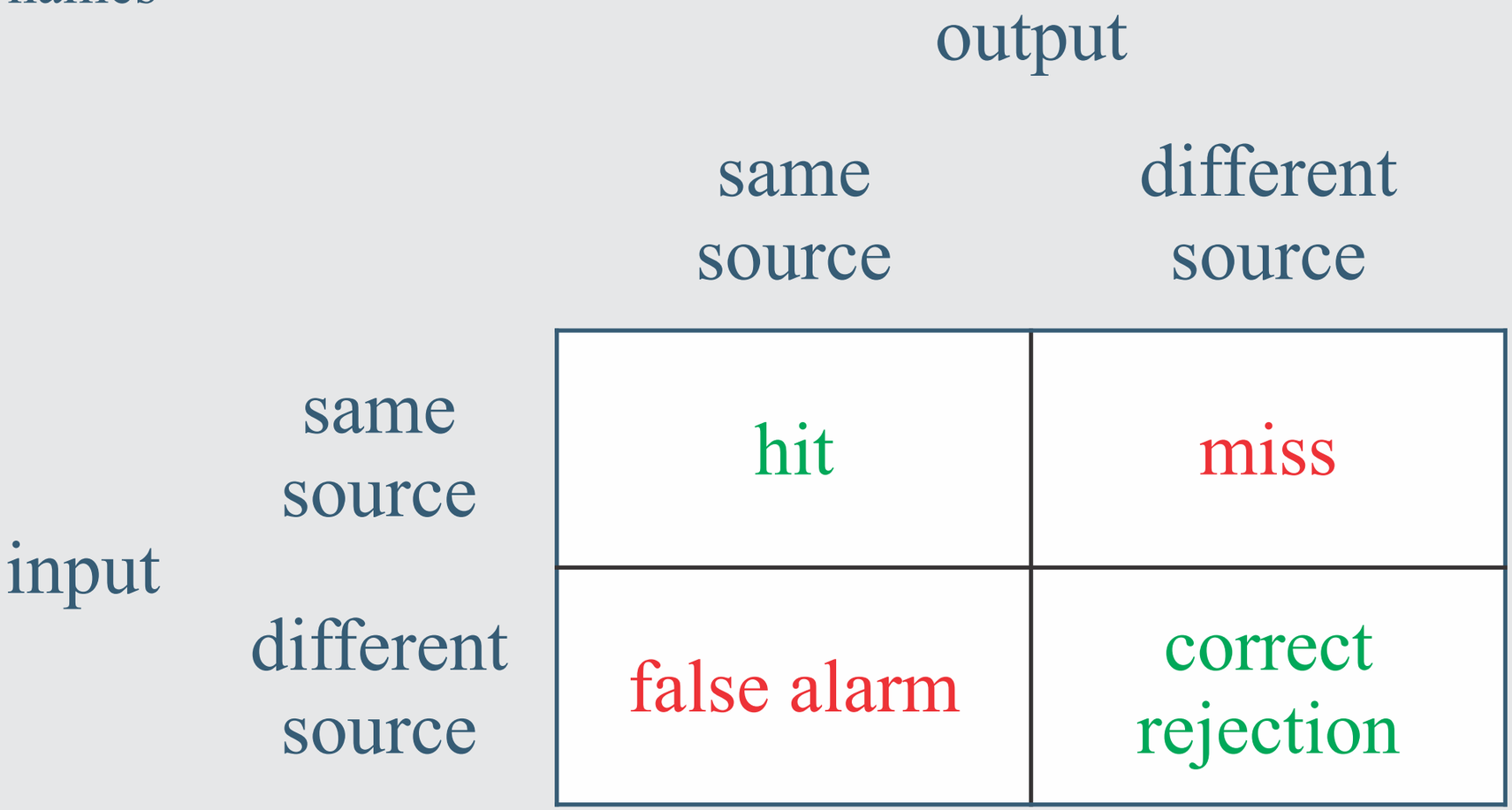
Validation metric

- Classification-error rate

input \ output	same source	different source
same source	correct	incorrect
different source	incorrect	correct

Validation metric

- Classification-error rate
  - names



Validation metric

- Classification-error rate
  - penalty values

output

input	same source	different source
same source	0	1
different source	1	0

Validation metric

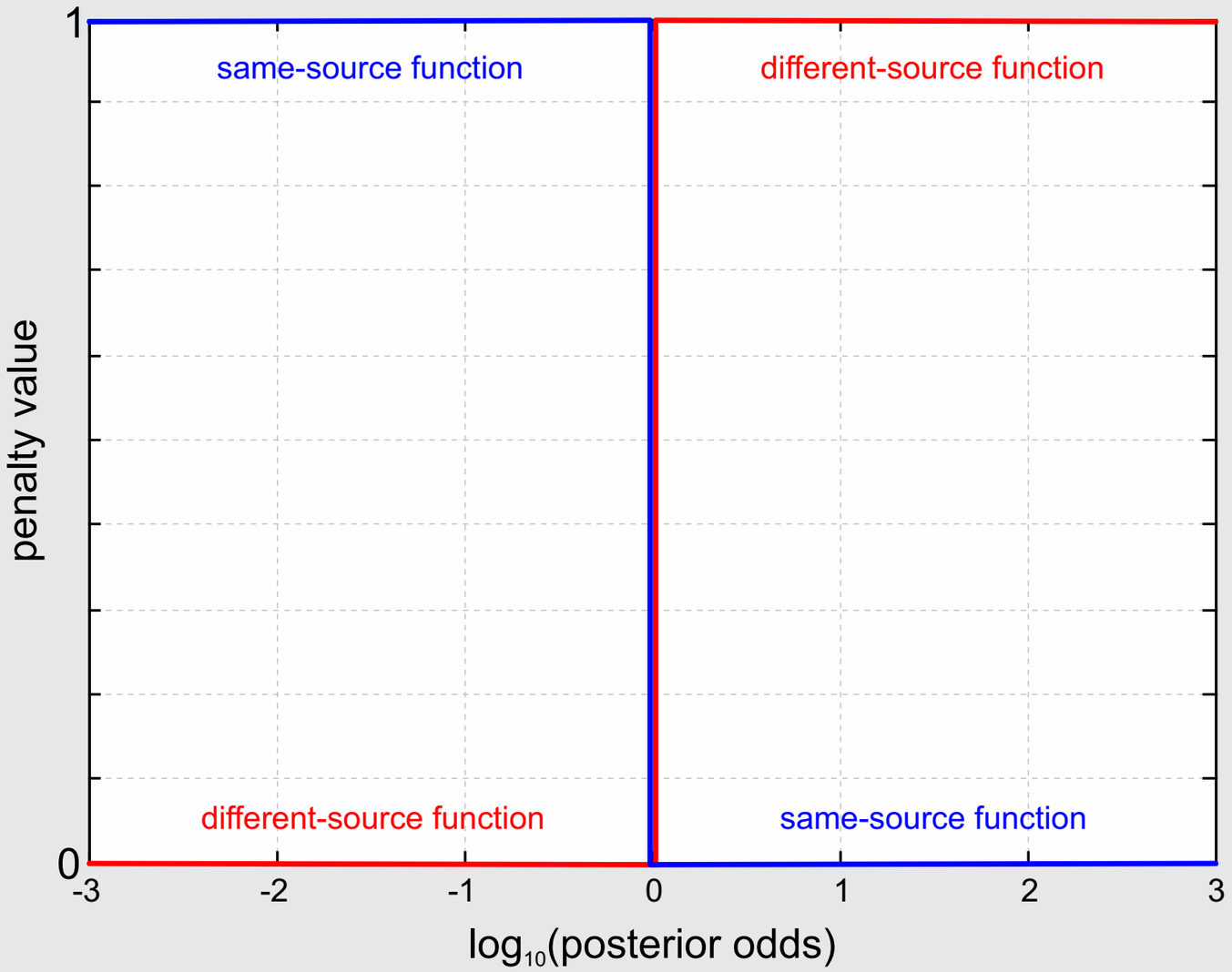
- Classification-error rate
  - formula

$$E_{\text{class}} = \frac{1}{2}\left(\frac{1}{N_{\text{s}}} \sum_{i=1}^{N_{\text{s}}} \begin{pmatrix} 0 \text{ if } y_i = \text{s} \\ 1 \text{ if } y_i = \text{d} \end{pmatrix} + \frac{1}{N_{\text{d}}} \sum_{j=1}^{N_{\text{d}}} \begin{pmatrix} 1 \text{ if } y_j = \text{s} \\ 0 \text{ if } y_j = \text{d} \end{pmatrix} \right)$$

miss: $y_i = \text{d}$ false alarm: $y_j = \text{s}$

Validation metric

- Penalty functions for calculating classification-error rate



Validation metric

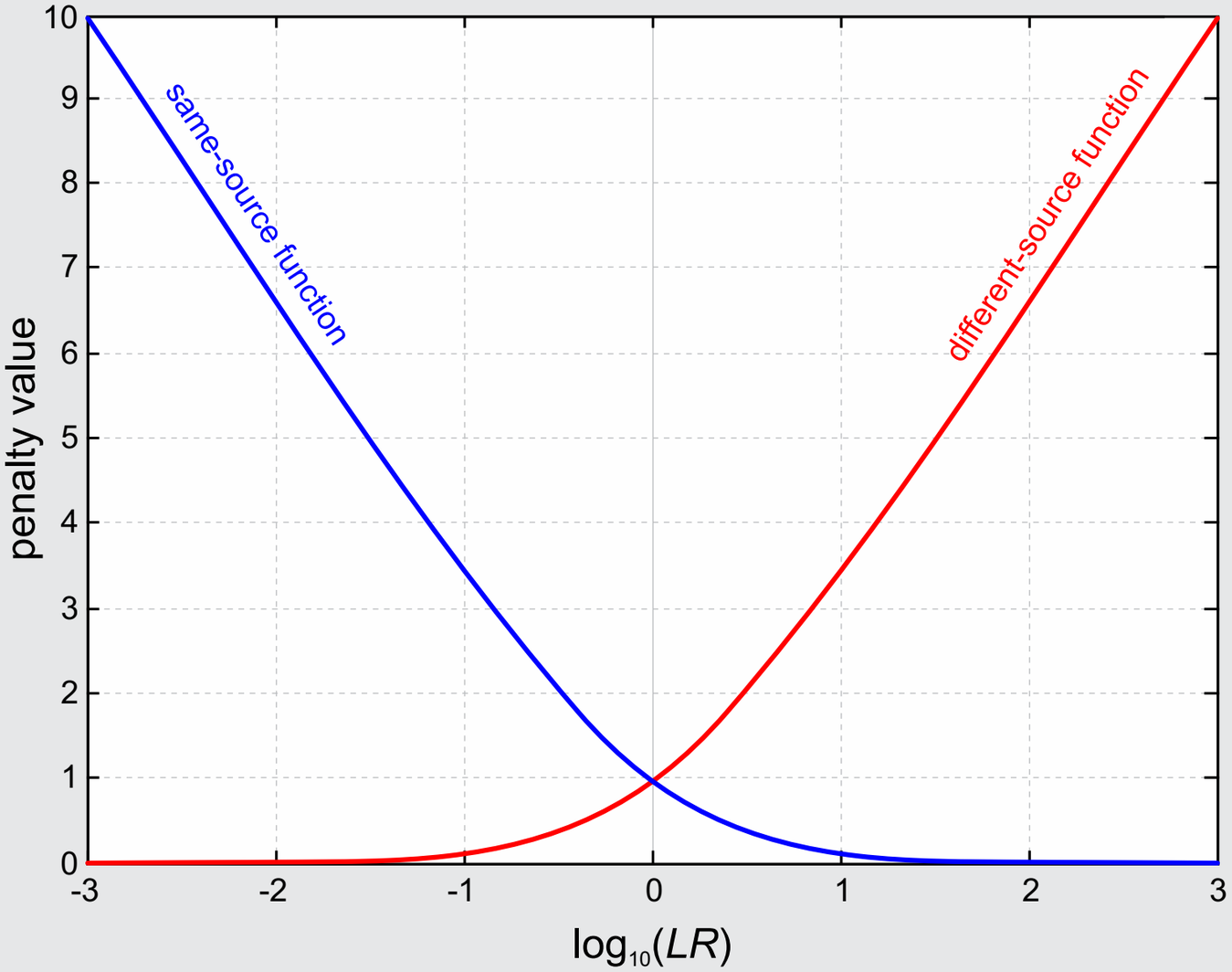
- Classification-error rate is not appropriate for assessing the performance of a system that outputs likelihood ratios because it is based on a **threshold applied to posterior probabilities**
  - It is not appropriate for a forensic practitioner to assess posterior probabilities
  - A threshold introduces a cliff-edge effect:
    - two values close to each other but on opposite sides of the threshold get treated differently
    - two values far from each other but on the same side of the threshold get treated the same

