Concepts of likelihood-ratio calculation

Calibration and validation of likelihood-ratio systems

Geoffrey Stewart Morrison

Forensic Data Science Laboratory
Aston Institute for Forensic Linguistics

Aston University
BIRMINGHAM UK

Workshop material

- Slides

https://forensic-data-science.net/workshops/LR_calc.html

Contents – Part I

- Concepts of likelihood-ratio calculation
  - Discrete data
  - Probability density (continuous data)
  - Specific-source likelihood ratios
  - Common-source likelihood ratios
  - Similarity-score-based likelihood ratios

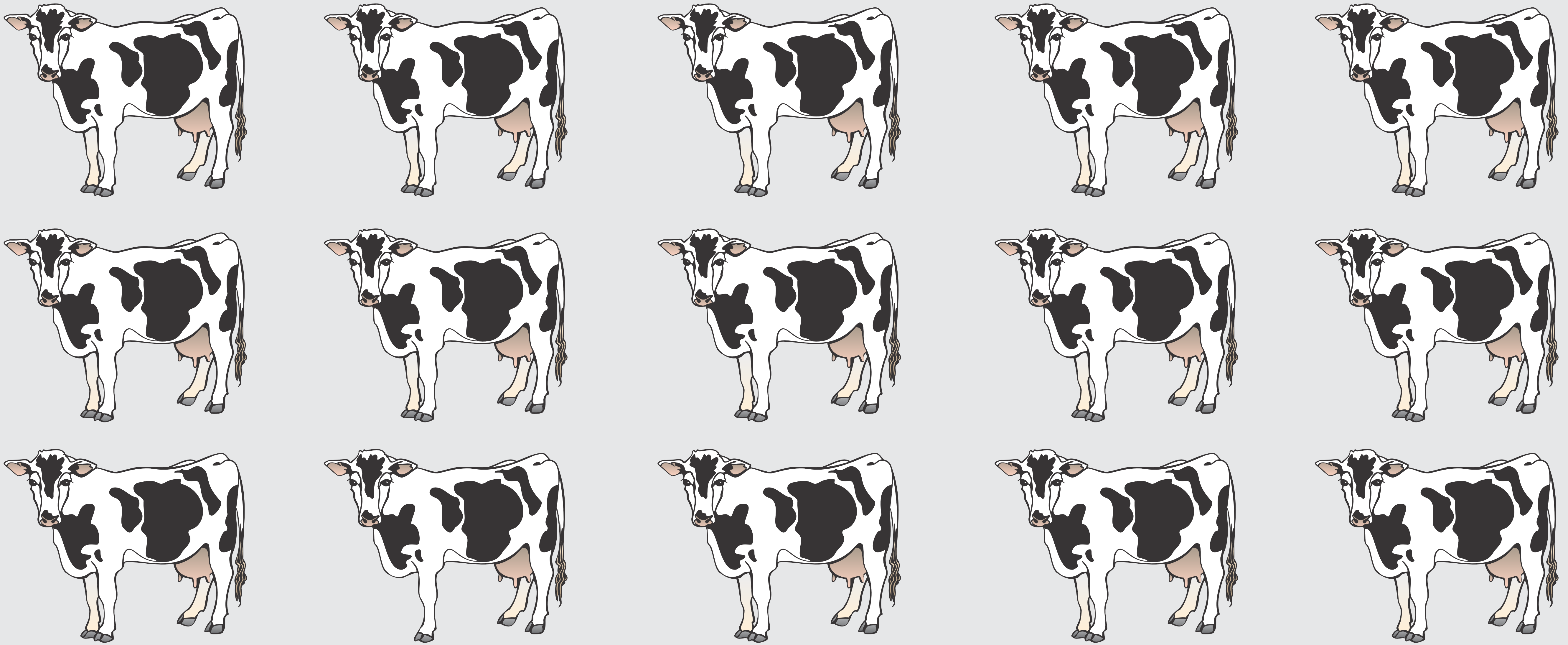
Contents – Part II

- 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)
- Consensus on Validation
  - Key points

Part I
Concepts of Likelihood-Ratio Calculation

Discrete Data

Remember the cows?



Prize geranium

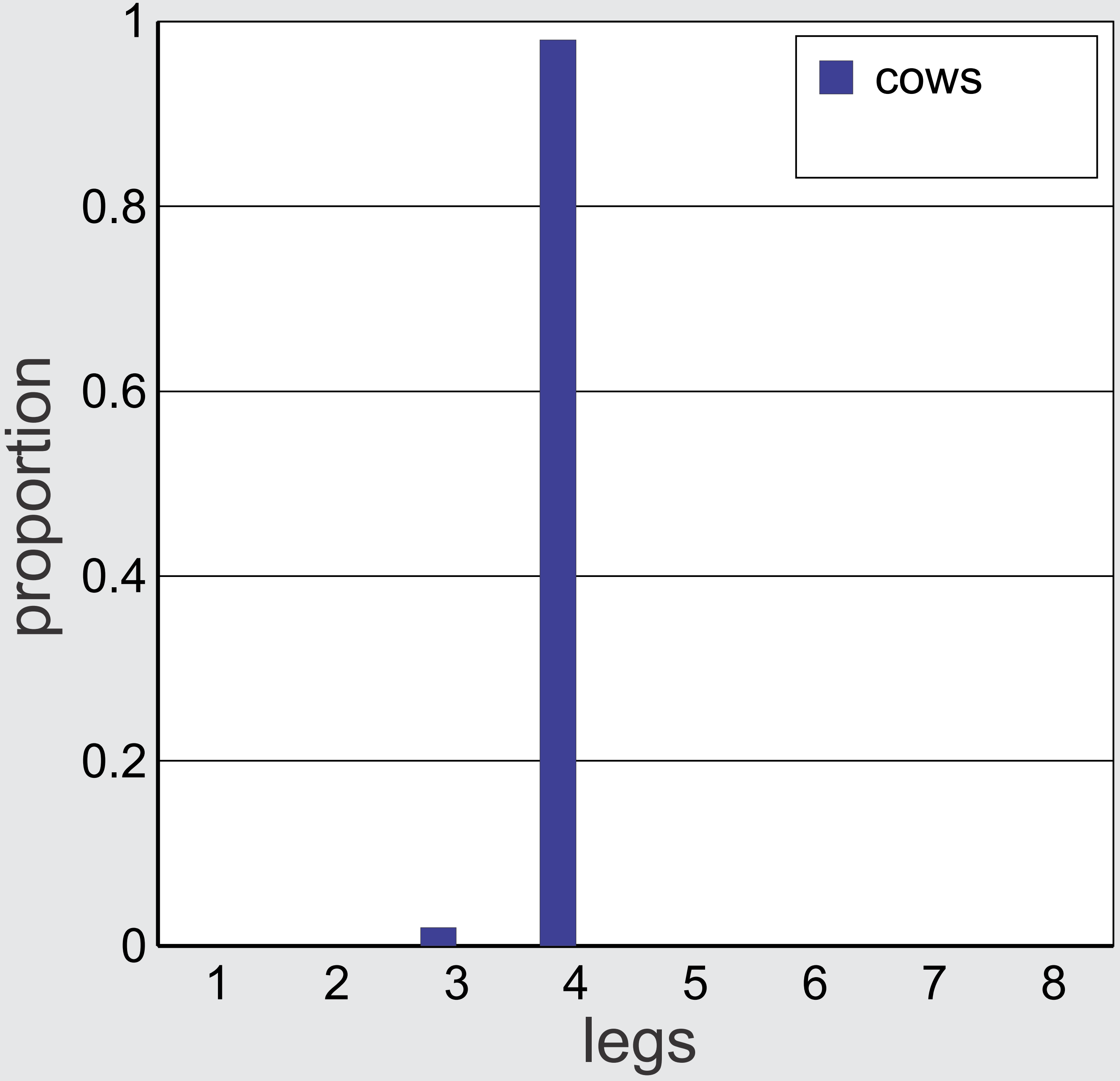
Ready to calculate a likelihood ratio?

$$\frac{\mathrm{p}(\,x\ \text{legs} \mid \text{cow}\,)}{\mathrm{p}(\,x\ \text{legs} \mid \text{not cow}\,)}$$

$$\frac{\mathrm{p}(\,E \mid H_1\,)}{\mathrm{p}(\,E \mid H_2\,)}$$

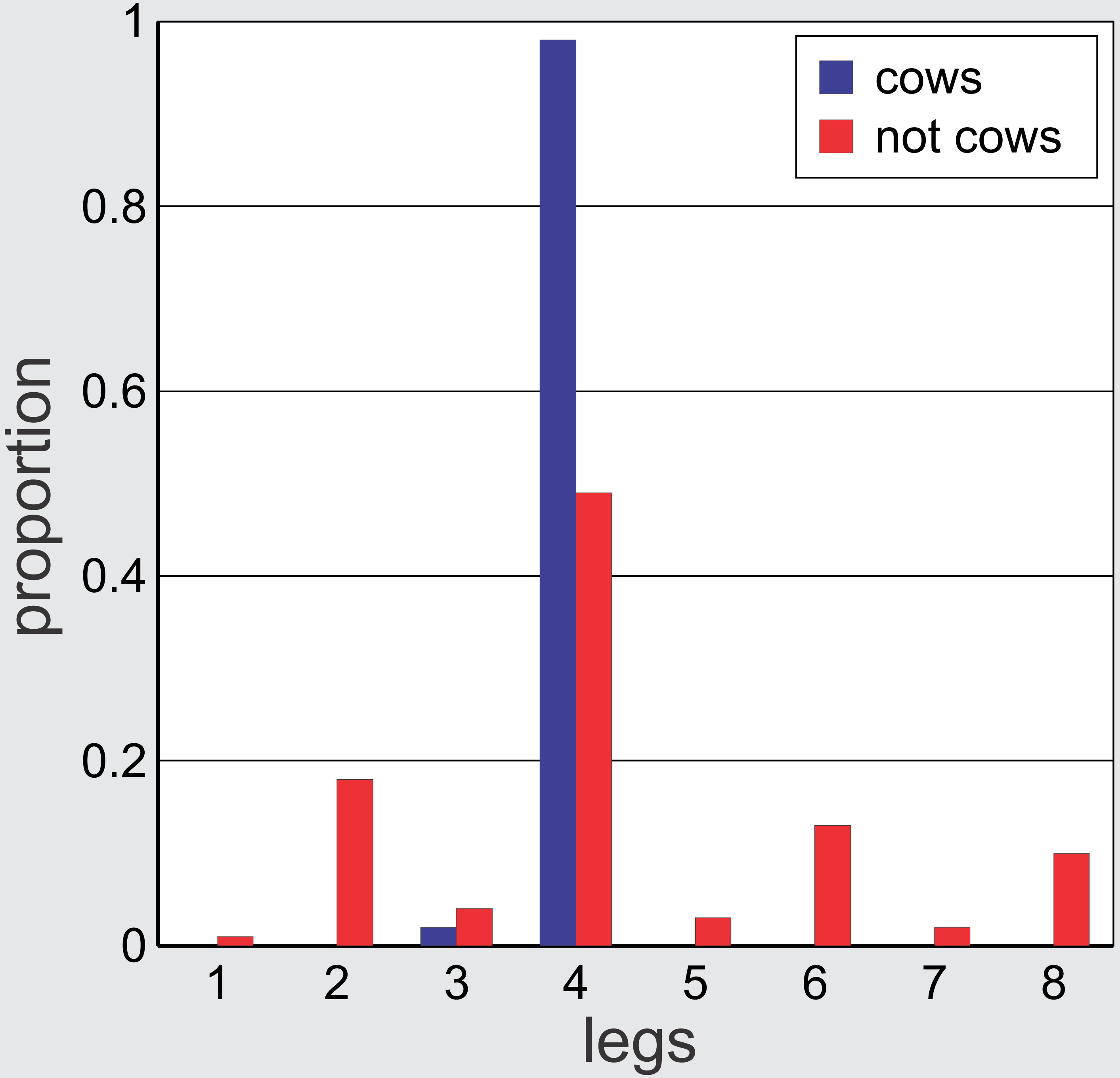
Discrete data

- Bar graph



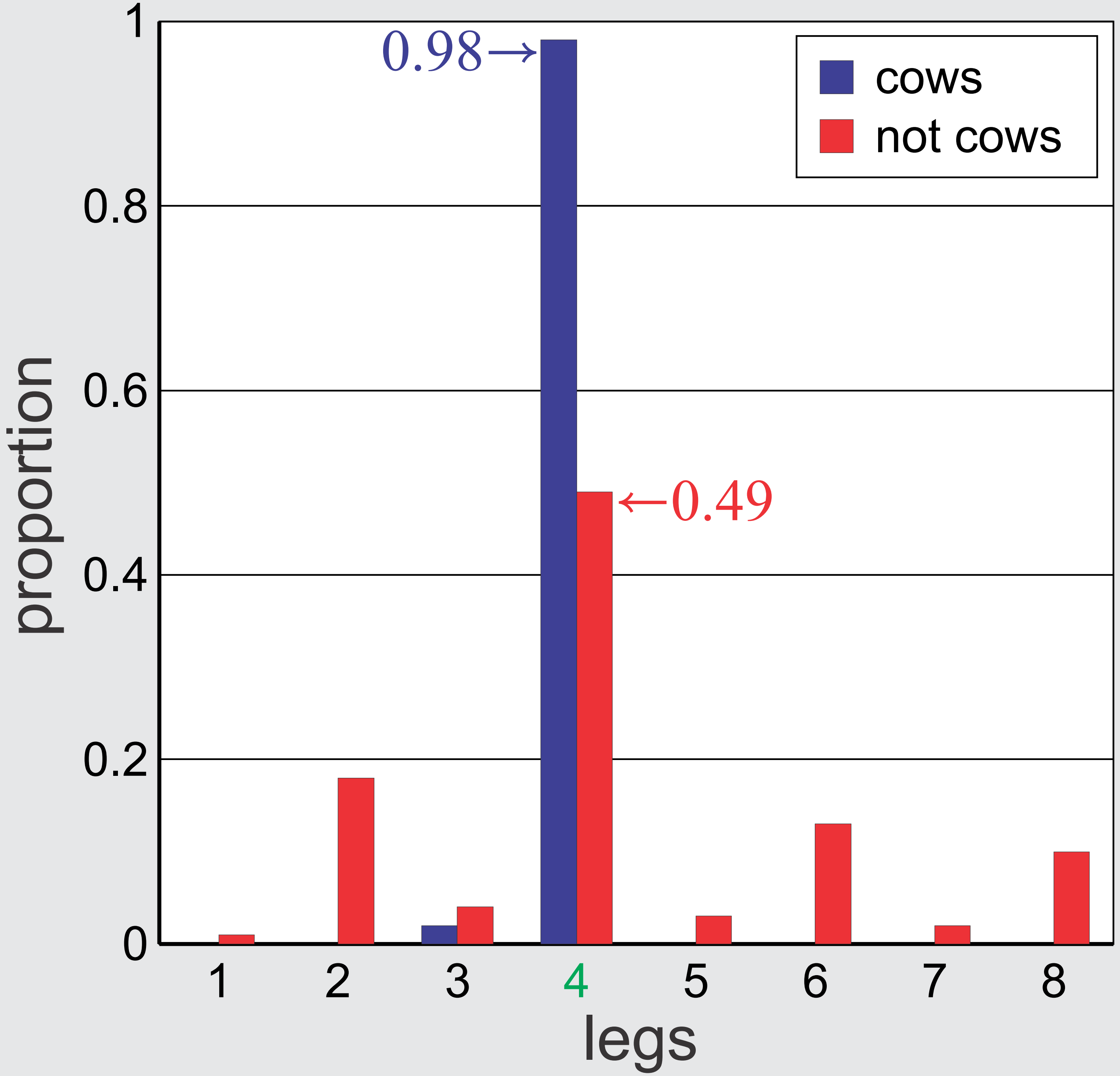
Discrete data

- Bar graph



Discrete data

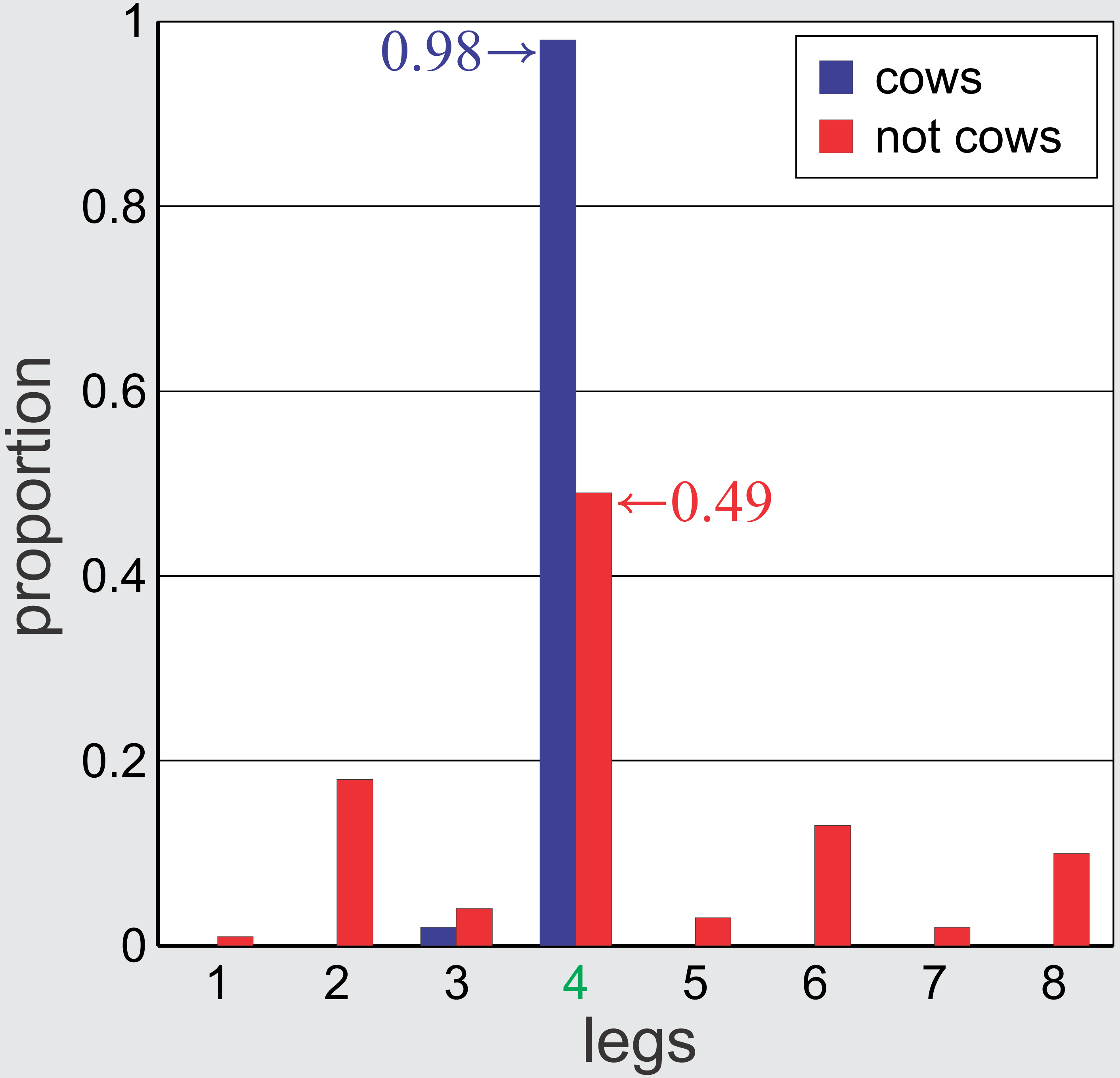
- Bar graph



$$\frac{p(\text{ 4 legs } | \text{ cow })}{p(\text{ 4 legs } | \text{ not a cow })}$$

Discrete data

- Bar graph



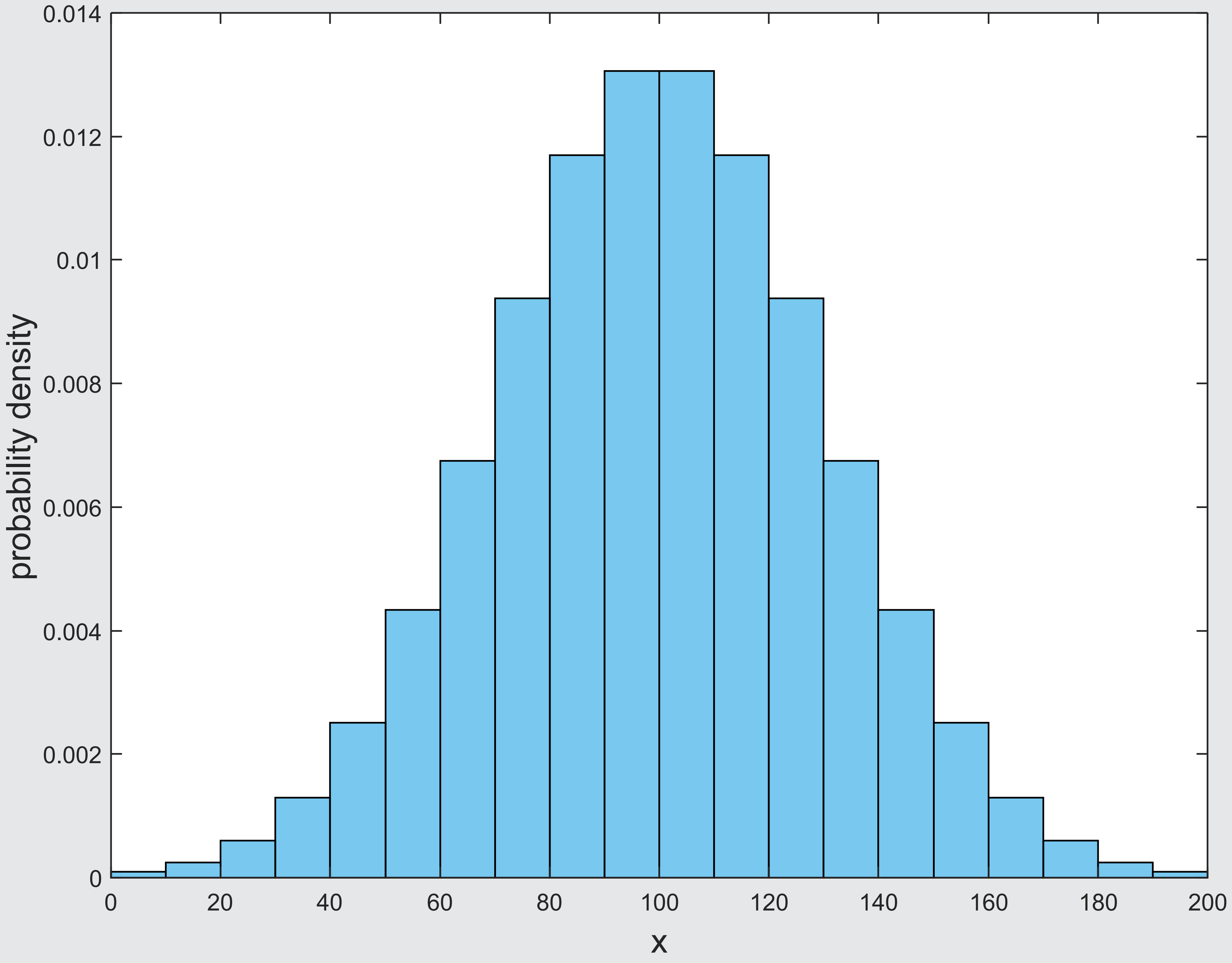
$$\frac{p(\text{ 4 legs } | \text{ cow })}{p(\text{ 4 legs } | \text{ not a cow })}$$