Validation metric

- For a system that outputs likelihood ratios, a metric of performance should be based on **likelihood-ratio values**
  - given a **same-source** input pair
    - the **larger** the likelihood-ratio value the **better** the performance
  - given a **different-source** input pair
    - the **smaller** the likelihood-ratio value the **better** the performance

Validation metric

- Penalty functions for calculating the **log-likelihood-ratio cost (C_{llr})**



Validation metric

- Formula for calculating C_{llr}

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{\text{s}}} \sum_{i=1}^{N_{\text{s}}} \log_2\left(1 + \frac{1}{LR_{\text{s}_i}}\right) + \frac{1}{N_{\text{d}}} \sum_{j=1}^{N_{\text{d}}} \log_2\left(1 + LR_{\text{d}_j}\right) \right)$$

Validation metric

- The **better the performance** of the system, the **smaller the C_{llr} value**
  - $C_{llr} > 0$
  - A system that always responds with a likelihood-ratio value of 1 irrespective of the input provides no useful information
    - the posterior odds will alway equal the prior odds
    - this system will have $C_{llr} = 1$

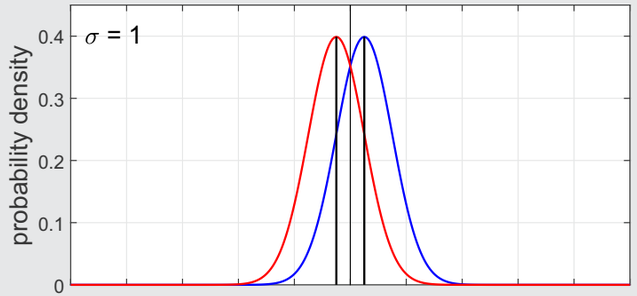
Validation metric

- The **better the performance** of the system, the **smaller the C_{llr} value**
  - $C_{llr} > 1$ can occur for an uncalibrated or miscalibrated system
    - this can be addressed by calibrating the system
  - A well-calibrated system will have $C_{llr} \leq 1$
    - but $C_{llr} \leq 1$ does not necessarily imply that the system is well calibrated
  - If $C_{llr} < 1$, the system is providing useful information

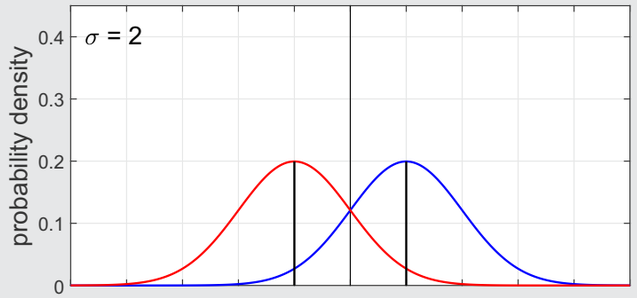
Validation metric

- Perfectly calibrated ln(LR) distributions
  - C_{llr} values

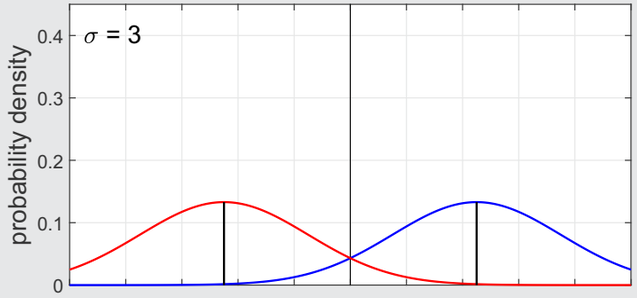
0.84



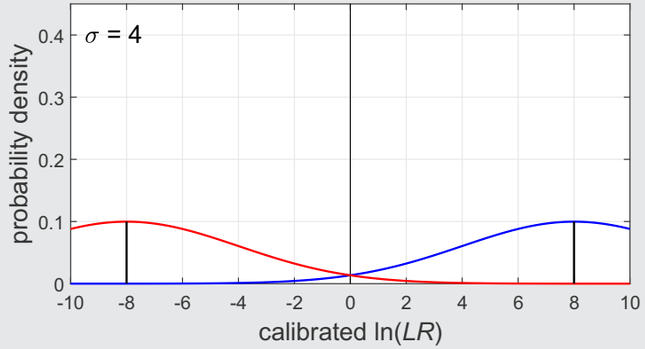
0.51



0.24



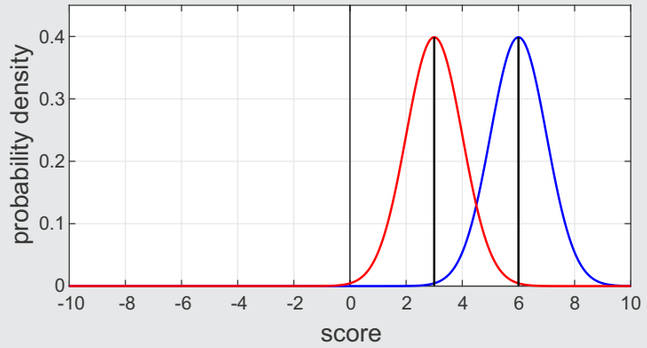
0.09



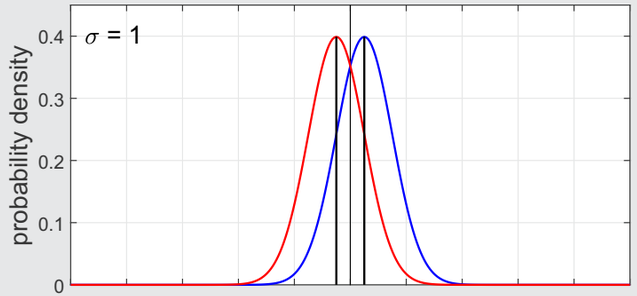
Validation metric

- Perfectly calibrated ln(LR) distributions
  - C_{llr} values
- Uncalibrated score distributions
  - C_{llr} value

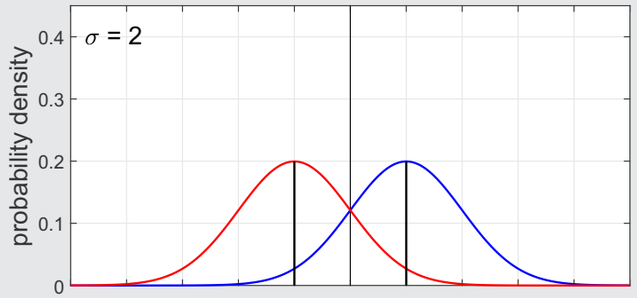
5.2



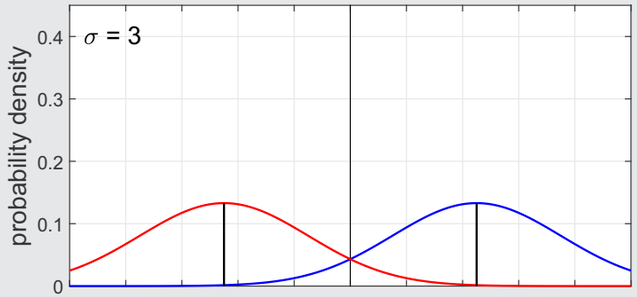
0.84



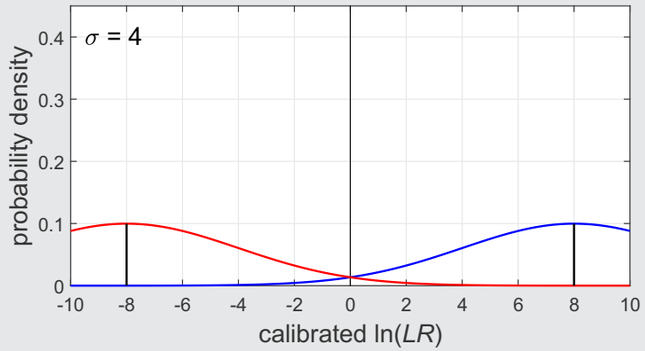
0.51



0.24



0.09



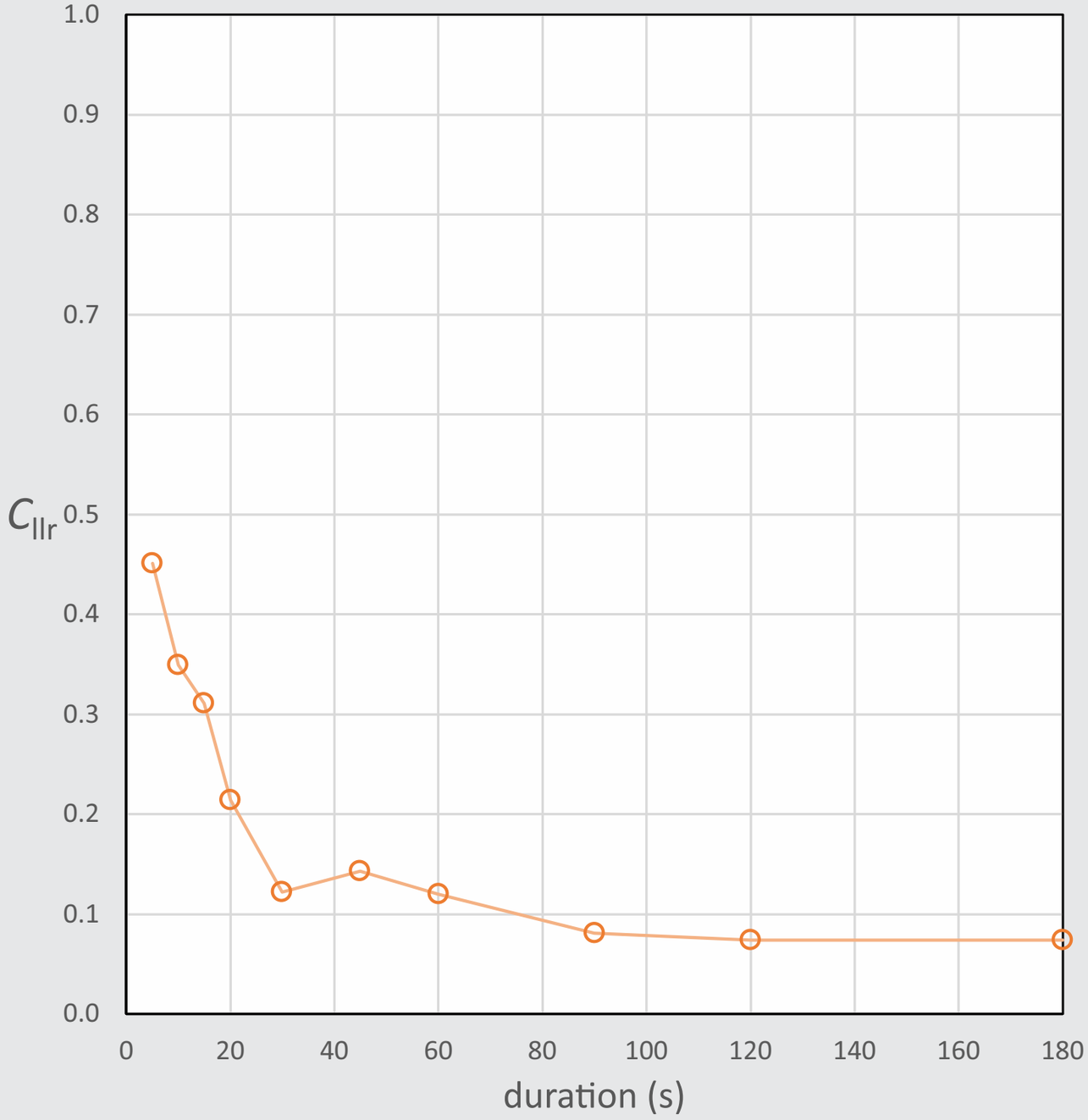
Validation metric

- Example C_{llr} values
  - different forensic-voice-comparison systems validated on the same case-relevant data