$$\frac{0.98}{0.49} = 2$$

Probability Density

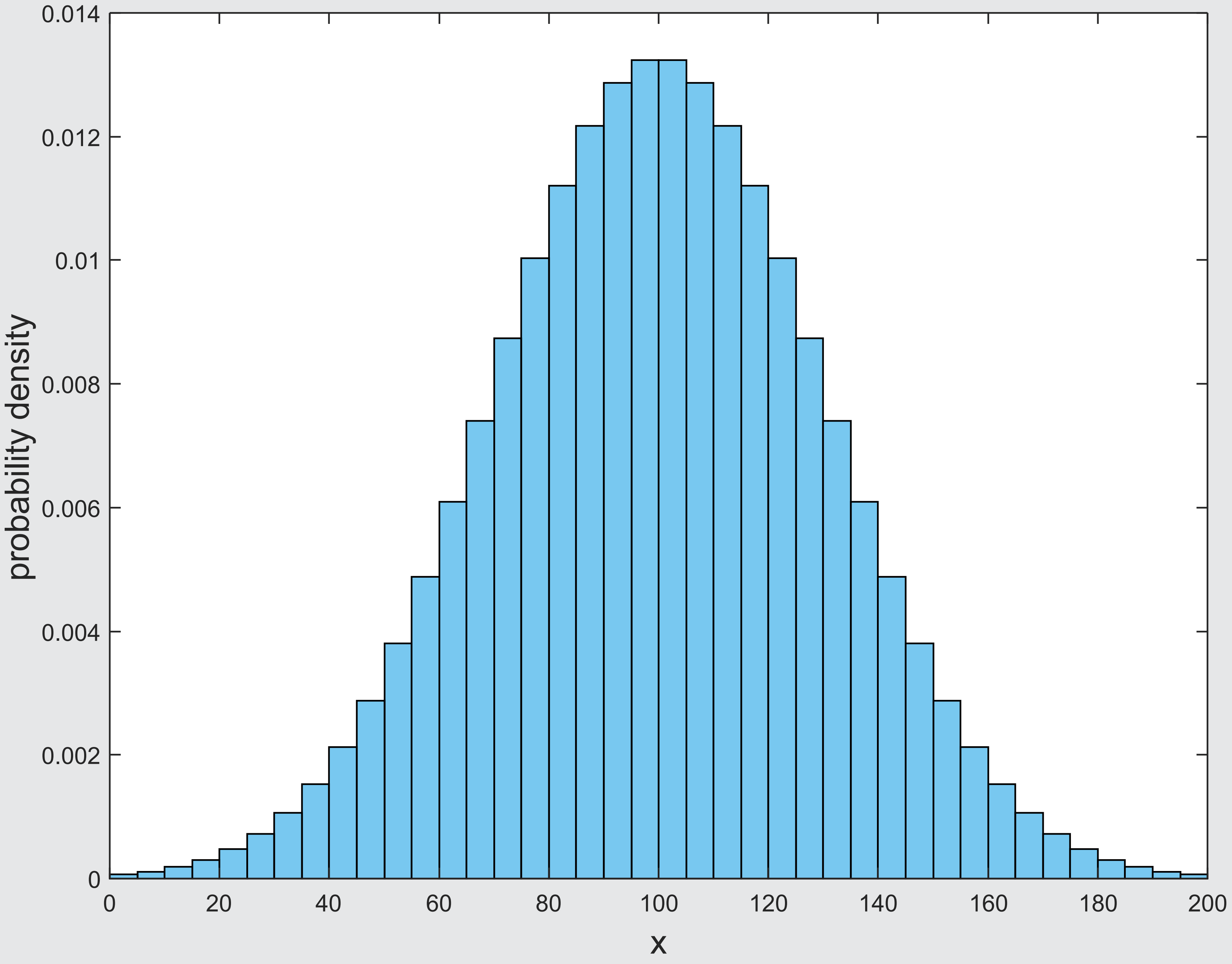
Continuous data

- histogram



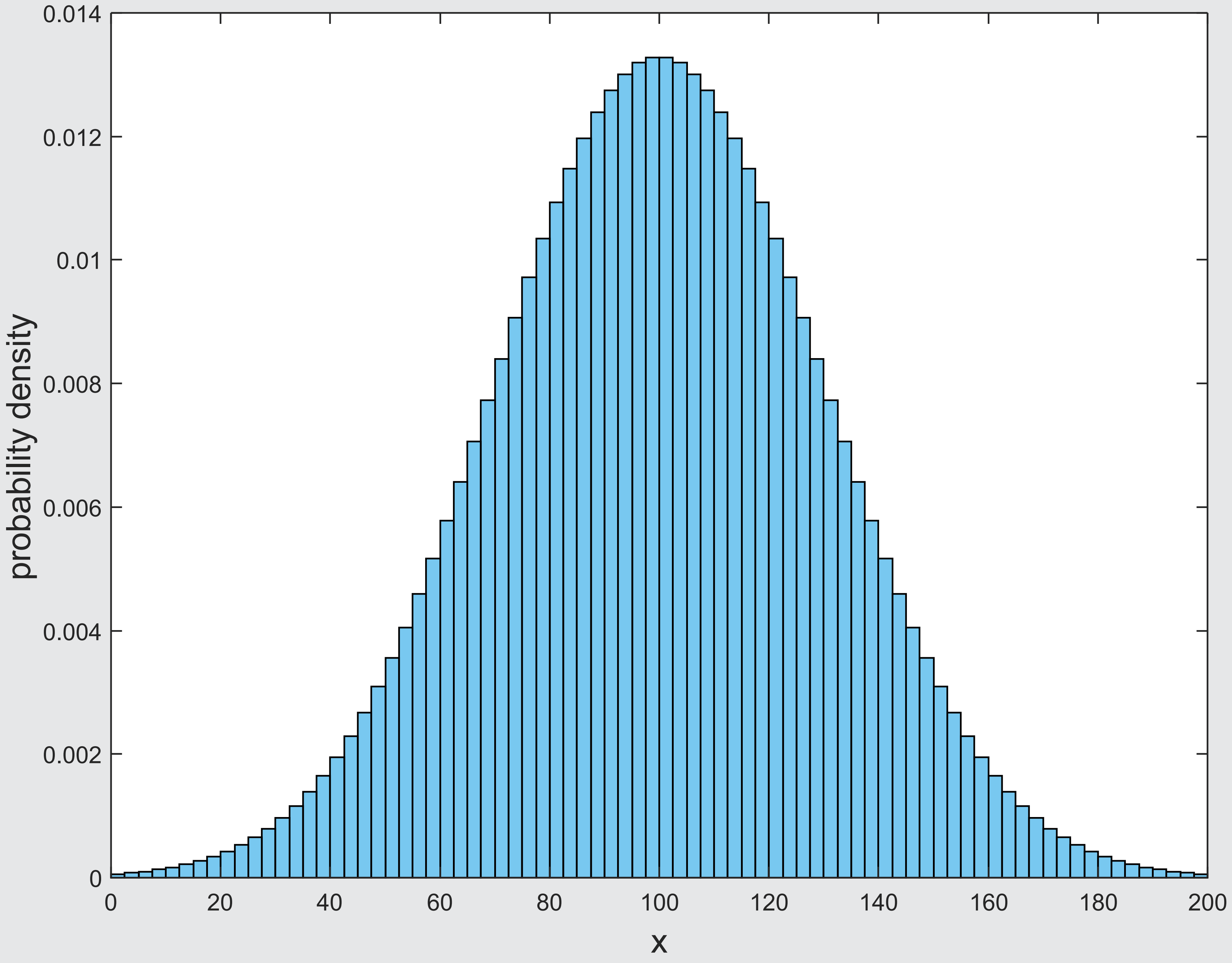
Continuous data

- histogram



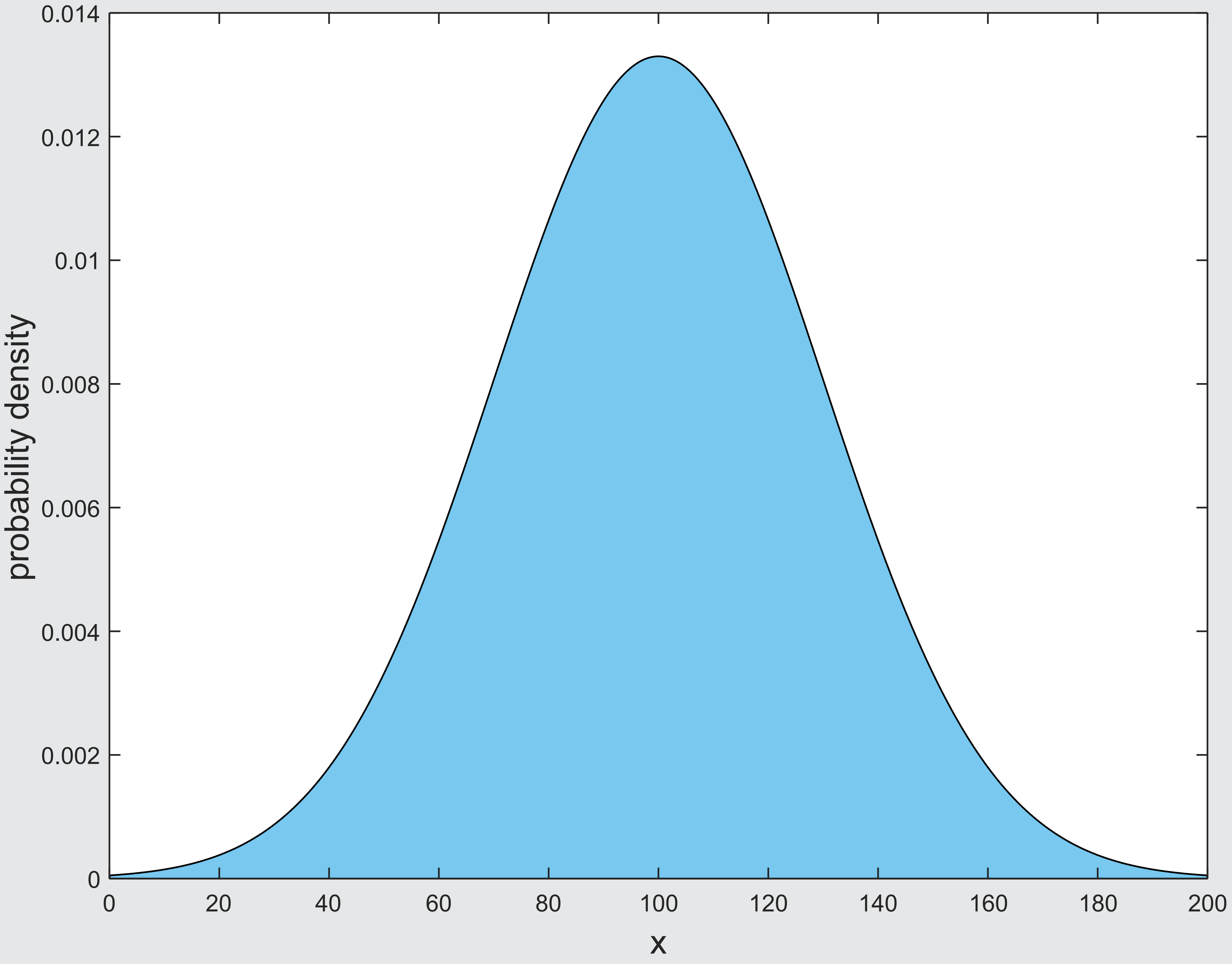
Continuous data

- histogram



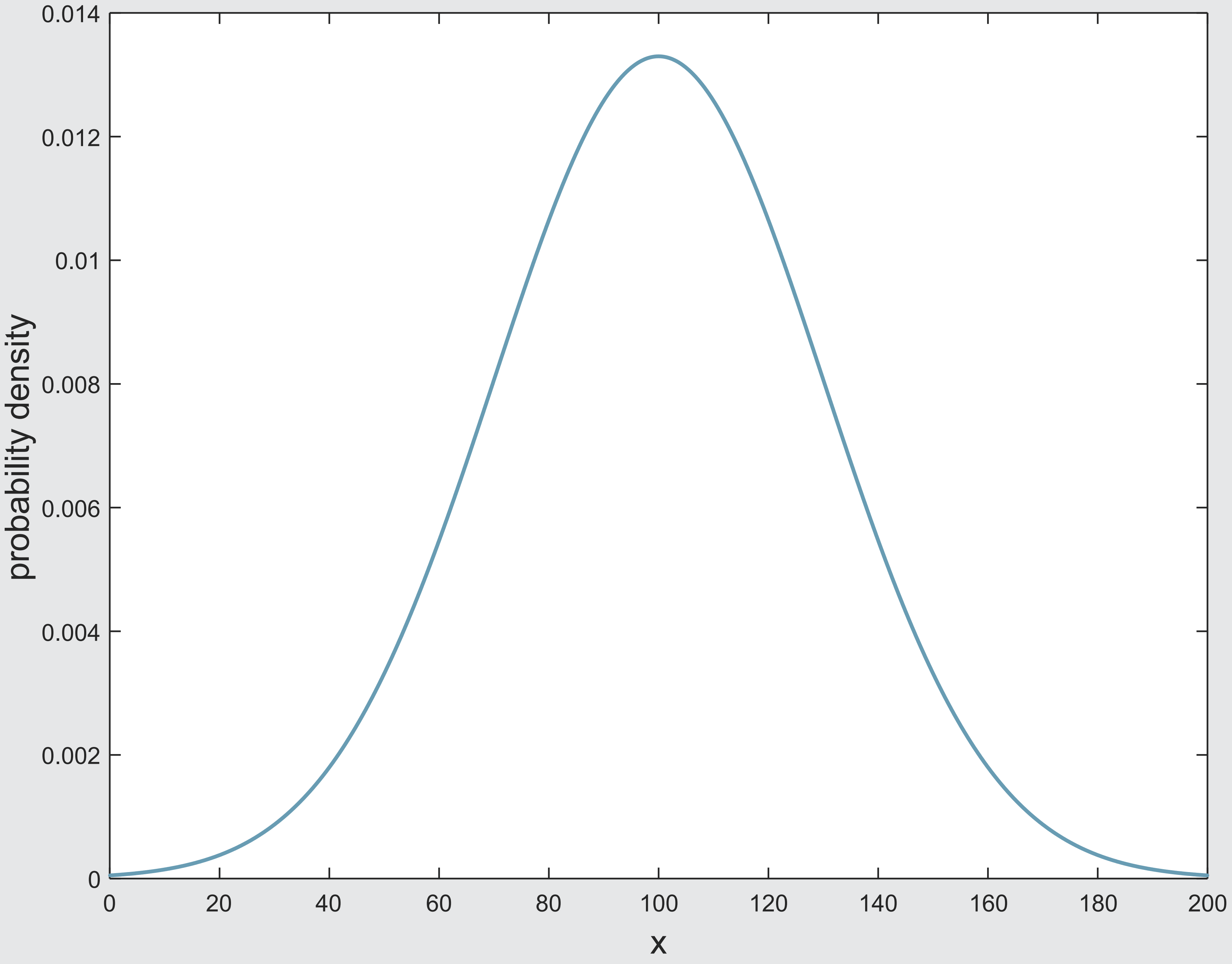
Continuous data

- histogram → probability-density function (pdf)



Continuous data

- probability-density function (pdf)
    - $\mu = 100$
    - $\sigma = 30$



Continuous data

- probability-density function (pdf)