System name	System type	C_{llr}
Batvox 3.1	GMM-UBM	0.59
MSR GMM-UBM	GMM-UBM	0.58
MSR GMM i-vector	GMM i-vector	0.45
Batvox 4.1	GMM i-vector	0.37
Nuance 9.2	GMM i-vector	0.29
VOCALISE 2017B	GMM i-vector	0.27
VOCALISE 2019A	x-vector	0.25
E3FS3α	x-vector	0.21
Phonexia BETA4	x-vector	0.21

Validation metric

- Example C_{llr} values
  - a forensic-voice-comparison system validated with questioned-speaker recordings of different durations



Validation graphic

Tippett plot

Validation graphic

- For a system that outputs likelihood ratios, a graphical representation of performance should be based on **likelihood-ratio values**
  - given a **same-source** input pair
    - the **larger** the likelihood-ratio value the **better** the performance
  - given a **different-source** input pair
    - the **smaller** the likelihood-ratio value the **better** the performance

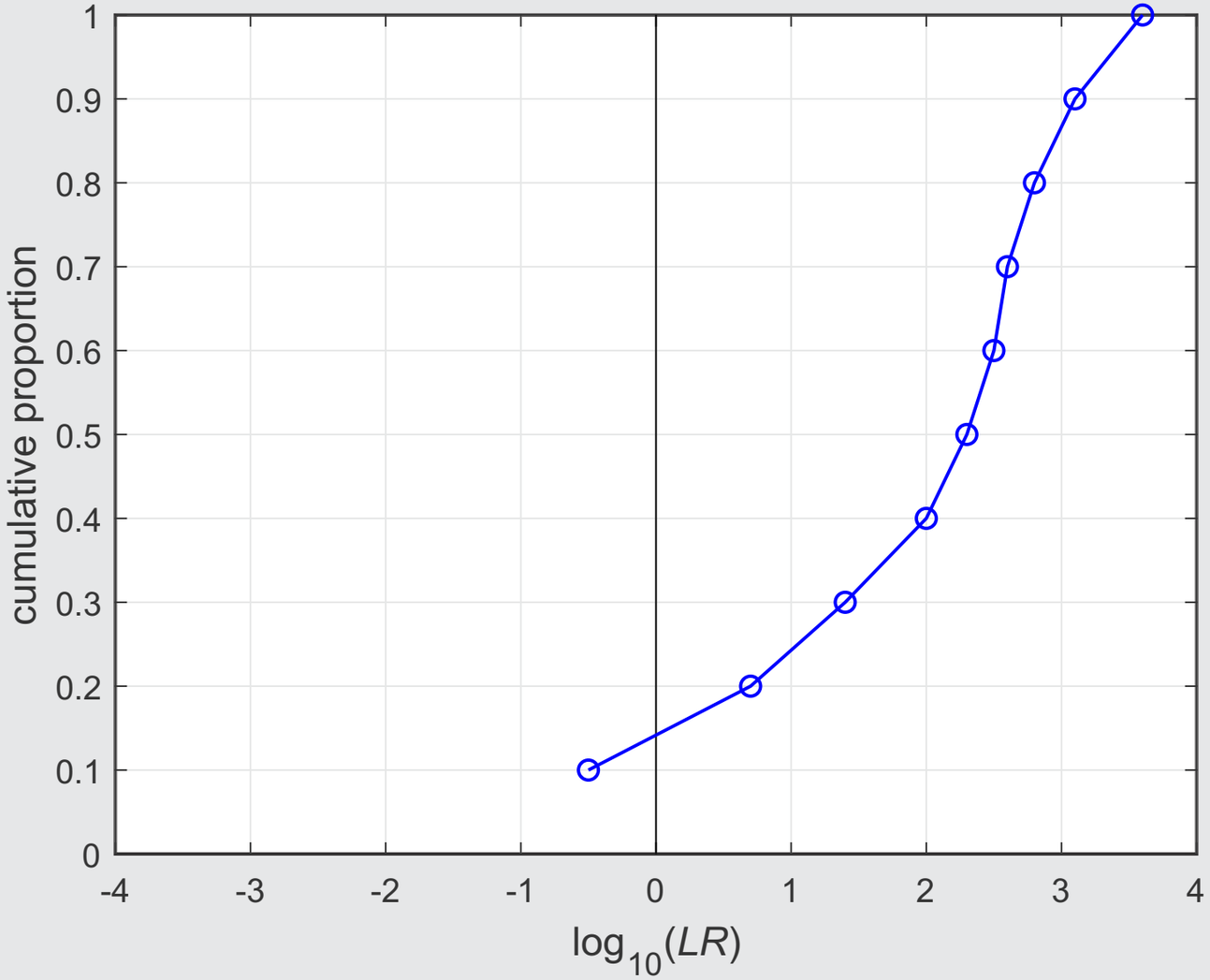
Validation graphic

- **Tippett plot**:
  - rank the log(LR) values resulting from same-source pairs from smallest to largest
  - plot the proportion of values that are ≤ each log(LR) value
    - value on y axis is the **proportion of same-source log likelihood ratio values** that are **smaller than** or equal to the value on the x axis

x	−0.5	0.7	1.4	2	2.3	2.5	2.6	2.8	3.1	3.6
y	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

Validation graphic

- **Tippett plot**:



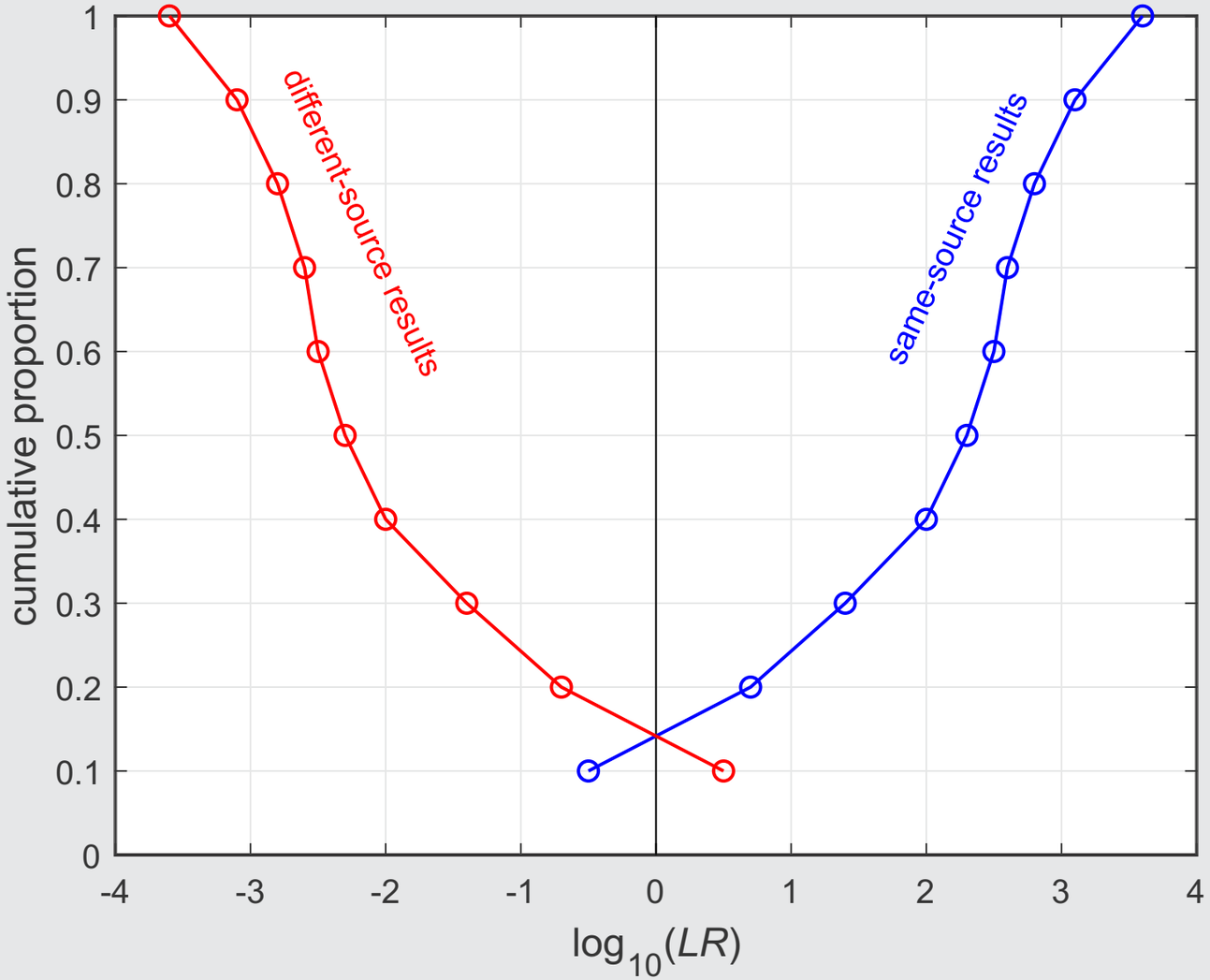
Validation graphic

- **Tippett plot**:
  - rank the log(LR) values resulting from different-source pairs from smallest to largest
  - plot the proportion of values that are ≥ each log(LR) value
    - value on y axis is the **proportion of different-source log likelihood ratio values** that are **larger than** or equal to the value on the x axis

x	−3.6	−3.1	−2.8	−2.6	−2.5	−2.3	−2	−1.4	−0.7	0.5
y	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1

Validation graphic

- **Tippett plot**:



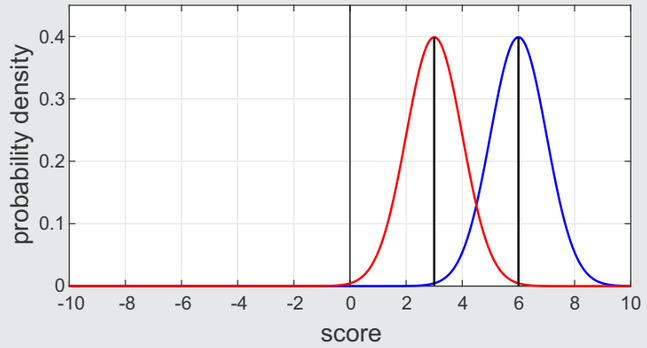
Validation graphic

- Tippett plots can be used to help:
    - decide whether the system is well calibrated or whether there is obvious bias in the validation results
    - decide whether the log-likelihood-ratio value calculated for the comparison of the actual questioned-source and known-source items in the case is supported by the validation results
        - values within the range of the validation results would be unambiguously supported
        - values just beyond the range of the validation results would be reasonable
        - values far beyond the range of the validation results would not be reasonable