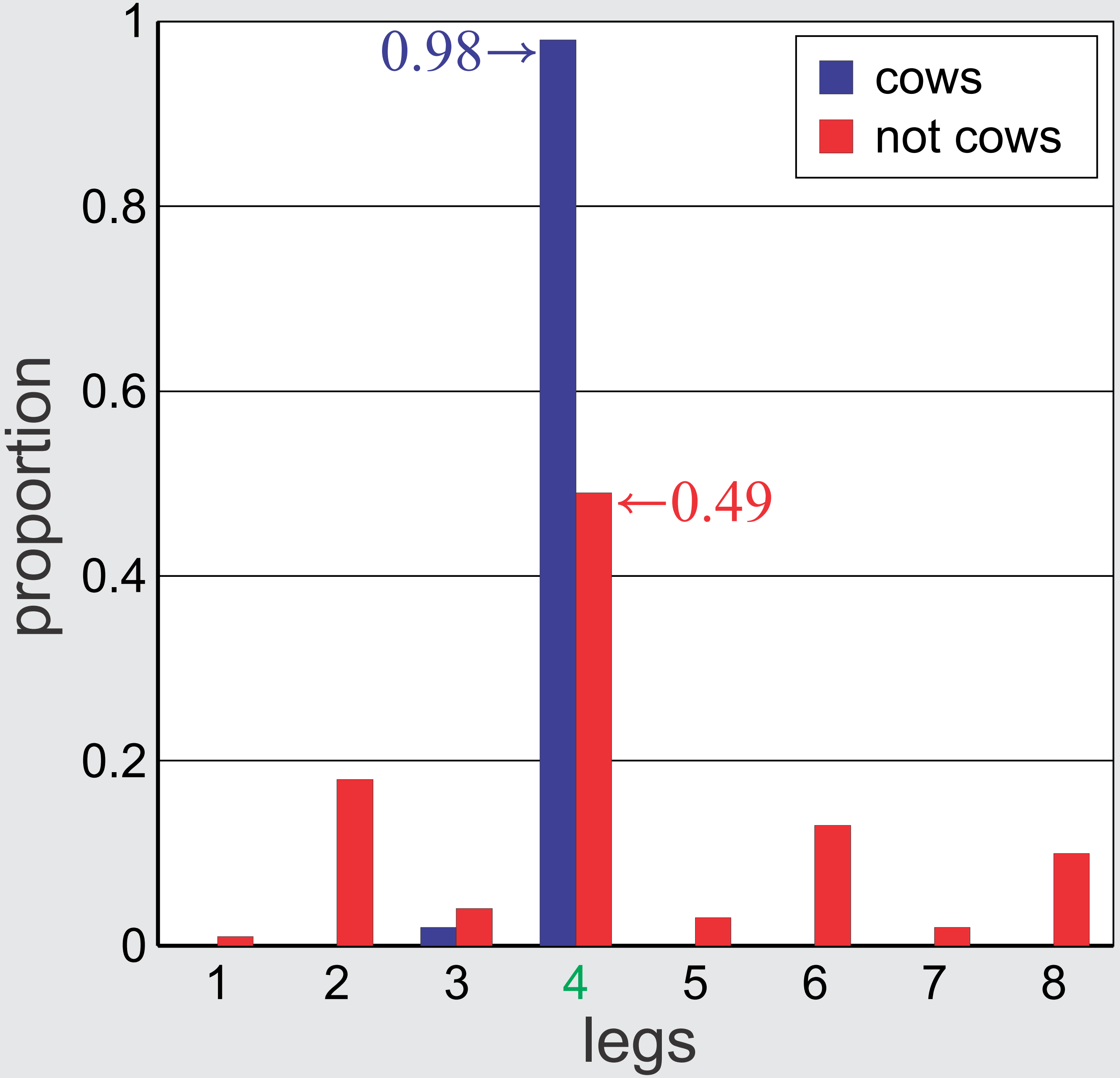
  - Gaussian distribution: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

  - Note the use of the symbol *f* instead of *p*

Specific-Source Likelihood Ratios

Discrete data

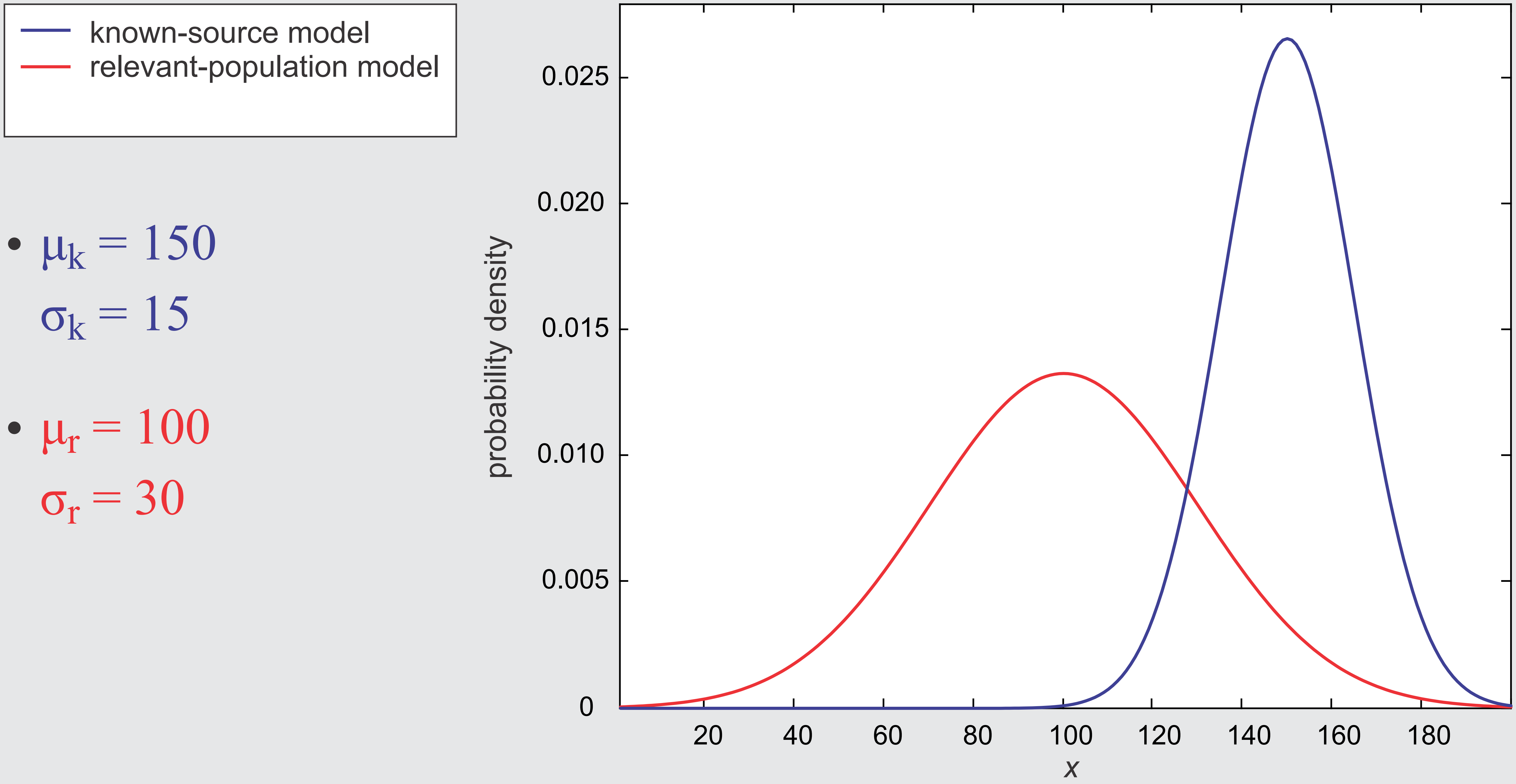
- Bar graph



$$\frac{p(\text{ 4 legs } | \text{ cow })}{p(\text{ 4 legs } | \text{ not a cow })}$$

$$\frac{0.98}{0.49} = 2$$

Continuous data



- $\mu_k = 150$
  $\sigma_k = 15$

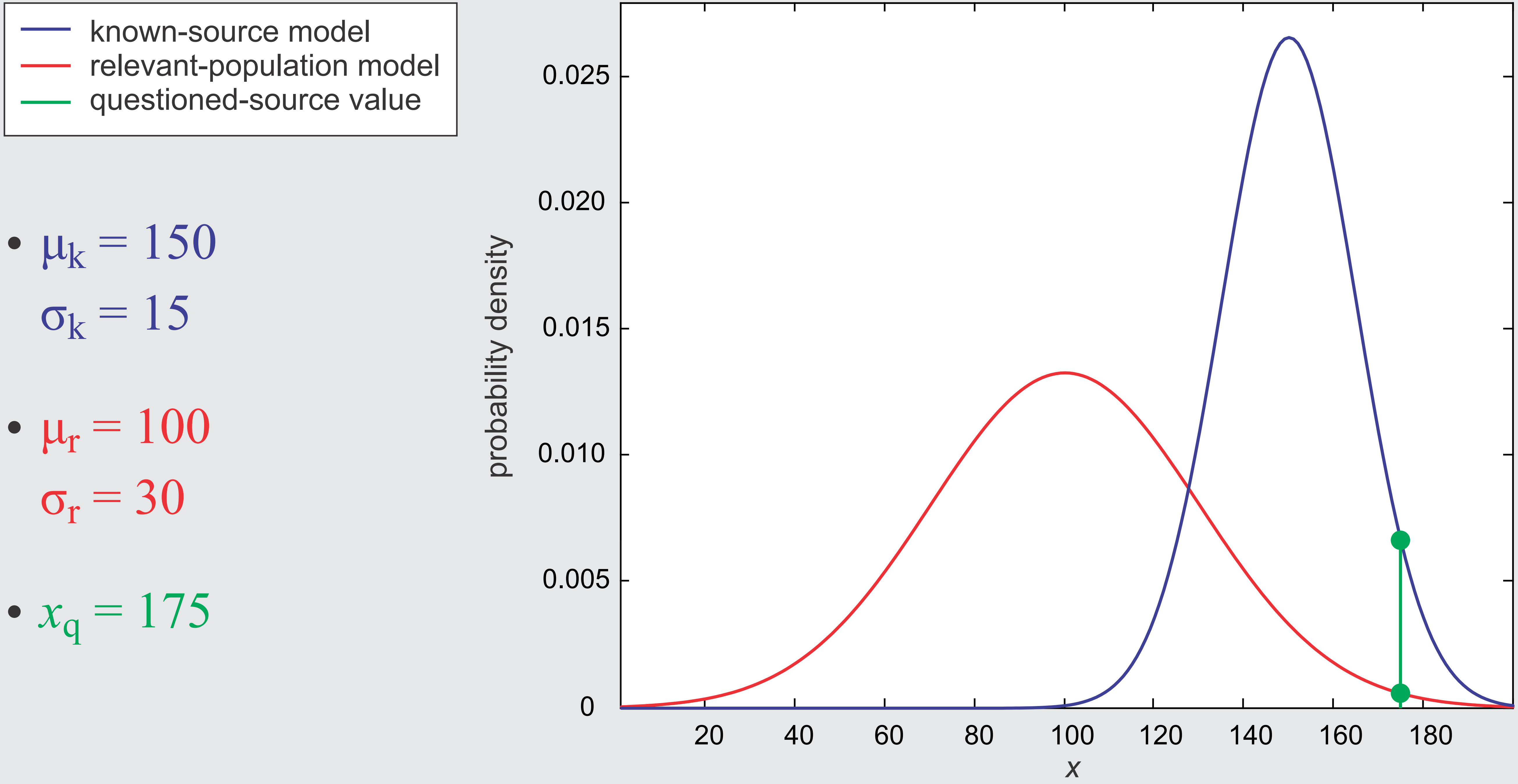
- $\mu_r = 100$
  $\sigma_r = 30$

Continuous data

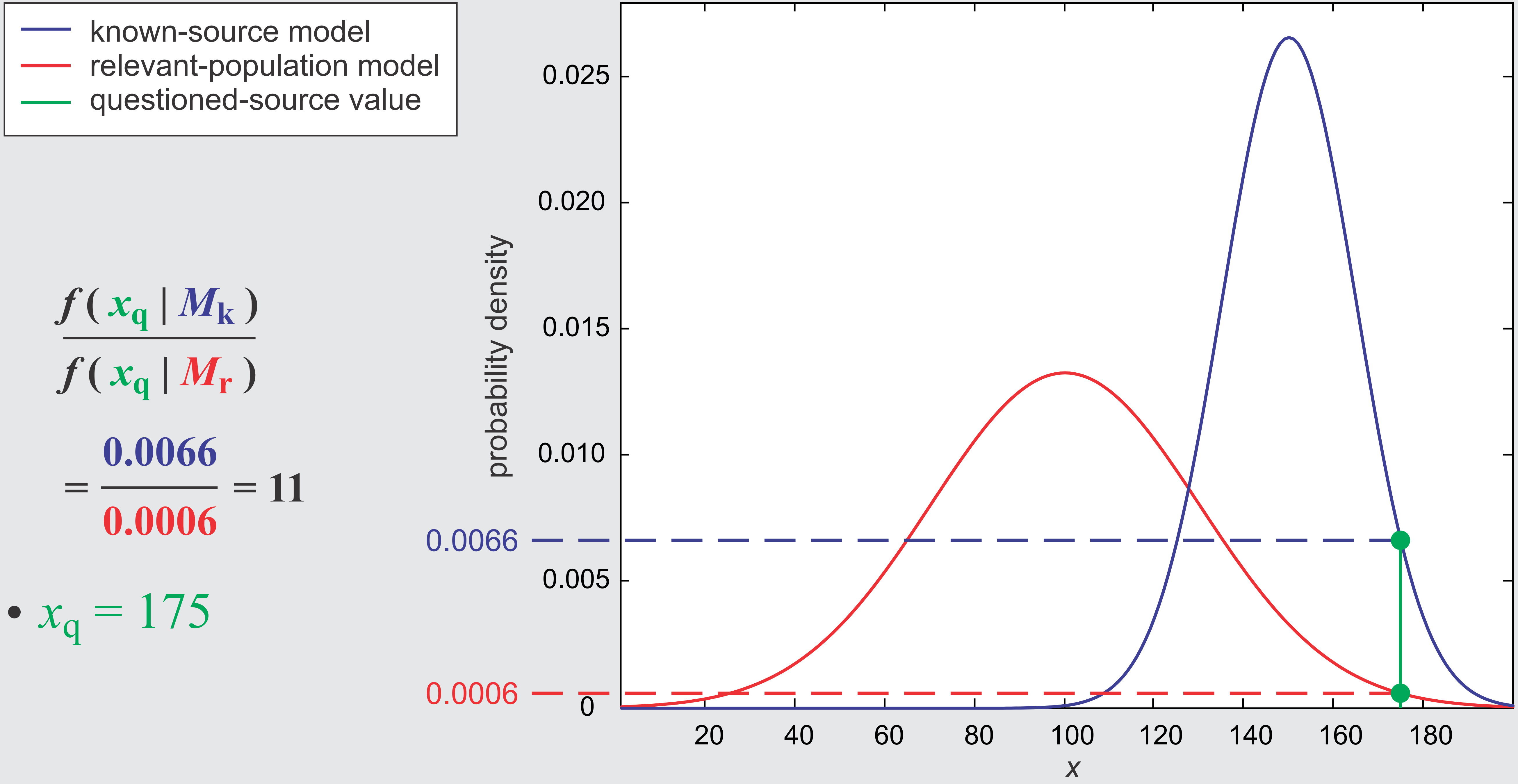
- $\mu_k = 150$
  $\sigma_k = 15$

- $\mu_r = 100$
  $\sigma_r = 30$

- $x_q = 175$



Continuous data



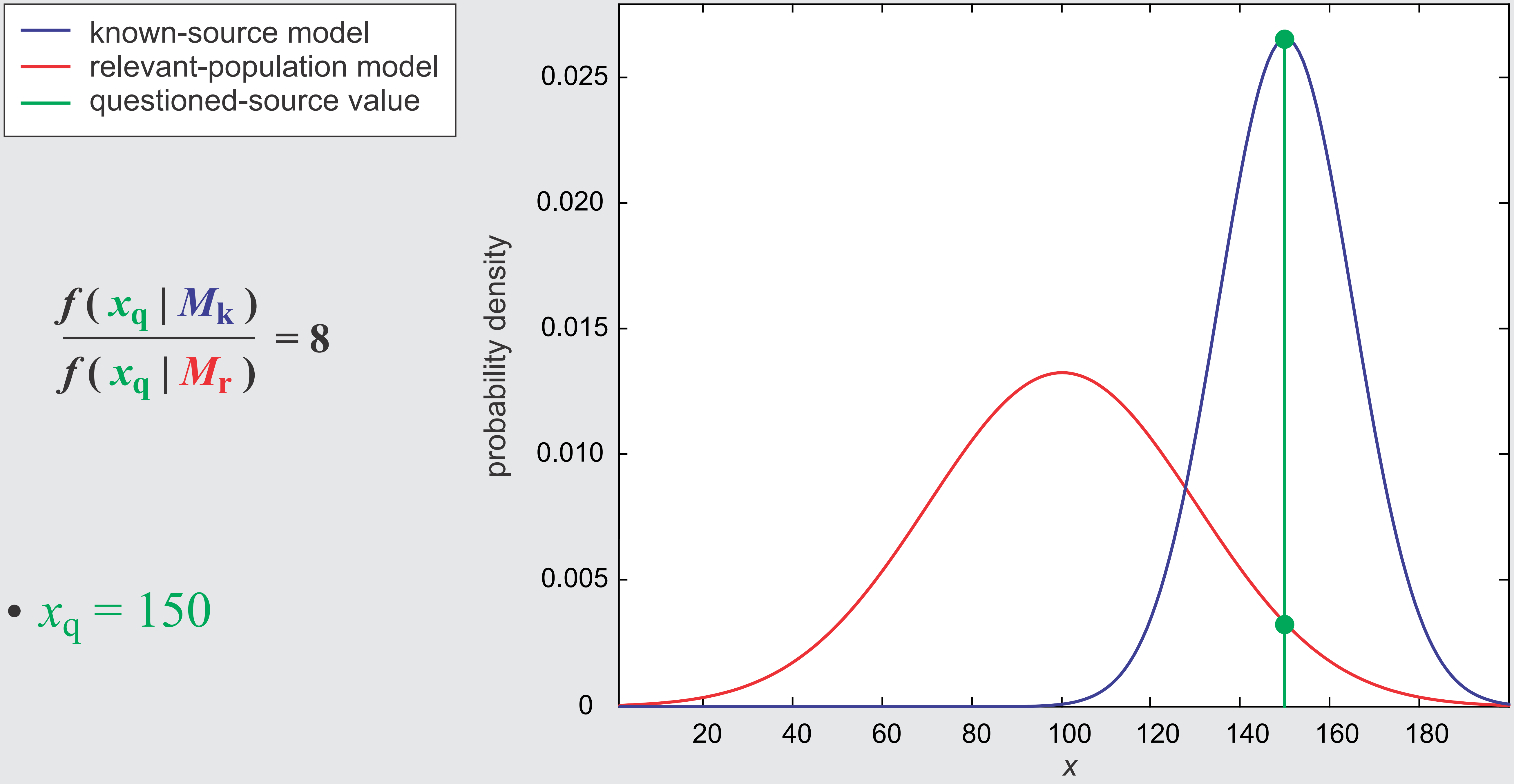
$$\frac{f(x_q \mid M_k)}{f(x_q \mid M_r)} = \frac{0.0066}{0.0006} = 11$$