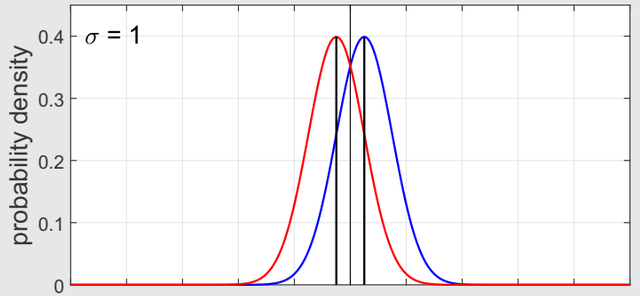
Validation graphic

- Perfectly calibrated ln(LR) distributions
  - C_{llr} values
- Uncalibrated score distributions
  - C_{llr} value

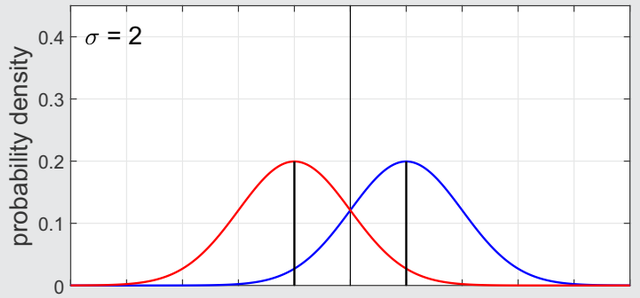
5.2



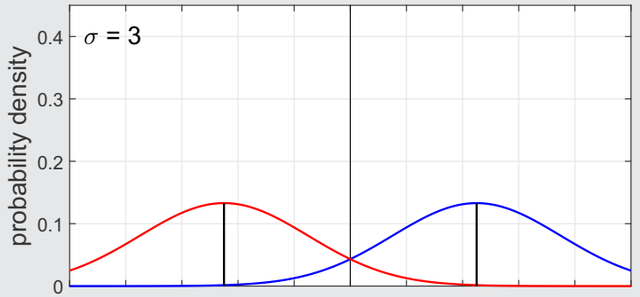
0.84



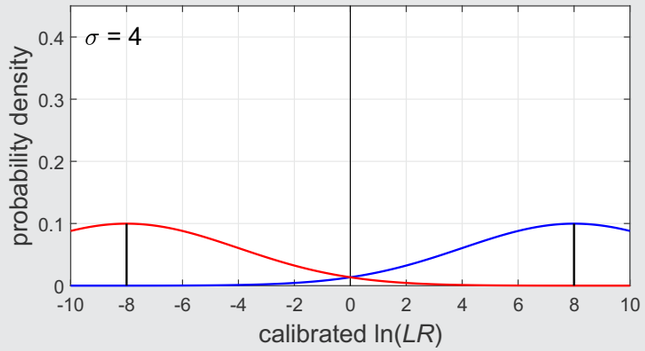
0.51



0.24



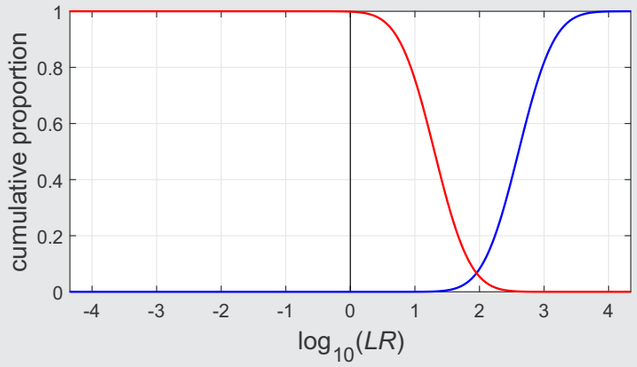
0.09



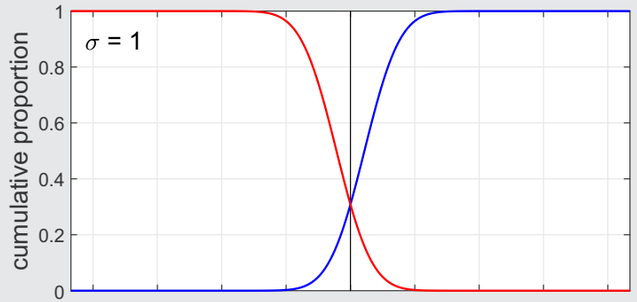
Validation graphic

- Tippett plots
  - C_{llr} values

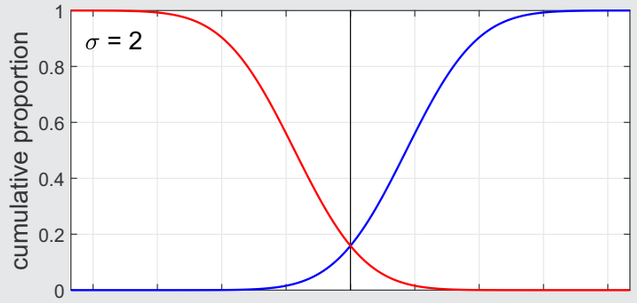
5.2



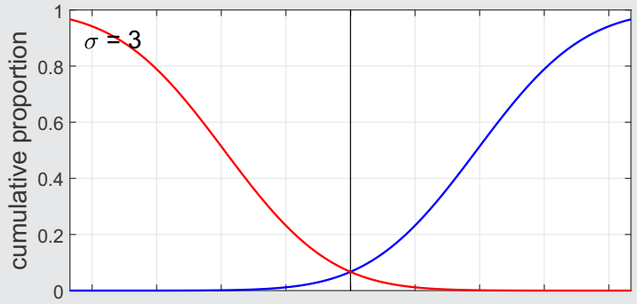
0.84



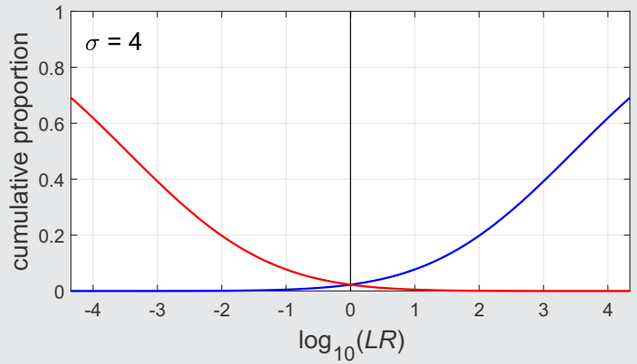
0.51



0.24



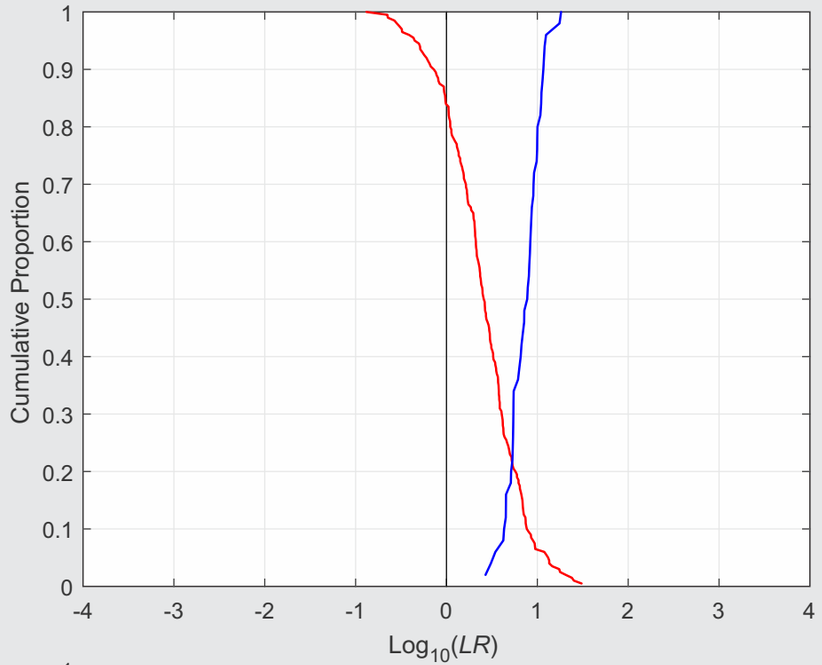
0.09



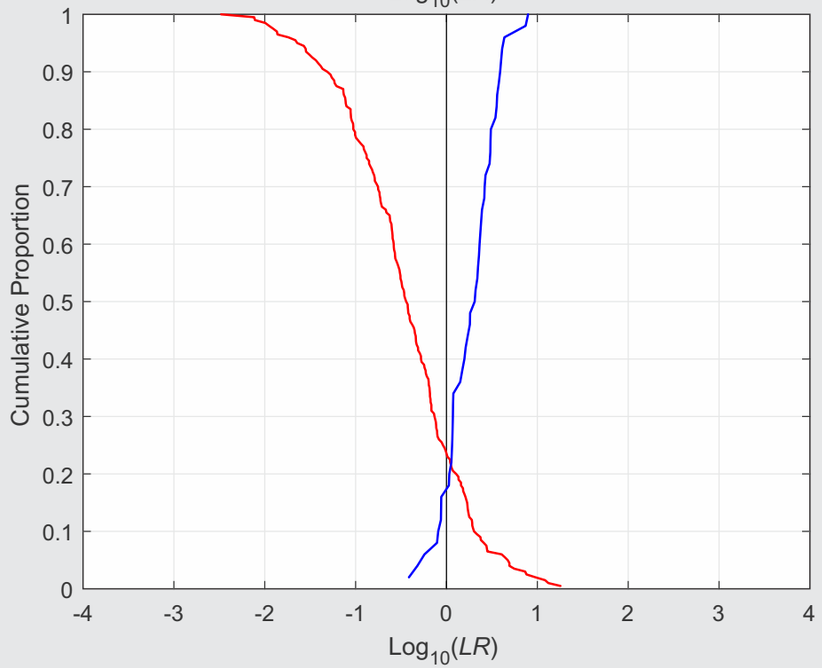
Validation graphic

- Example Tippett plots
  - C_{llr} values

1.07



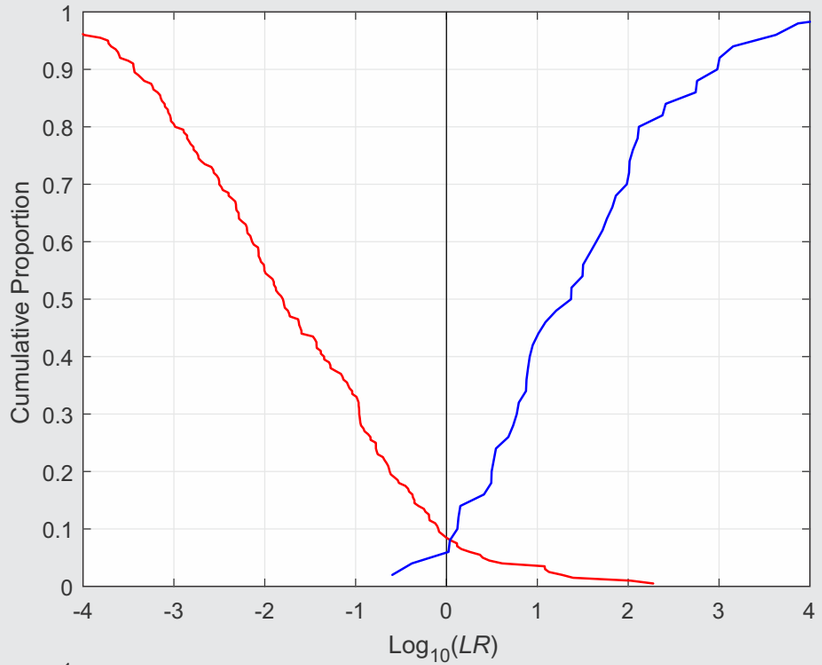
0.70



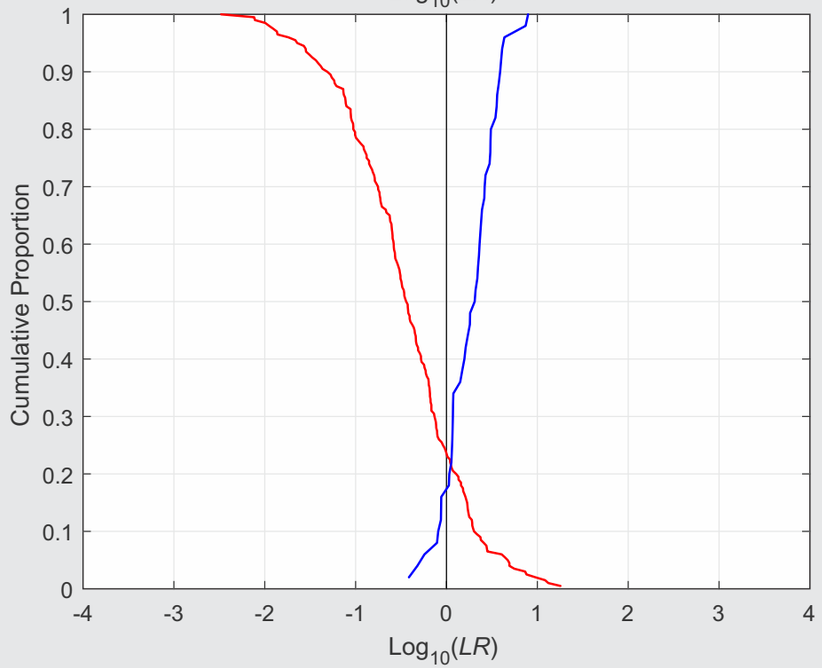
Validation graphic

- Example Tippett plots
  - C_{llr} values

0.31



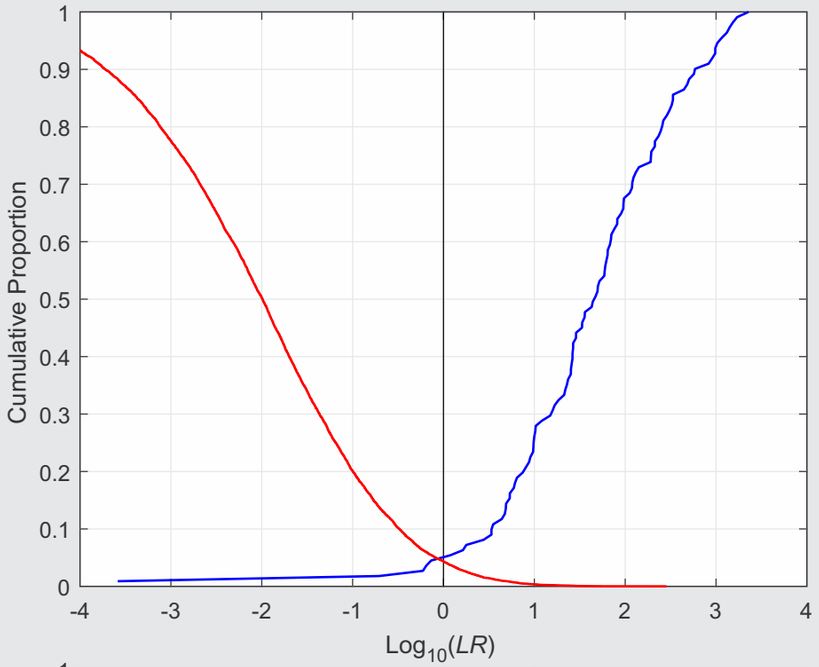
0.70



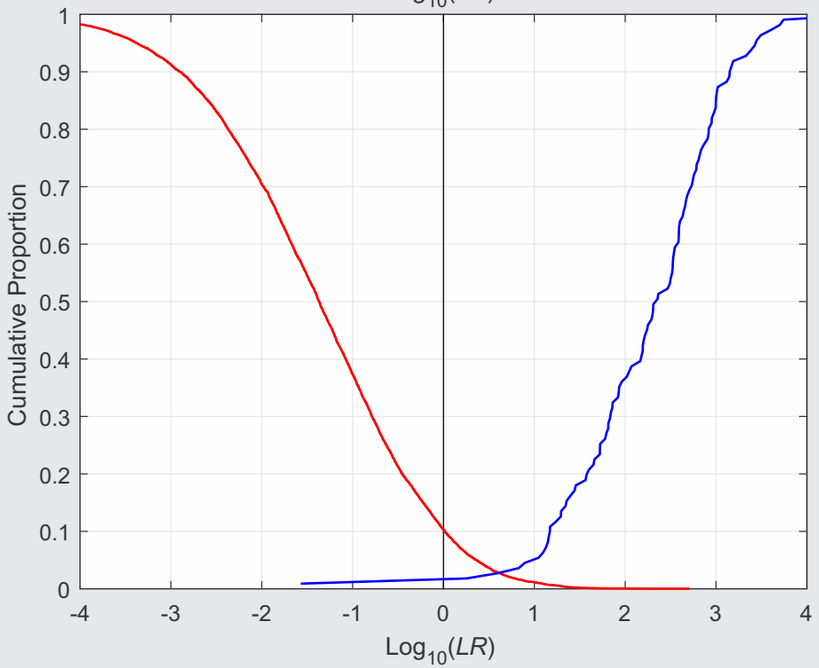