- $x_q = 175$

Continuous data

$$\frac{f(x_q \mid M_k)}{f(x_q \mid M_r)} = 8$$

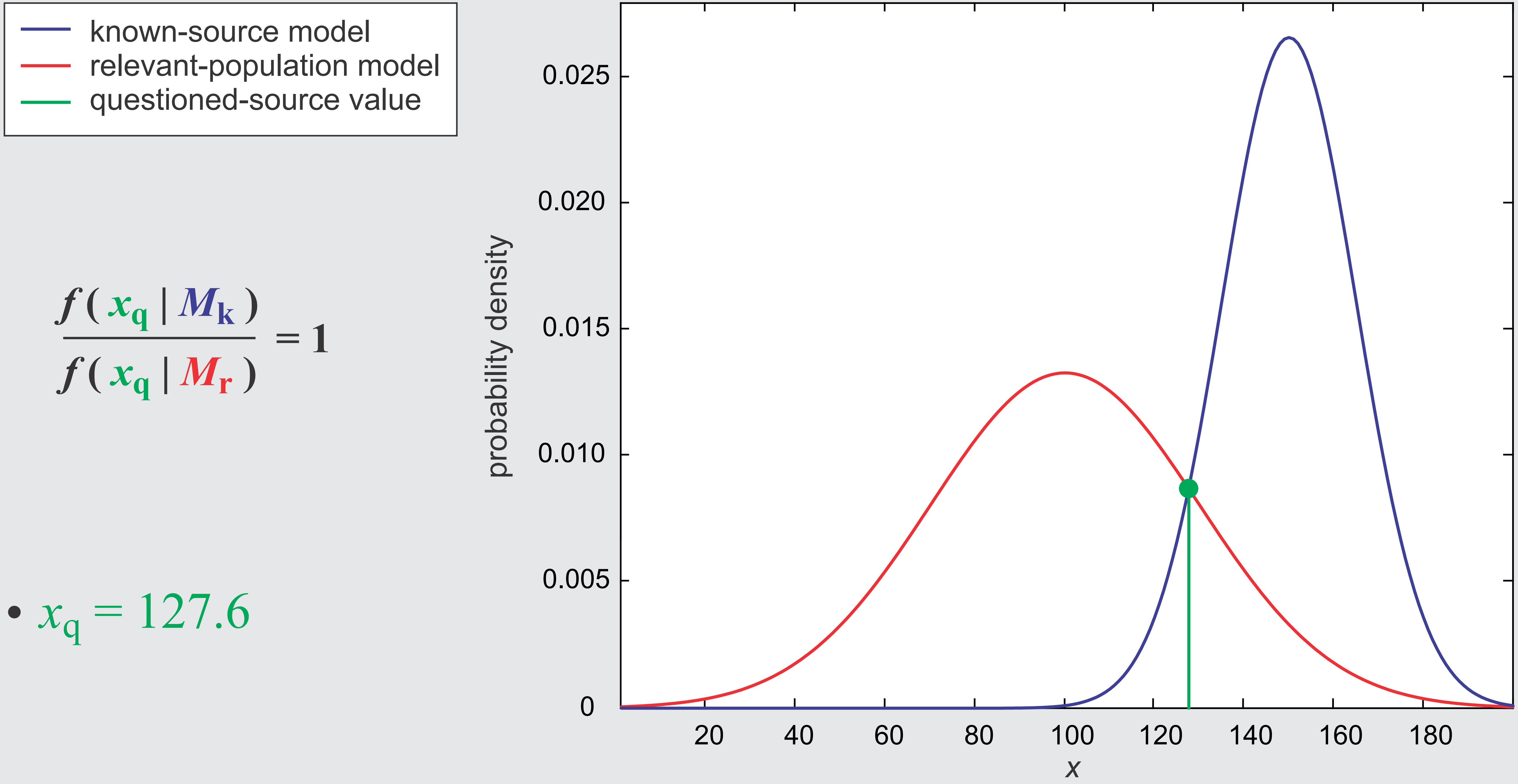
- $x_q = 150$



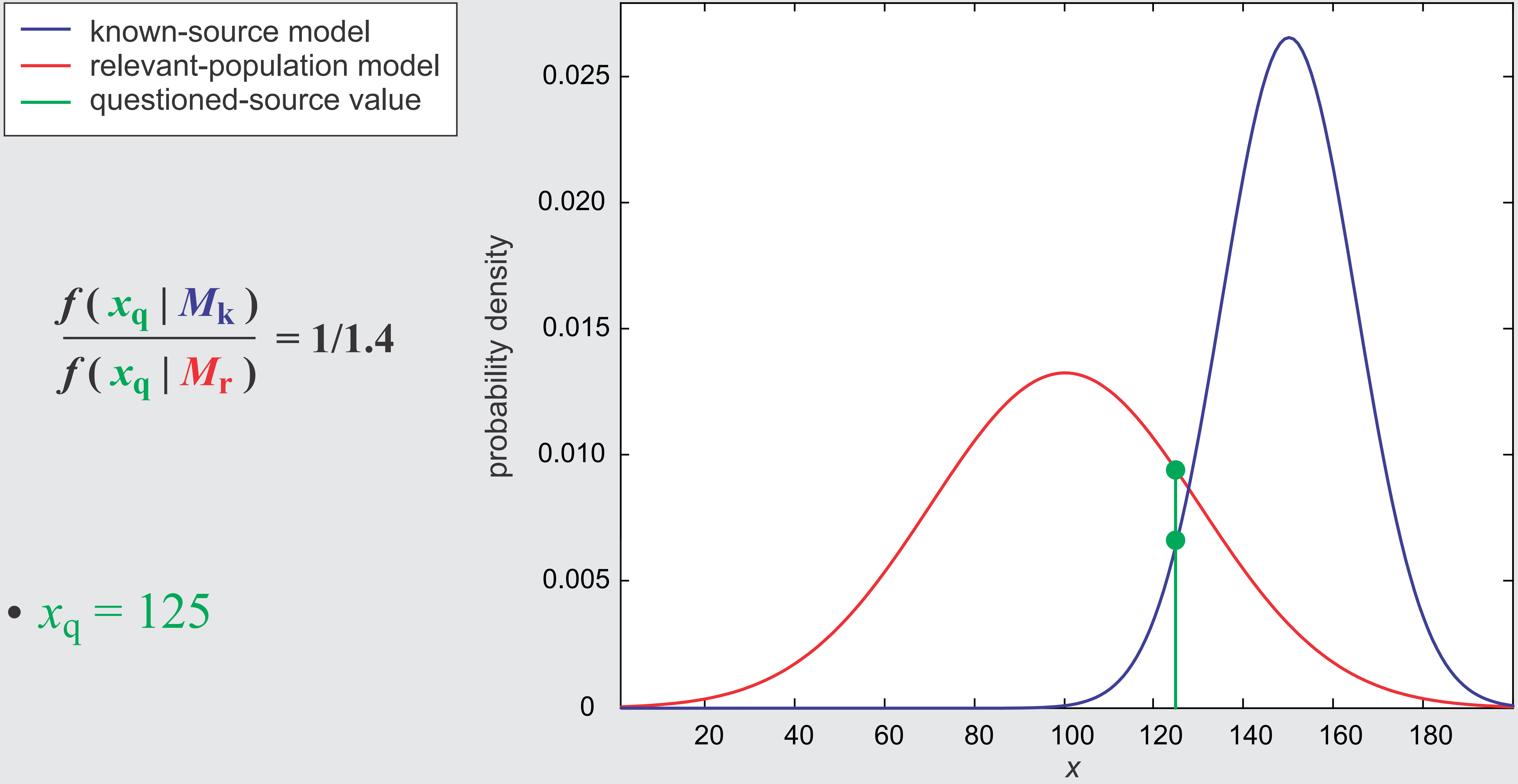
Continuous data

$$\frac{f(x_q \mid M_k)}{f(x_q \mid M_r)} = 1$$

- $x_q = 127.6$



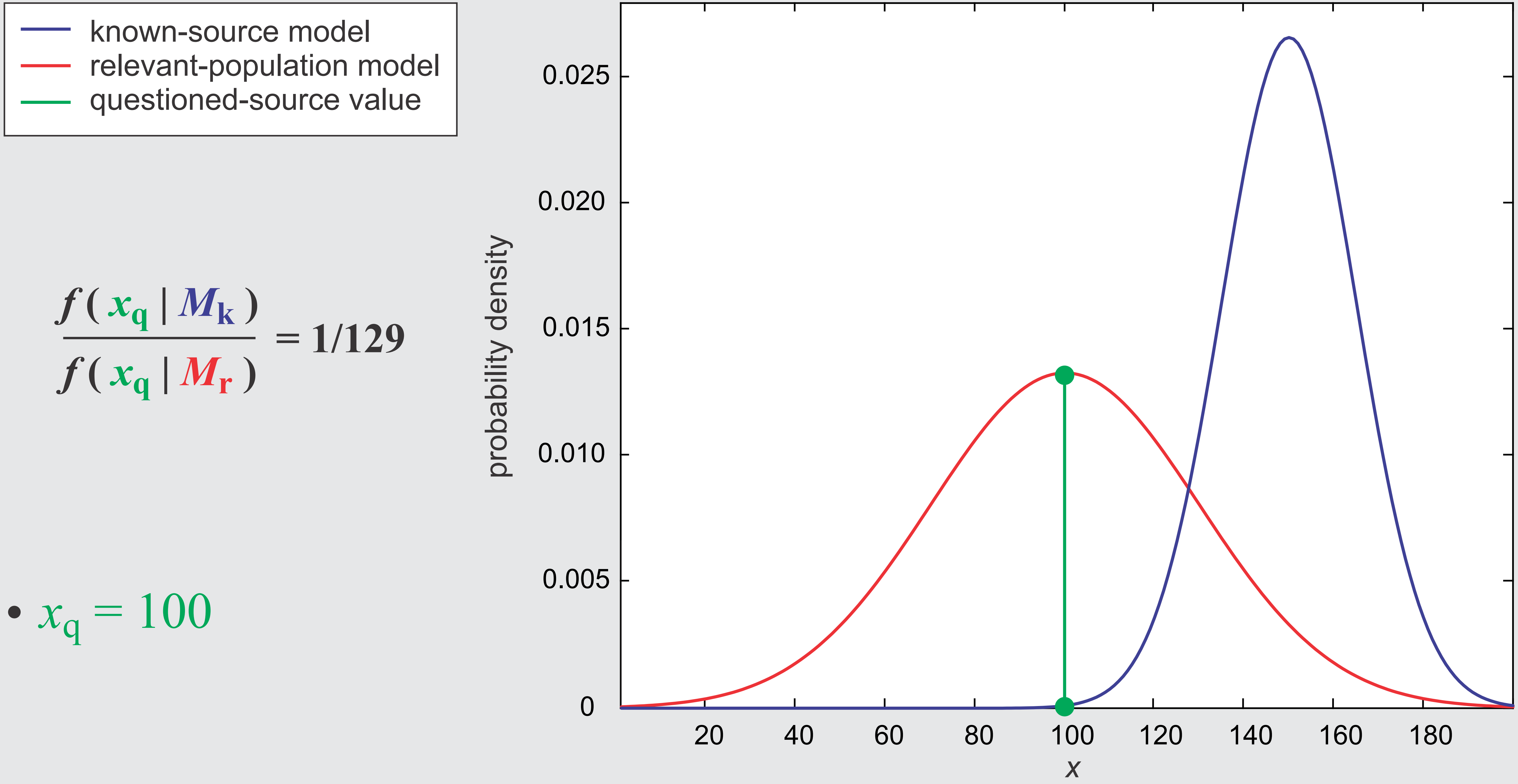
Continuous data



$$\frac{f(x_q \mid M_k)}{f(x_q \mid M_r)} = 1/1.4$$

- $x_q = 125$

Continuous data



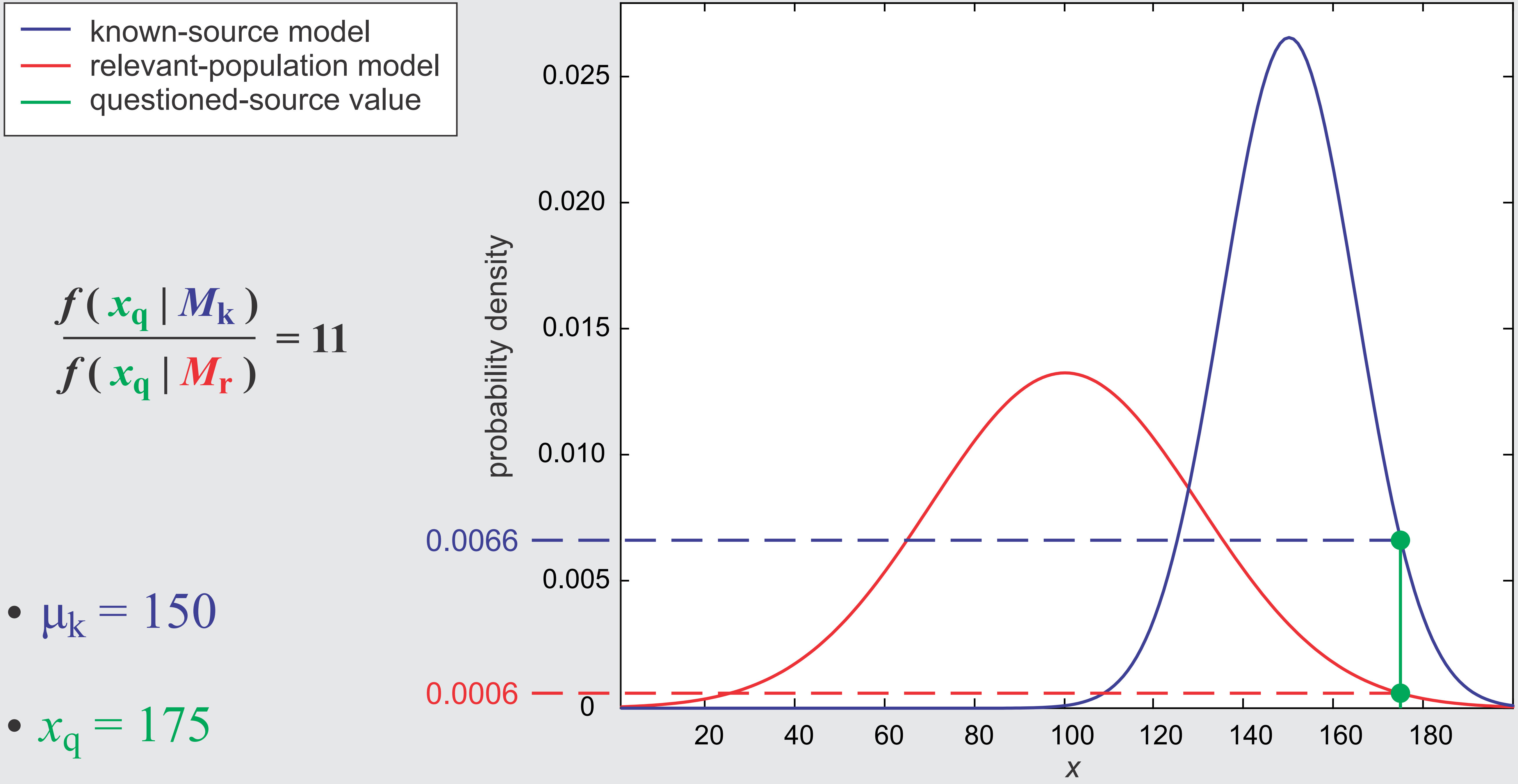
$$\frac{f(x_q \mid M_k)}{f(x_q \mid M_r)} = 1/129$$