Validation graphic

- Example Tippett plots
  - different variants of a forensic-voice-comparison system validated on the same case-relevant data
  - C_{llr} values

0.21



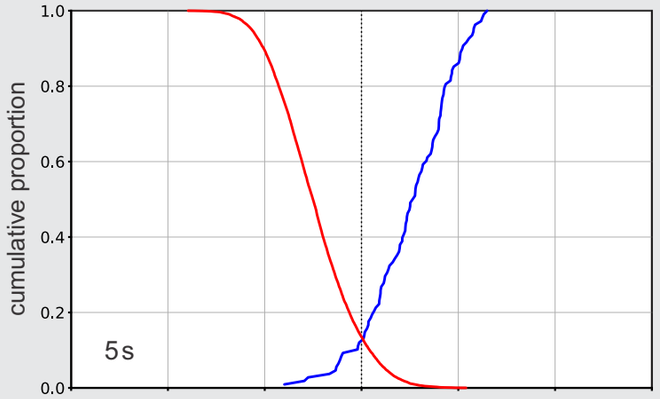
0.21



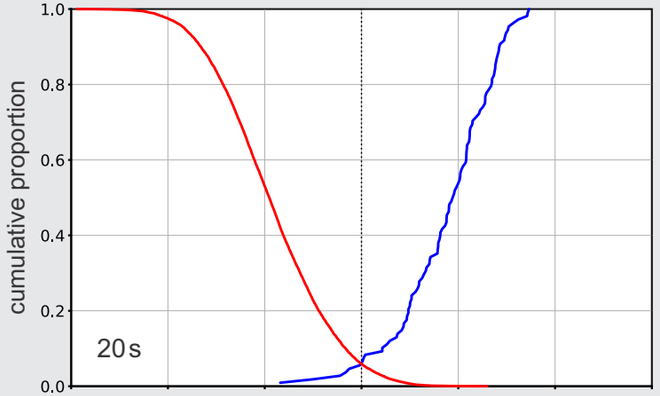
Validation graphic

- Example Tippett plots
  - a forensic-voice-comparison system validated with questioned-speaker recordings of different durations
  - C_{llr} values

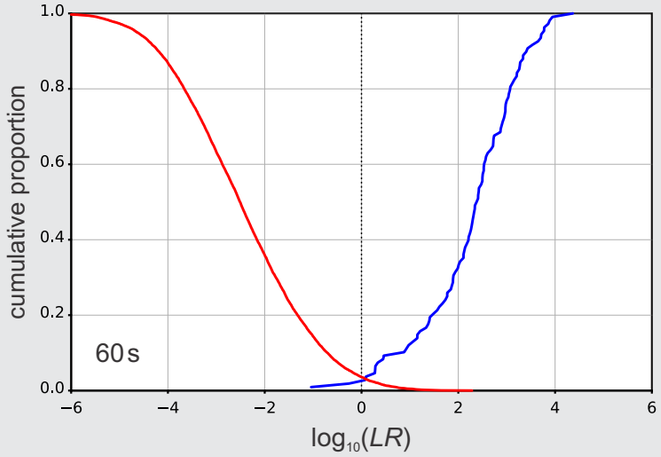
0.45



0.21



0.12



Bi-Gaussianized calibration

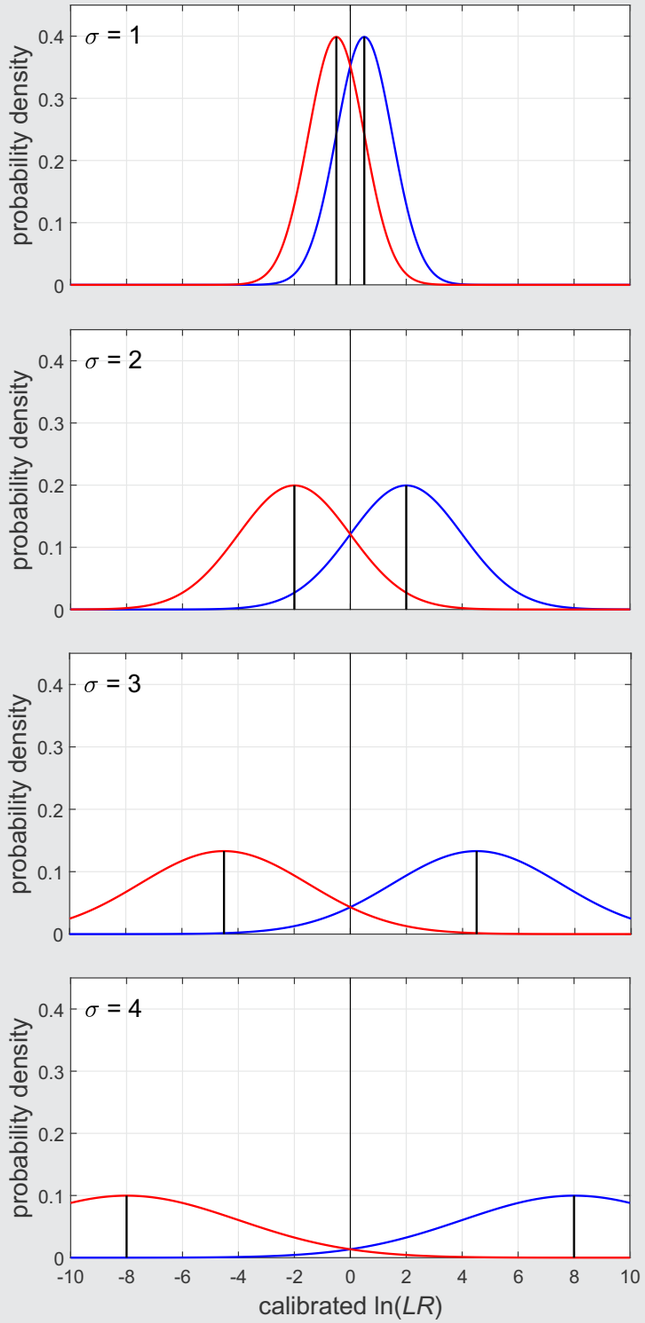
Well-calibrated likelihood ratios

- Perfectly calibrated ln(LR) distributions

- Both same-source and different-source distributions are Gaussian, and they have the same variance

$$\mu_{\mathrm{d}} = -\frac{\sigma^2}{2} \qquad \mu_{\mathrm{s}} = +\frac{\sigma^2}{2}$$

- Perfectly-calibrated bi-Gaussian systems



Bi-Gaussianized calibration

- Logistic-regression calibration applies a linear transformation in the log-likelihood-ratio space.

- Unless the distributions of the different-source and same-source uncalibrated log likelihood ratios are both Gaussian and have the same variance, the calibrated log likelihood ratios could be far from a perfectly calibrated bi-Gaussian system.

- Bi-Gaussianized calibration applies a non-linear (but still monotonic) transformation designed to bring the distributions closer to those of a perfectly-calibrated bi-Gaussian system.

Bi-Gaussianized calibration

1. Calculate uncalibrated likelihood ratios (scores) for training data and test data.

2. Calibrate the training-data output of Step 1 using logistic regression.

3. Calculate C_{llr} for the output of Step 2.

4. Determine the σ^2 of the perfectly-calibrated bi-Gaussian system with the C_{llr} calculated at Step 3.

5. Ignoring same-source and different-source labels, determine the mapping function from the empirical cumulative distribution of the training-data output of Step 1 to the cumulative distribution of the two-Gaussian mixture with the σ^2 determined at Step 4.

6. Apply the mapping function determined at Step 5 to the test-data output of Step 1.

Bi-Gaussianized calibration

1. Calculate uncalibrated likelihood ratios (scores) for training data and test data.

2. Calibrate the training-data output of Step 1 using logistic regression.

3. Calculate C_{llr} for the output of Step 2.

4. Determine the σ^2 of the perfectly-calibrated bi-Gaussian system with the C_{llr} calculated at Step 3.

5. Ignoring same-source and different-source labels, determine the mapping function from the empirical cumulative distribution of the training-data output of Step 1 to the cumulative distribution of the two-Gaussian mixture with the σ^2 determined at Step 4.

6. Apply the mapping function determined at Step 5 to the test-data output of Step 1.

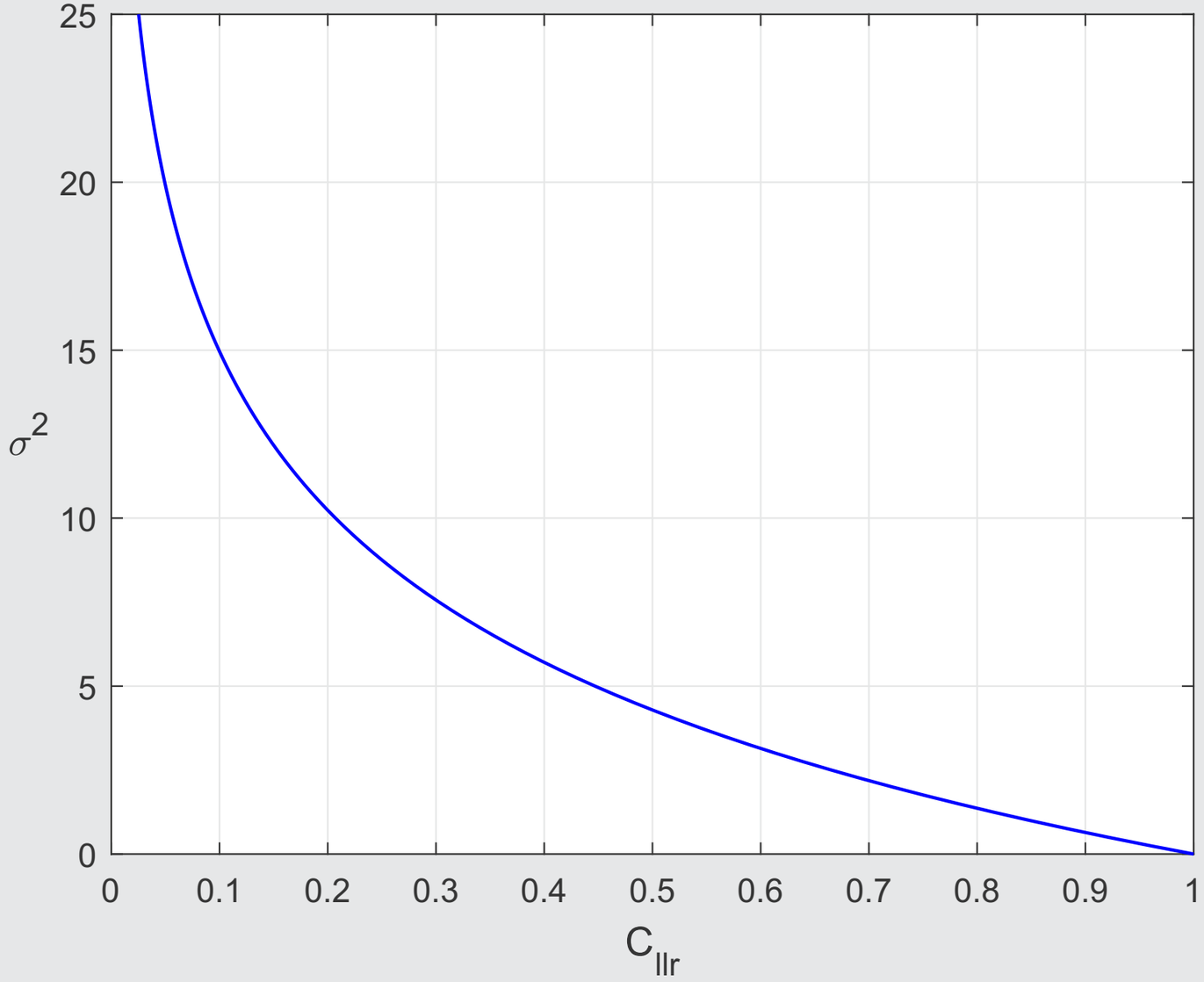
Relationship between C_{llr} and σ^2

- for a perfectly-calibrated bi-Gaussian system

$$\sigma^2 = -\frac{\ln\left(\frac{\ln(C_{llr})}{b} + 1\right)}{c}$$

$b = 17.7$

$c = 9.33 \times 10^{-3}$



Bi-Gaussianized calibration

1. Calculate uncalibrated likelihood ratios (scores) for training data and test data.

2. Calibrate the training-data output of Step 1 using logistic regression.

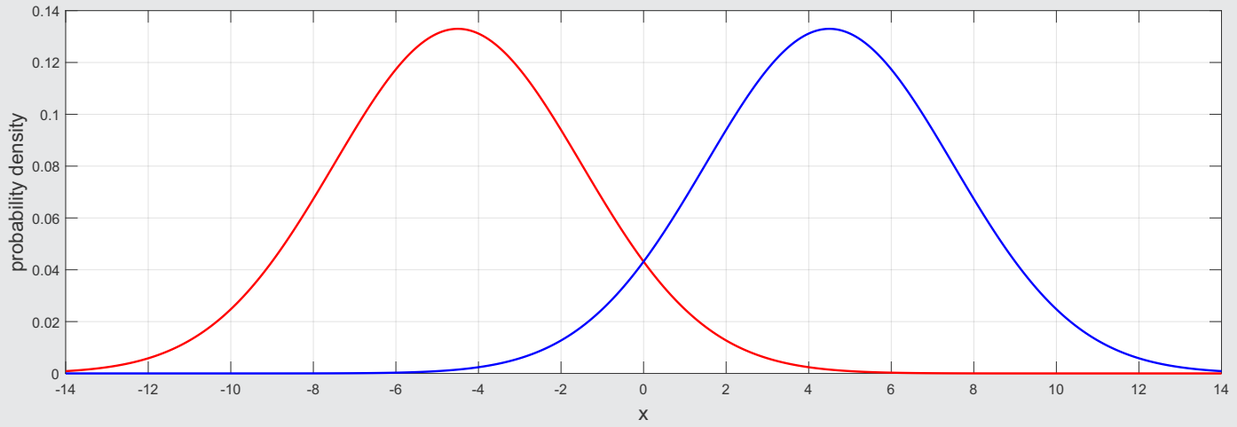
3. Calculate C_{llr} for the output of Step 2.

4. Determine the σ^2 of the perfectly-calibrated bi-Gaussian system with the C_{llr} calculated at Step 3.

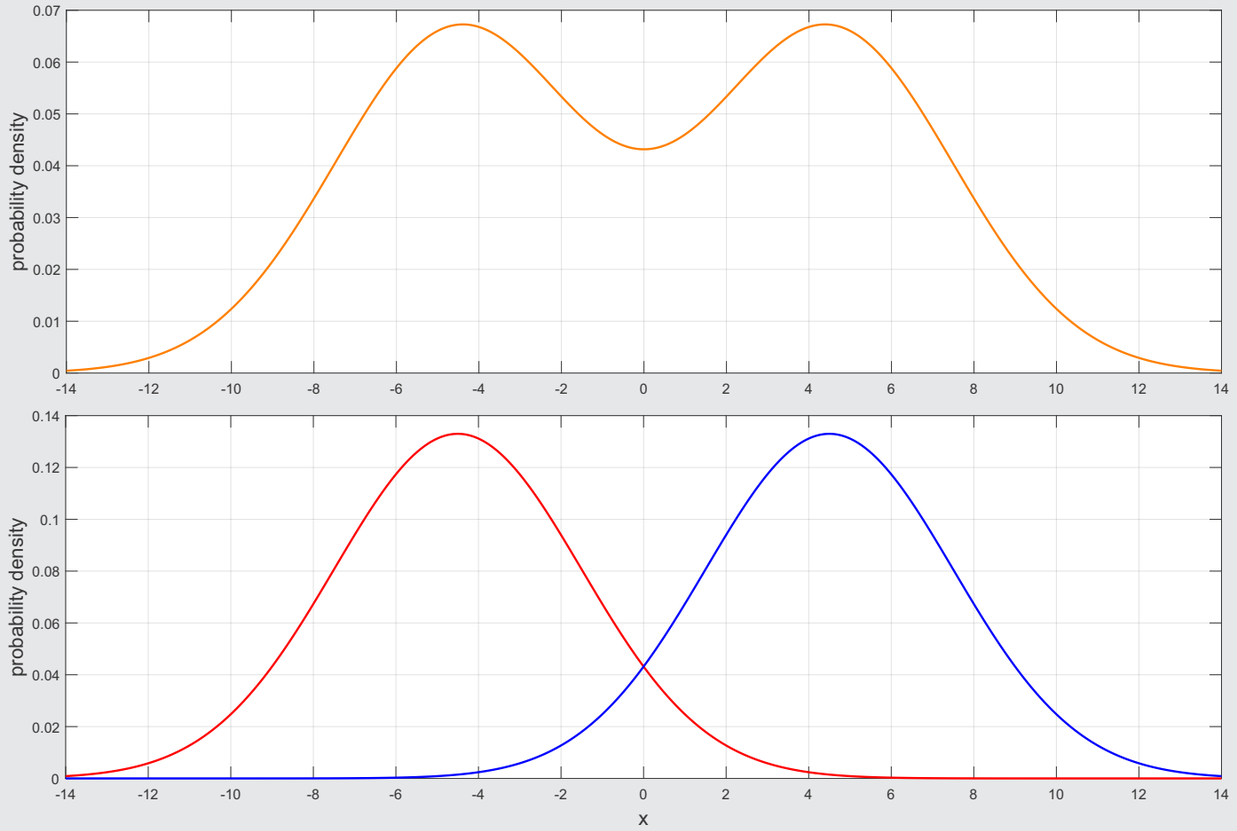
5. Ignoring same-source and different-source labels, determine the mapping function from the empirical cumulative distribution of the training-data output of Step 1 to the cumulative distribution of the two-Gaussian mixture with the σ^2 determined at Step 4.