- $x_q = 100$

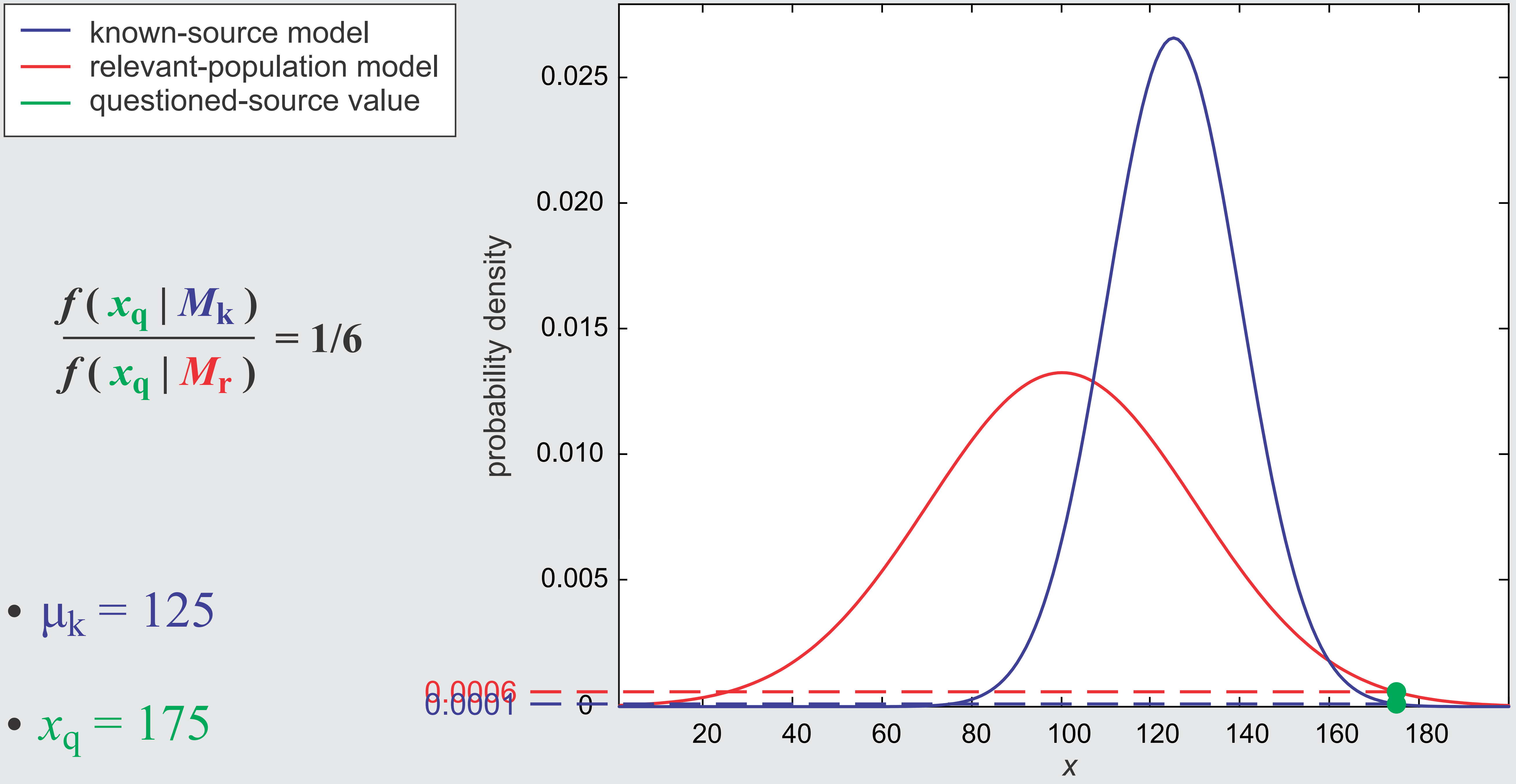
Continuous data

$$\frac{f(x_q \mid M_k)}{f(x_q \mid M_r)} = 11$$

- $\mu_k = 150$
- $x_q = 175$



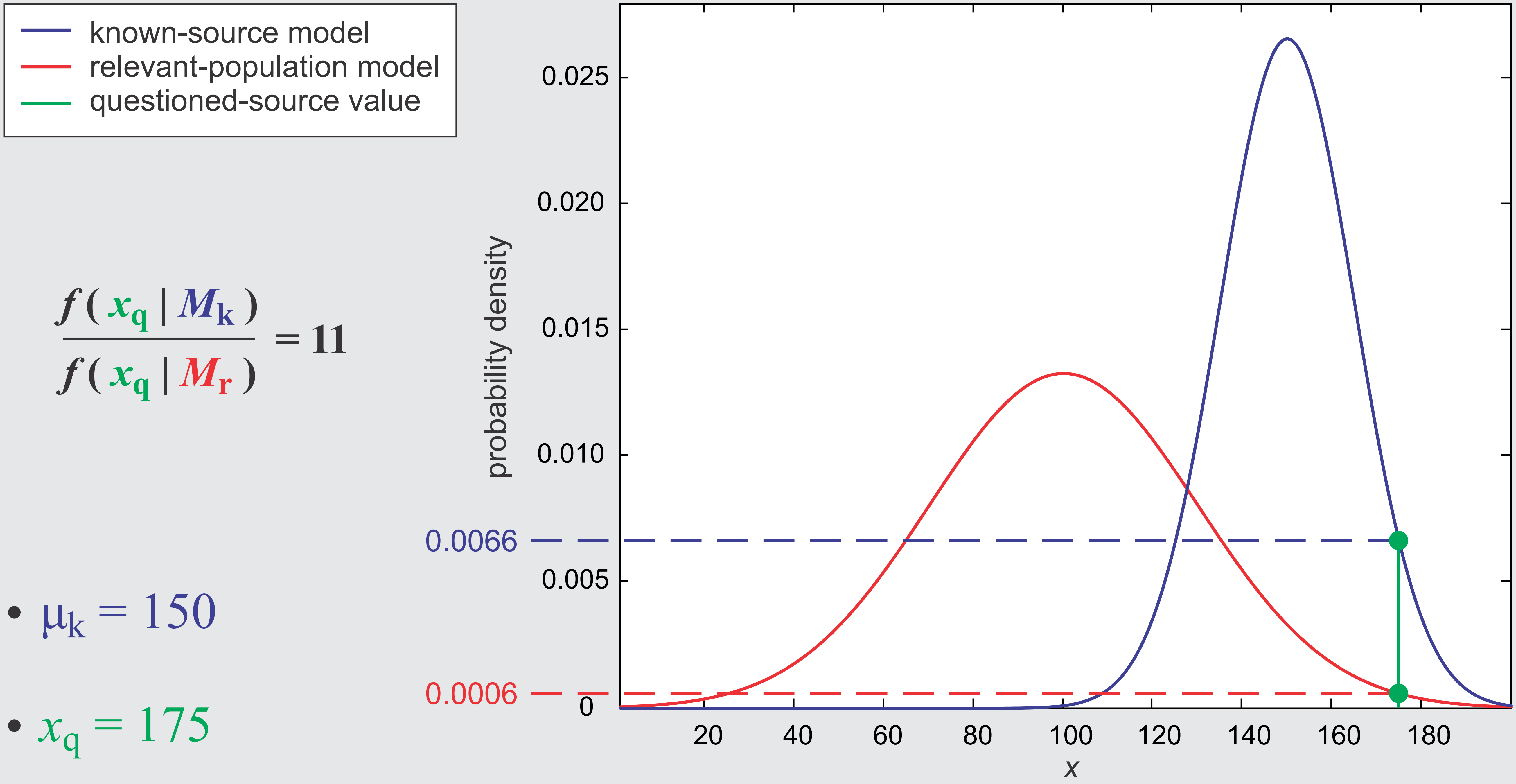
Continuous data



$$\frac{f(x_q \mid M_k)}{f(x_q \mid M_r)} = 1/6$$

- $\mu_k = 125$
- $x_q = 175$

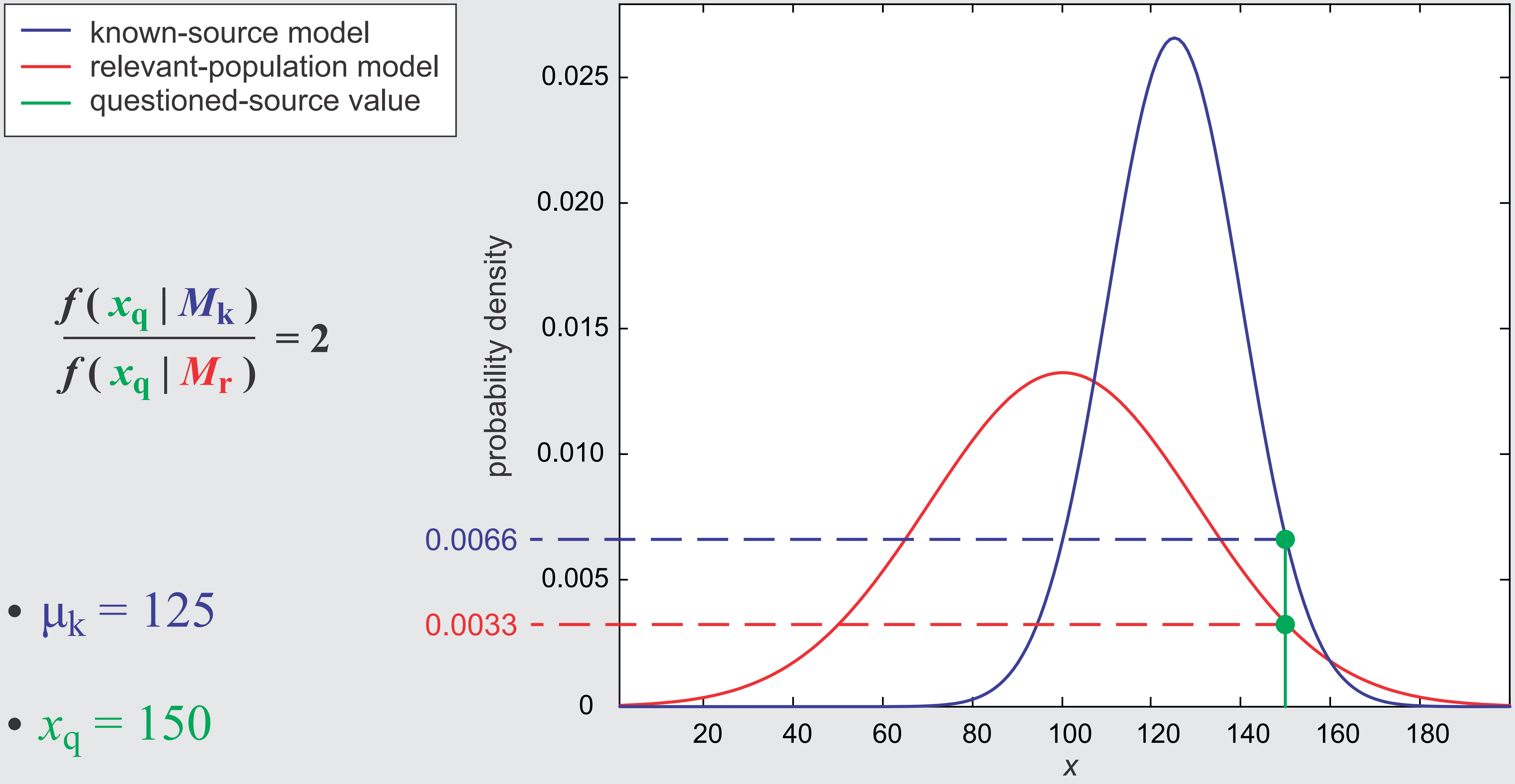
Continuous data



$$\frac{f(x_q \mid M_k)}{f(x_q \mid M_r)} = 11$$

- $\mu_k = 150$
- $x_q = 175$

Continuous data



$$\frac{f(x_q \mid M_k)}{f(x_q \mid M_r)} = 2$$

- $\mu_k = 125$
- $x_q = 150$

Specific-source likelihood ratio

H_1: The item of questioned source came from the specific known source

versus

H_2: The item of questioned source came not from the specific known source but from some other source selected at random from the relevant population

Specific-source likelihood ratio

- What is the probability* of obtaining the measured properties of the item of questioned source if it came from the specific known source?

divided by

- What is the probability* of obtaining the measured properties of the item of questioned source if it came not from the specific known source but from some other source selected at random from the relevant population?

* technically: probability density, likelihood

Specific-source likelihood ratio

$$\Lambda = \frac{f\left(x_{\mathrm{q}} \middle| M_{\mathrm{k}}\right)}{f\left(x_{\mathrm{q}} \middle| M_{\mathrm{r}}\right)}$$

$$\Lambda = \frac{f\left(x_{\mathrm{q}} \middle| \mu_{\mathrm{k}}, \sigma_{\mathrm{k}}^{2}\right)}{f\left(x_{\mathrm{q}} \middle| \mu_{\mathrm{r}}, \sigma_{\mathrm{r}}^{2}\right)} = \frac{f\left(x_{\mathrm{q}} \middle| \mu_{\mathrm{k}}, \sigma_{\mathrm{w}}^{2}\right)}{f\left(x_{\mathrm{q}} \middle| \mu_{\mathrm{r}}, \sigma_{\mathrm{w}}^{2} + \sigma_{\mathrm{b}}^{2}\right)}$$

Specific-source likelihood ratio

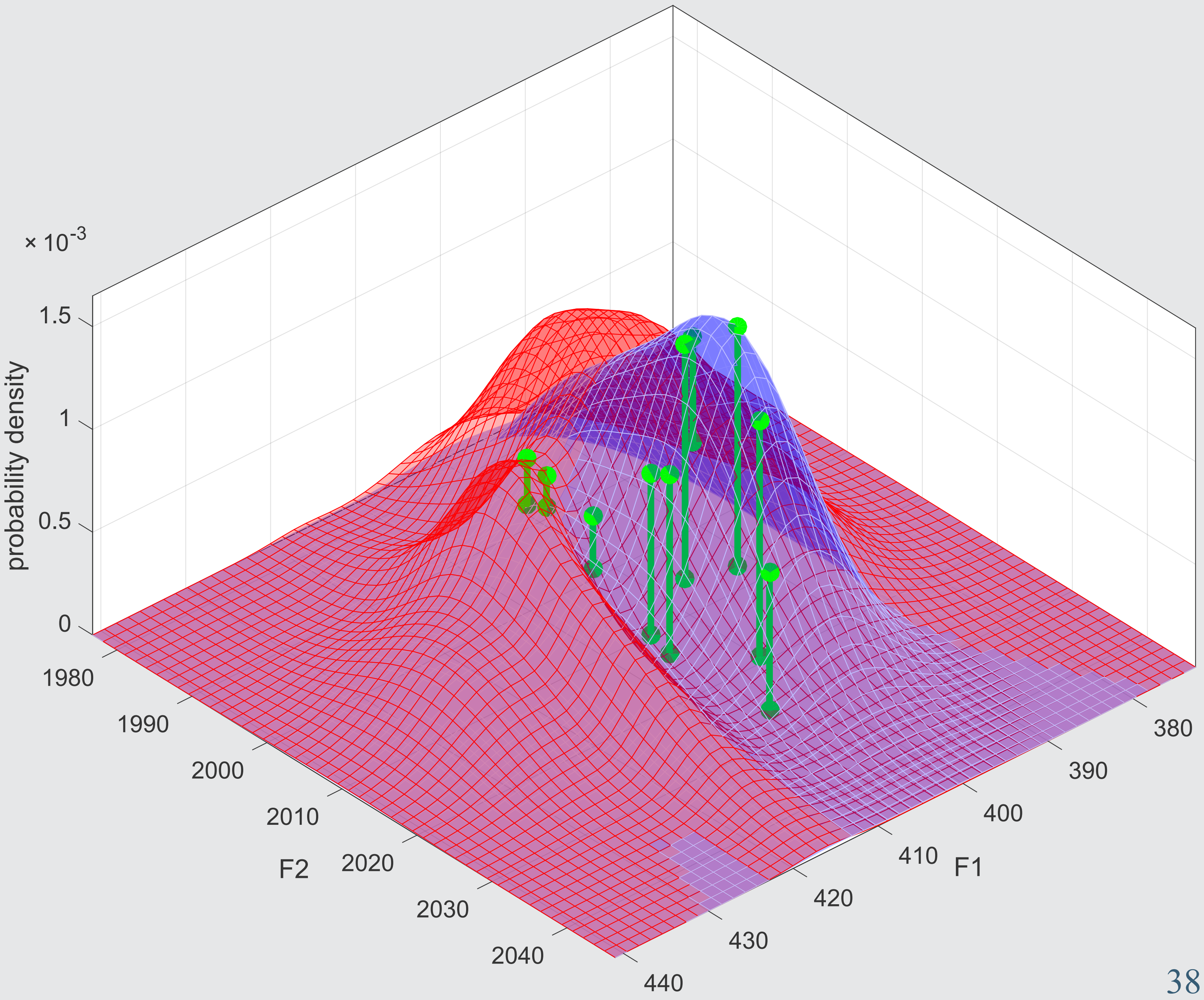
$$\Lambda = \frac{f\left(x_{\mathrm{q}} \middle| M_{\mathrm{k}}\right)}{f\left(x_{\mathrm{q}} \middle| M_{\mathrm{r}}\right)}$$

$$\Lambda = \frac{f\left(x_{\mathrm{q}} \middle| \mu_{\mathrm{k}}, \sigma_{\mathrm{k}}^{2}\right)}{f\left(x_{\mathrm{q}} \middle| \mu_{\mathrm{r}}, \sigma_{\mathrm{r}}^{2}\right)} = \frac{f\left(x_{\mathrm{q}} \middle| \mu_{\mathrm{k}}, \sigma_{\mathrm{w}}^{2}\right)}{f\left(x_{\mathrm{q}} \middle| \mu_{\mathrm{r}}, \sigma_{\mathrm{w}}^{2} + \sigma_{\mathrm{b}}^{2}\right)}$$

- The smaller the within-source variance compared to the between-source variance, the further from 1 likelihood ratios can potentially be.