6. Apply the mapping function determined at Step 5 to the test-data output of Step 1.

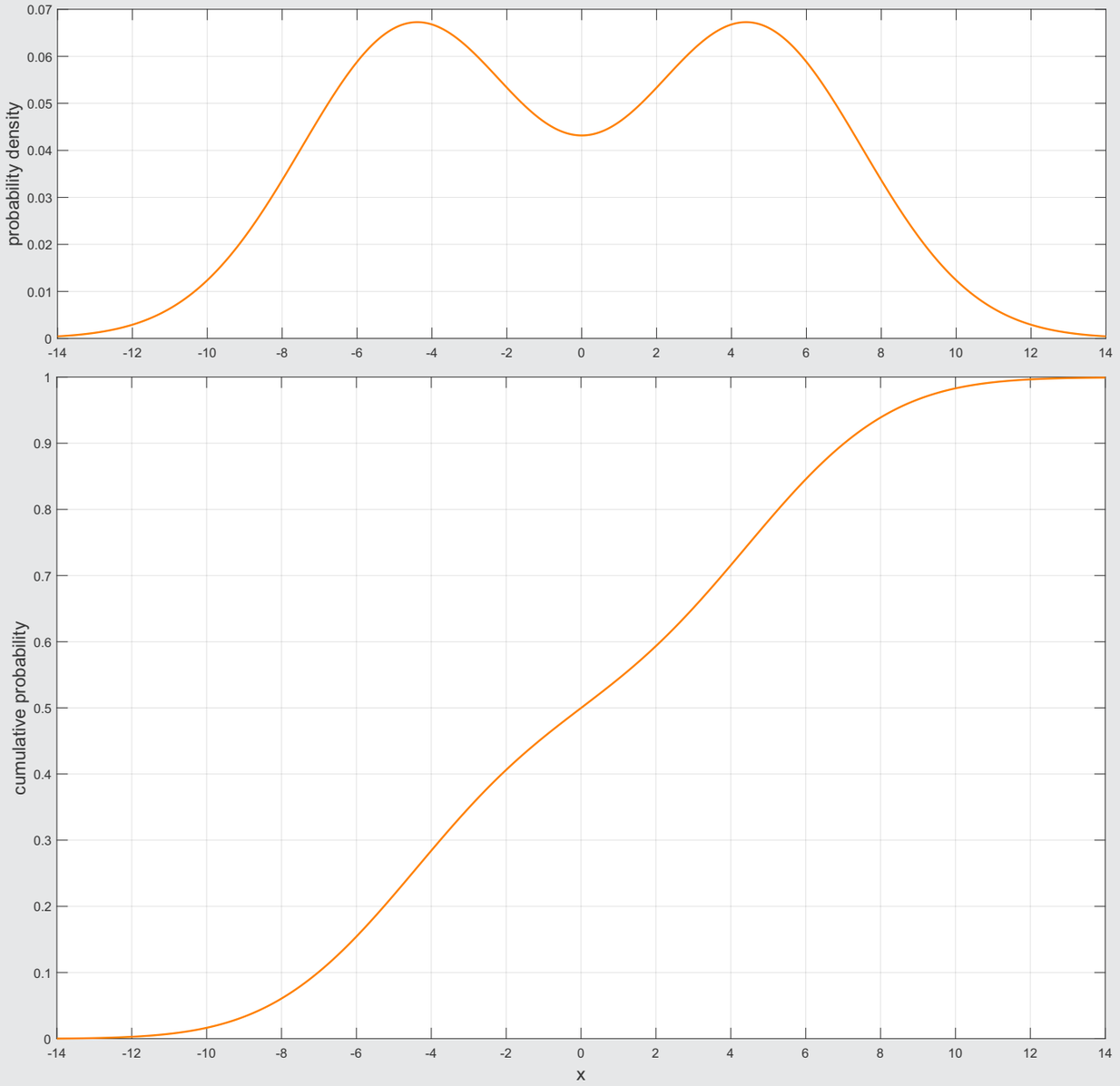
Cumulative distribution of Gaussian mixture



Cumulative distribution of Gaussian mixture

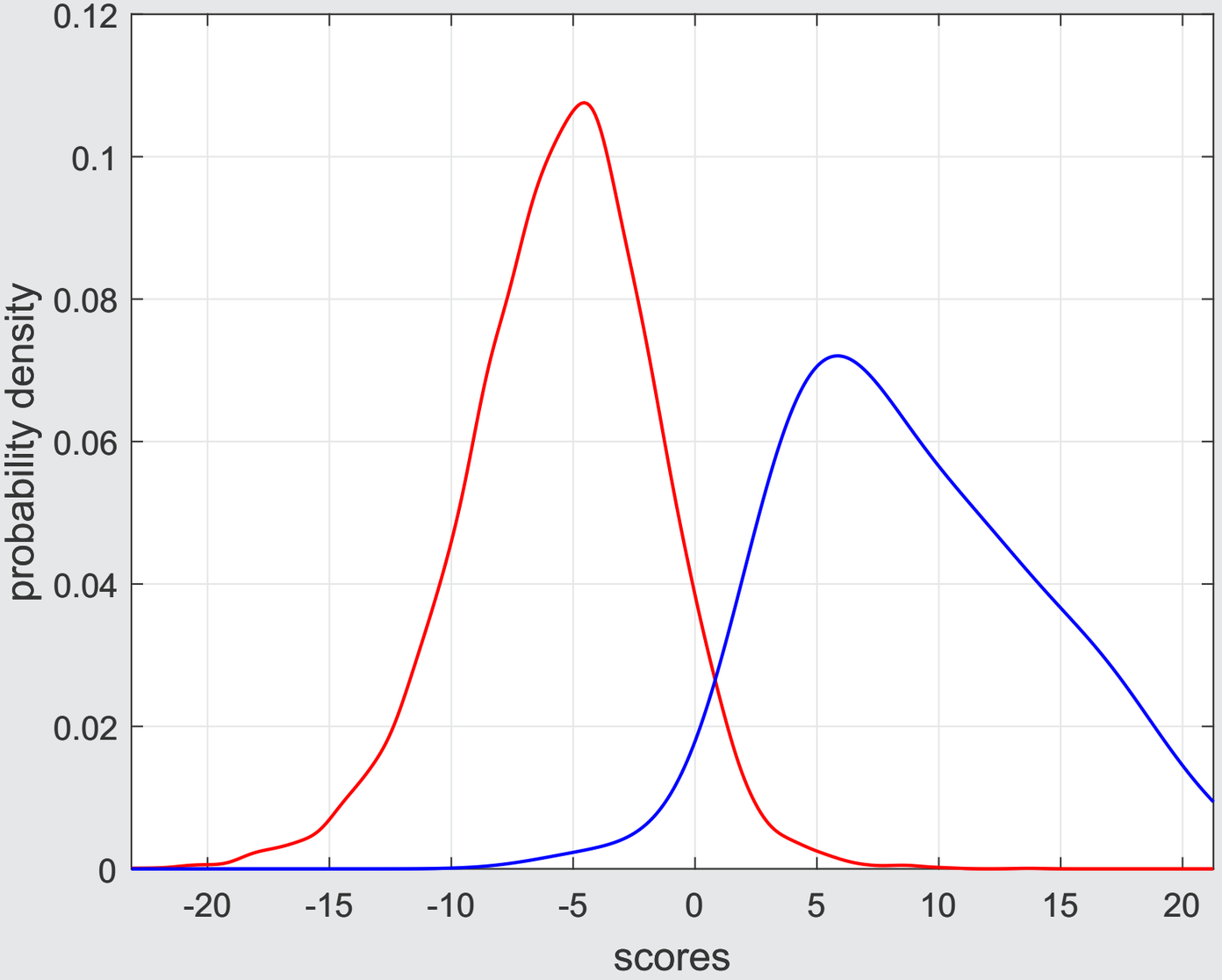


Cumulative distribution of Gaussian mixture



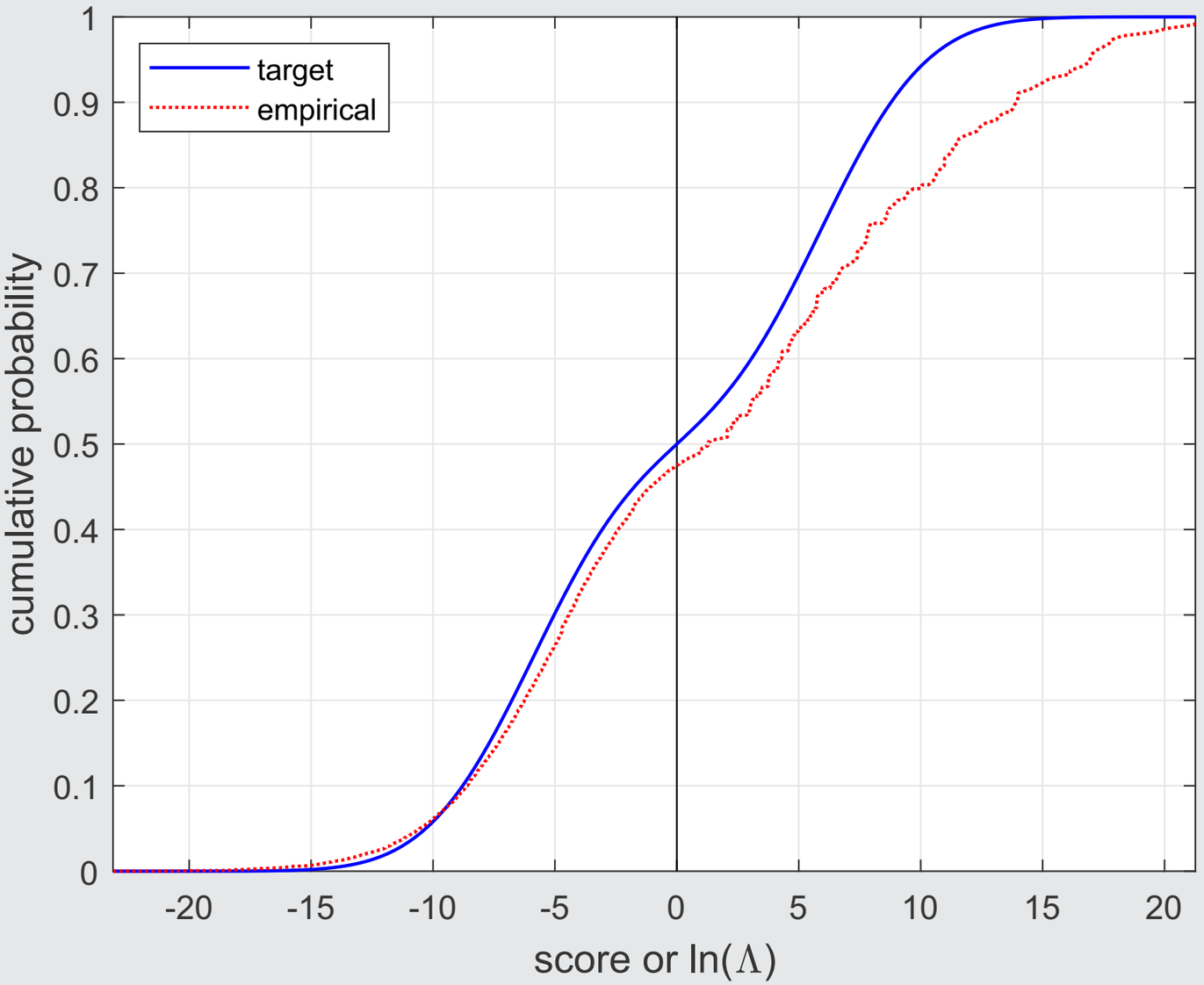
Bi-Gaussianized calibration

- forensic-voice-comparison data
- $C_{\text{llr}} = 0.172$
- target $\sigma = 3.44$



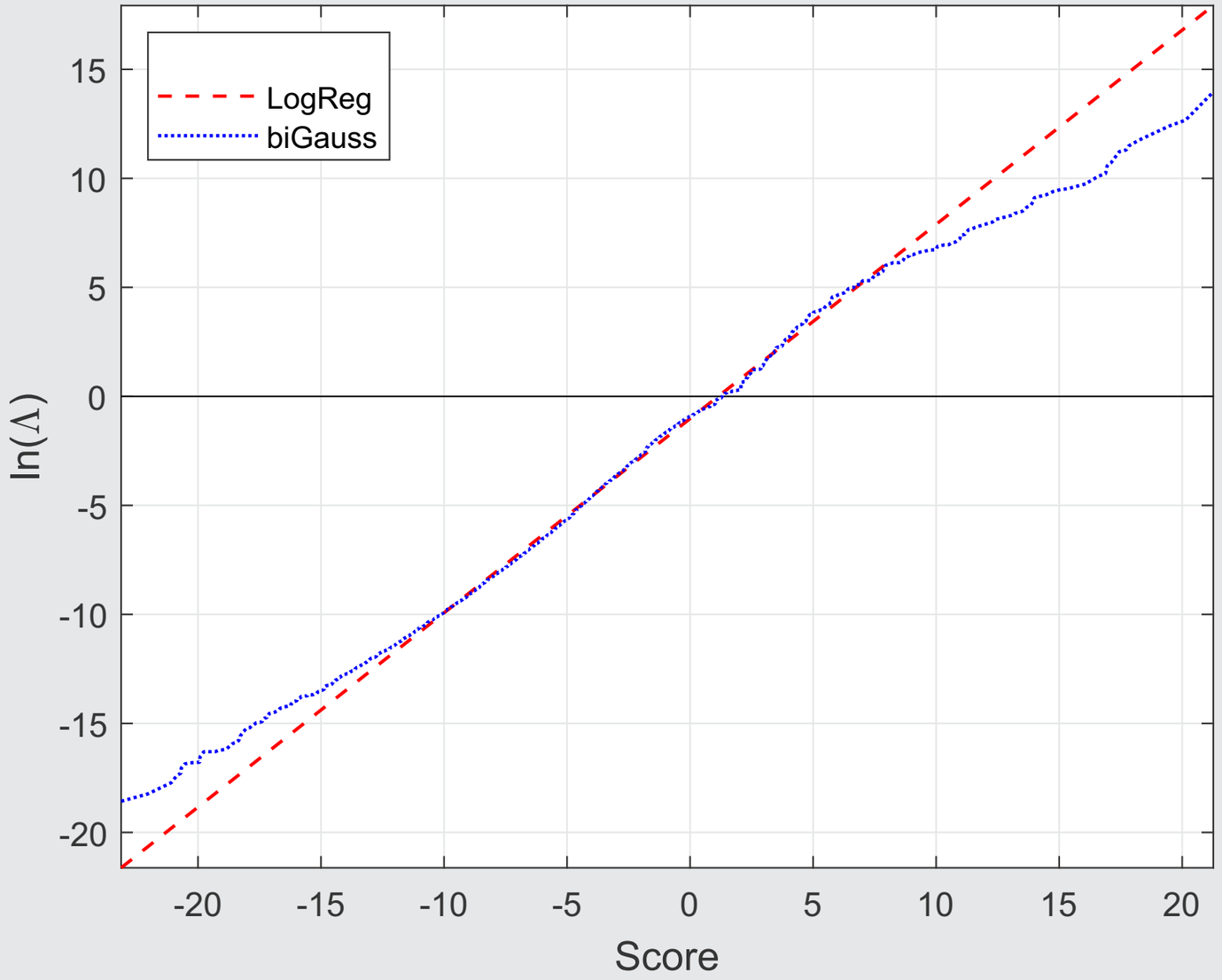
Bi-Gaussianized calibration

- forensic-voice-comparison data
- $C_{\text{llr}} = 0.172$
- target $\sigma = 3.44$



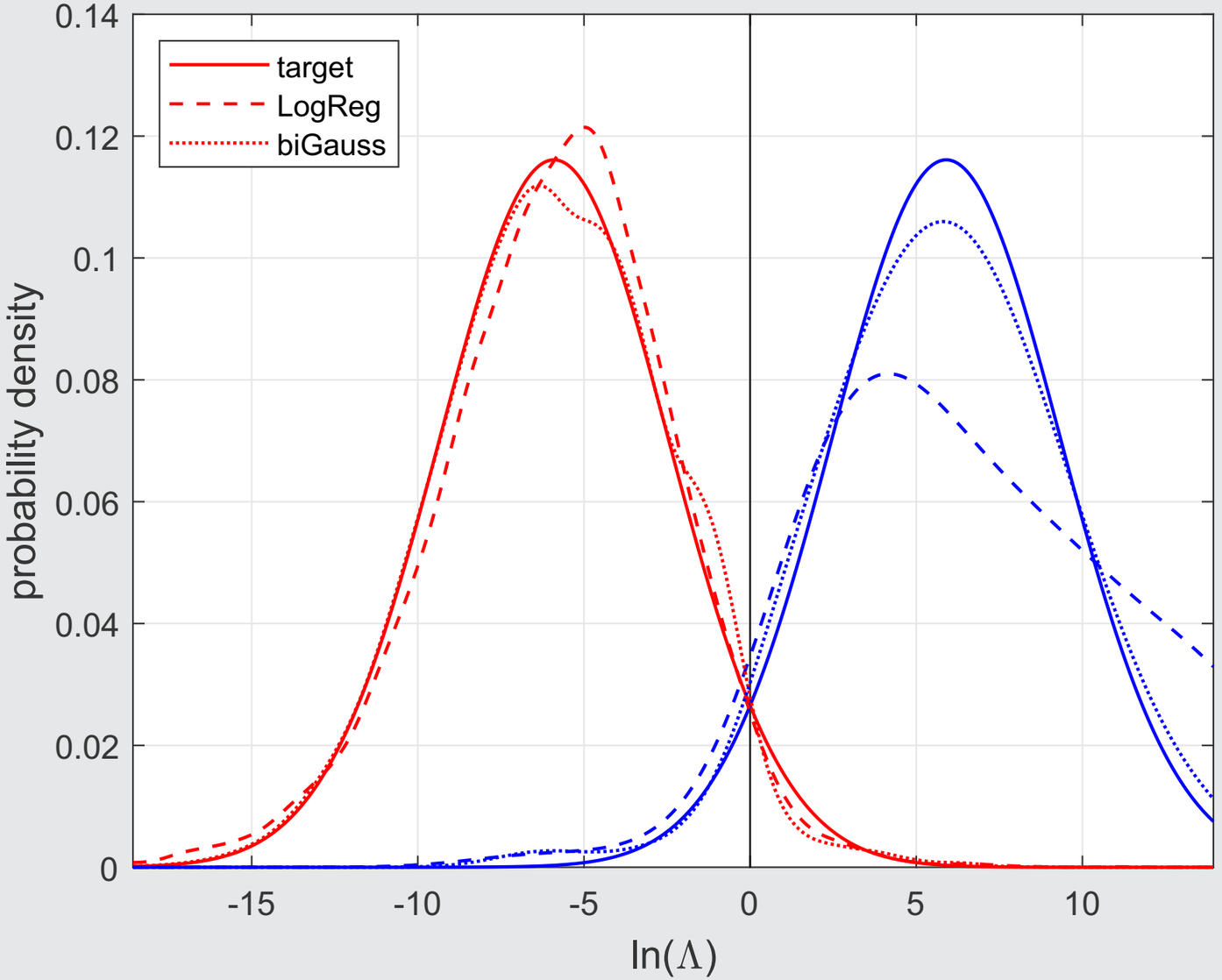
Bi-Gaussianized calibration

- forensic-voice-comparison data
- $C_{\text{llr}} = 0.172$
- target $\sigma = 3.44$



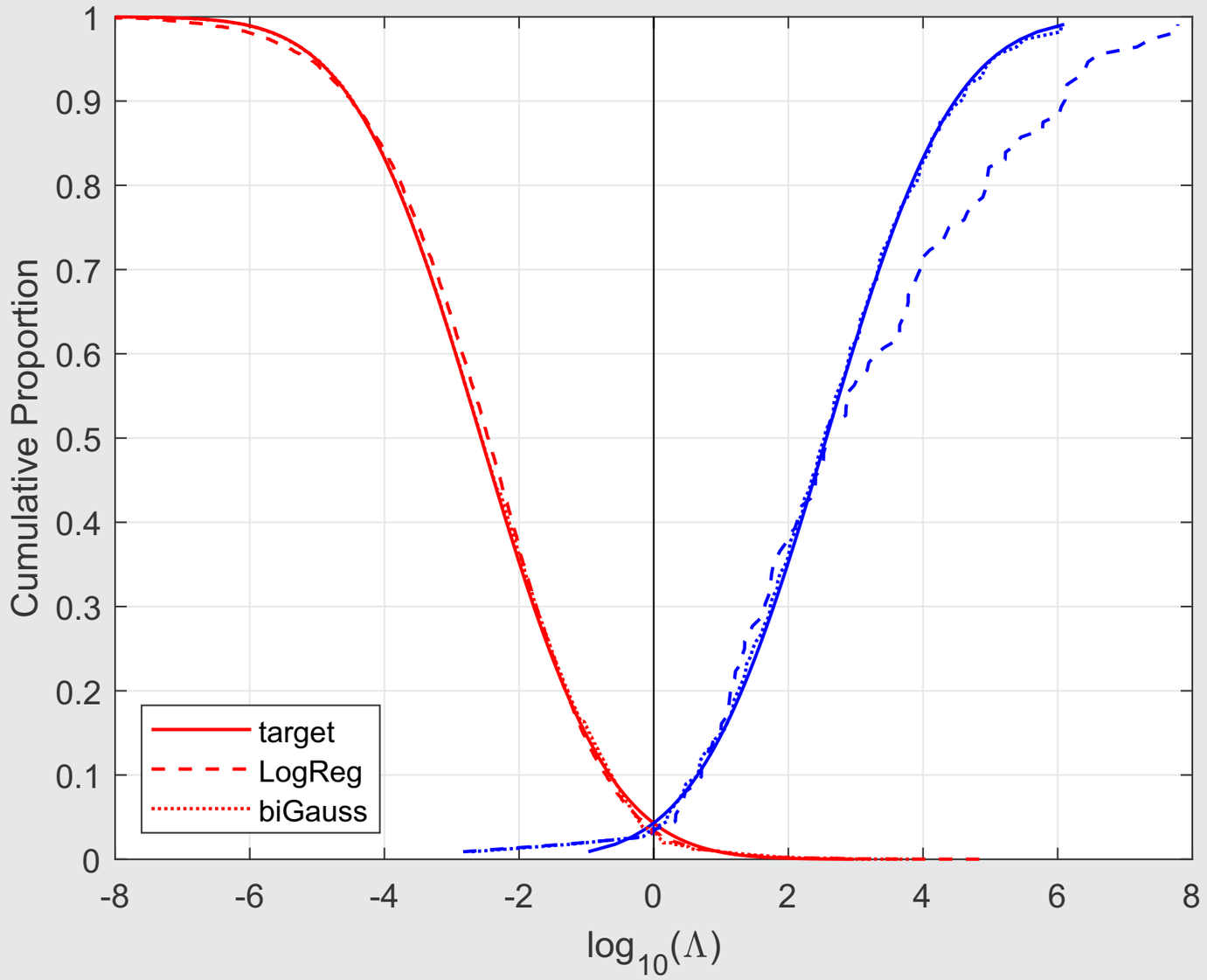
Bi-Gaussianized calibration

- forensic-voice-comparison data
- $C_{llr} = 0.172$
- target $\sigma = 3.44$



Bi-Gaussianized calibration

- forensic-voice-comparison data
- $C_{\text{llr}} = 0.172$
- target $\sigma = 3.44$



Thank You