Continuous data

- multivariate data
- more complex distributions



Common-Source Likelihood Ratios

Specific-source likelihood ratio

H_1: The item of questioned source came from the specific known source

versus

H_2: The item of questioned source came not from the specific known source but from some other source selected at random from the relevant population

Common-source likelihood ratio

H_1: The items of questioned- and known-source both came from the same source (a source selected at random from the relevant population)

versus

H_2: The items of questioned- and known-source each came from a different source (each a source selected at random from the relevant population)

Specific-source likelihood ratio

- What is the probability* of obtaining the measured properties of the item of questioned source if it came from the specific known source?

  divided by

- What is the probability* of obtaining the measured properties of the item of questioned source if it came not from the specific known source but from some other source selected at random from the relevant population?

* technically: probability density, likelihood

Common-source likelihood ratio

- What is the probability* of obtaining the measured properties of the items of questioned- and known-source if they both came from the same source (a source selected at random from the relevant population)?

  divided by

- What is the probability* of obtaining the measured properties of the items of questioned- and known-source if they each came from a different source (each a source selected at random from the relevant population)?

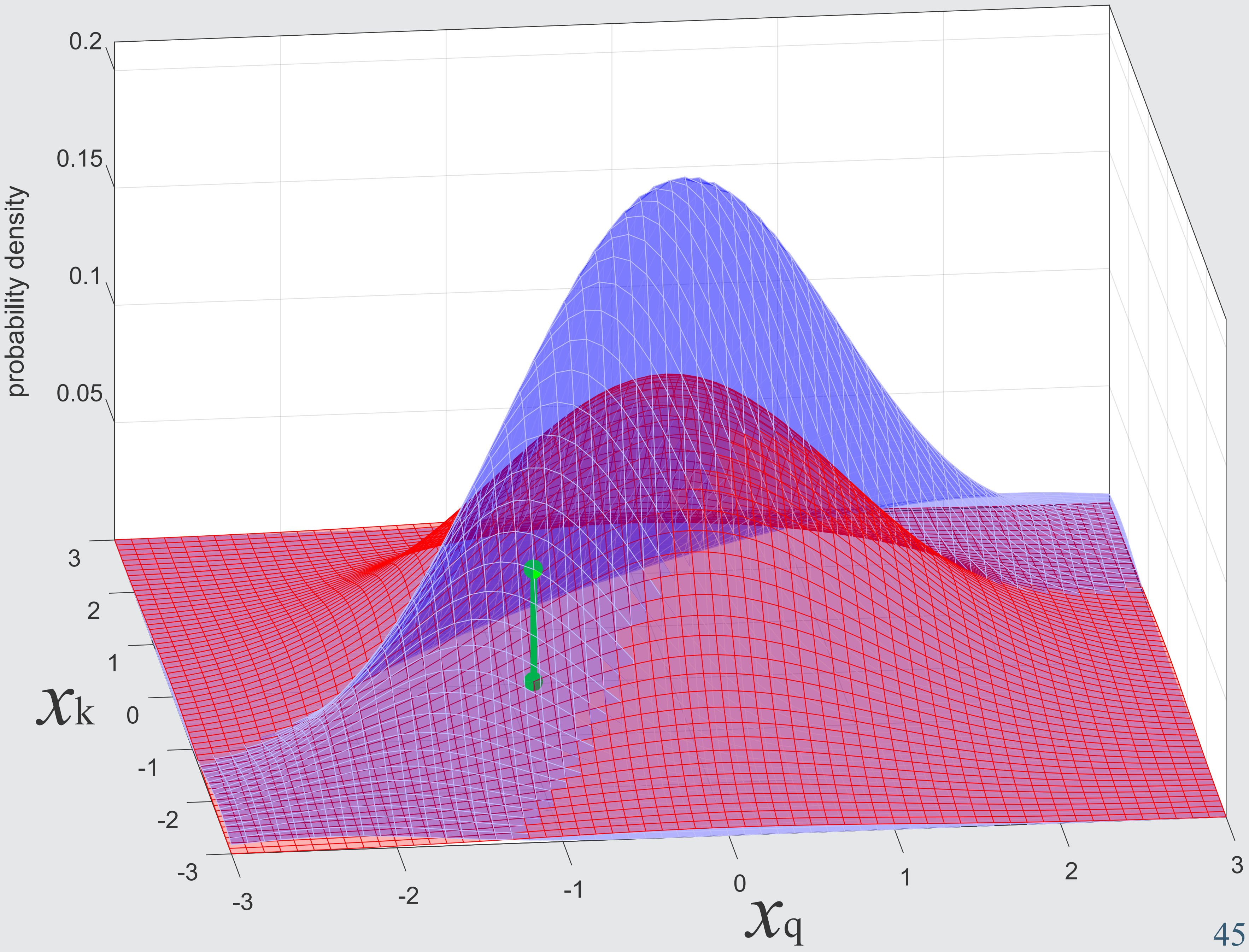
* technically: probability density, likelihood

Common-source likelihood ratio

$$\Lambda = \frac{f\left(x_{\mathrm{q}}, x_{\mathrm{k}} \middle| M_{\mathrm{s}}\right)}{f\left(x_{\mathrm{q}} \middle| M_{\mathrm{d}}\right) f\left(x_{\mathrm{k}} | M_{\mathrm{d}}\right)}$$

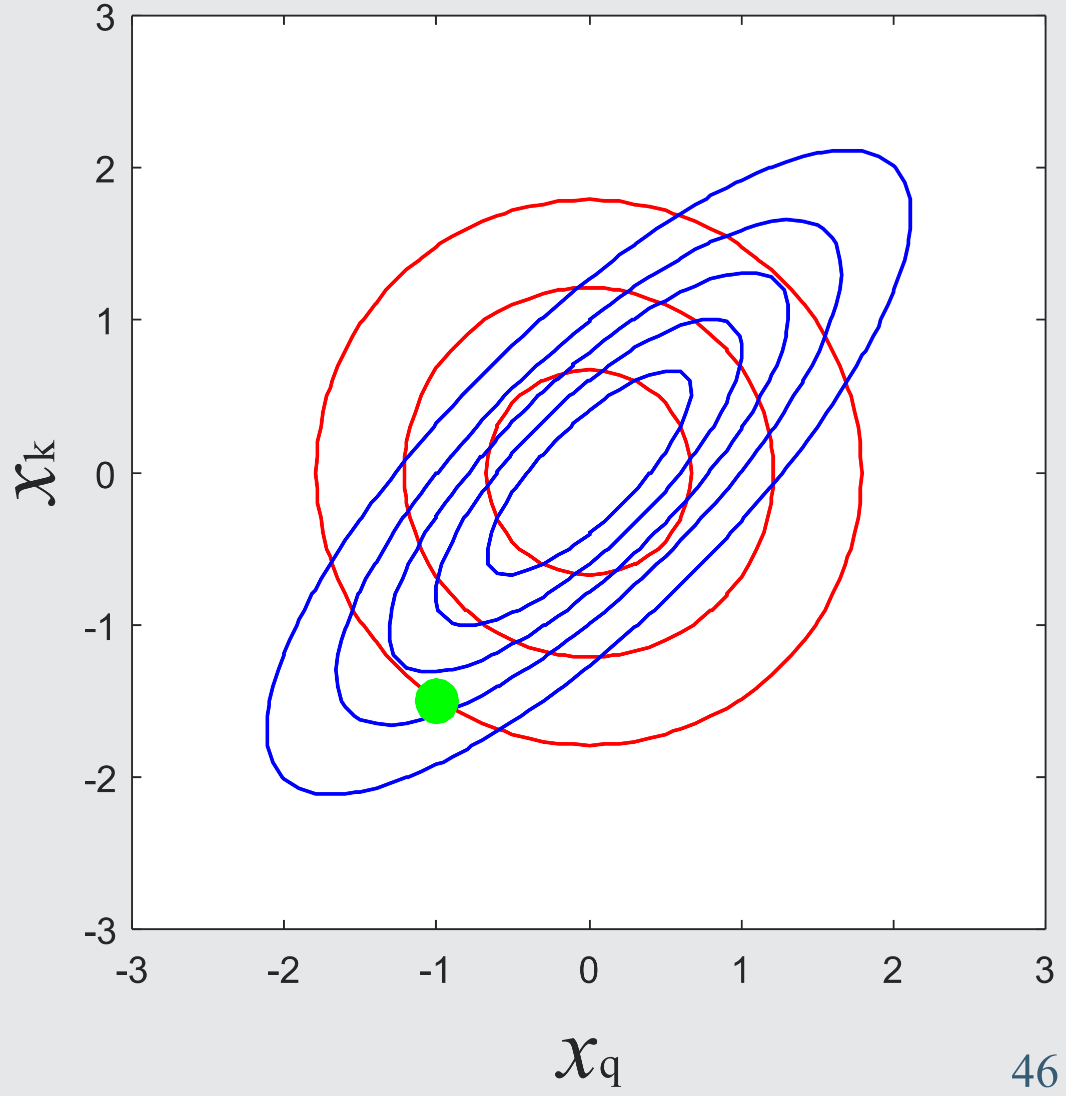
$$\Lambda = \frac{f\left(\begin{bmatrix} x_{\mathrm{q}} \\ x_{\mathrm{k}} \end{bmatrix} \middle| \begin{bmatrix} \mu_{\mathrm{r}} \\ \mu_{\mathrm{r}} \end{bmatrix}, \begin{bmatrix} \sigma_{\mathrm{w}}^2 + \sigma_{\mathrm{b}}^2 & \sigma_{\mathrm{b}}^2 \\ \sigma_{\mathrm{b}}^2 & \sigma_{\mathrm{w}}^2 + \sigma_{\mathrm{b}}^2 \end{bmatrix}\right)}{f\left(\begin{bmatrix} x_{\mathrm{q}} \\ x_{\mathrm{k}} \end{bmatrix} \middle| \begin{bmatrix} \mu_{\mathrm{r}} \\ \mu_{\mathrm{r}} \end{bmatrix}, \begin{bmatrix} \sigma_{\mathrm{w}}^2 + \sigma_{\mathrm{b}}^2 & 0 \\ 0 & \sigma_{\mathrm{w}}^2 + \sigma_{\mathrm{b}}^2 \end{bmatrix}\right)}$$

Common-source likelihood ratio



Common-source likelihood ratio

$$\Lambda = \frac{f\left(\begin{bmatrix} x_\mathrm{q} \\ x_\mathrm{k} \end{bmatrix} \middle| \begin{bmatrix} \mu_\mathrm{r} \\ \mu_\mathrm{r} \end{bmatrix}, \begin{bmatrix} \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 & \sigma_\mathrm{b}^2 \\ \sigma_\mathrm{b}^2 & \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 \end{bmatrix}\right)}{f\left(\begin{bmatrix} x_\mathrm{q} \\ x_\mathrm{k} \end{bmatrix} \middle| \begin{bmatrix} \mu_\mathrm{r} \\ \mu_\mathrm{r} \end{bmatrix}, \begin{bmatrix} \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 & 0 \\ 0 & \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 \end{bmatrix}\right)}$$



Similarity-Score-Based Likelihood Ratios

Specific-source likelihood ratio

- What is the probability* of obtaining the measured properties of the item of questioned source if it came from the specific known source?

divided by

- What is the probability* of obtaining the measured properties of the item of questioned source if it came not from the specific known source but from some other source selected at random from the relevant population?

* technically: probability density, likelihood

Common-source likelihood ratio

- What is the probability* of obtaining the measured properties of the items of questioned and known source if they both came from the same source (a source selected at random from the relevant population)?

  divided by

- What is the probability* of obtaining the measured properties of the items of questioned and known source if they each came from a different source (each a source selected at random from the relevant population)?

* technically: probability density, likelihood

Similarity-score-based likelihood ratio

- What is the probability* of obtaining the measured degree of similarity between the items of questioned and known source if they both came from the same source (a source selected at random from the relevant population)?

  divided by

- What is the probability* of obtaining the measured degree of similarity between the items of questioned and known source if they each came from a different source (each a source selected at random from the relevant population)?

* technically: probability density, likelihood

Similarity-score-based likelihood ratio

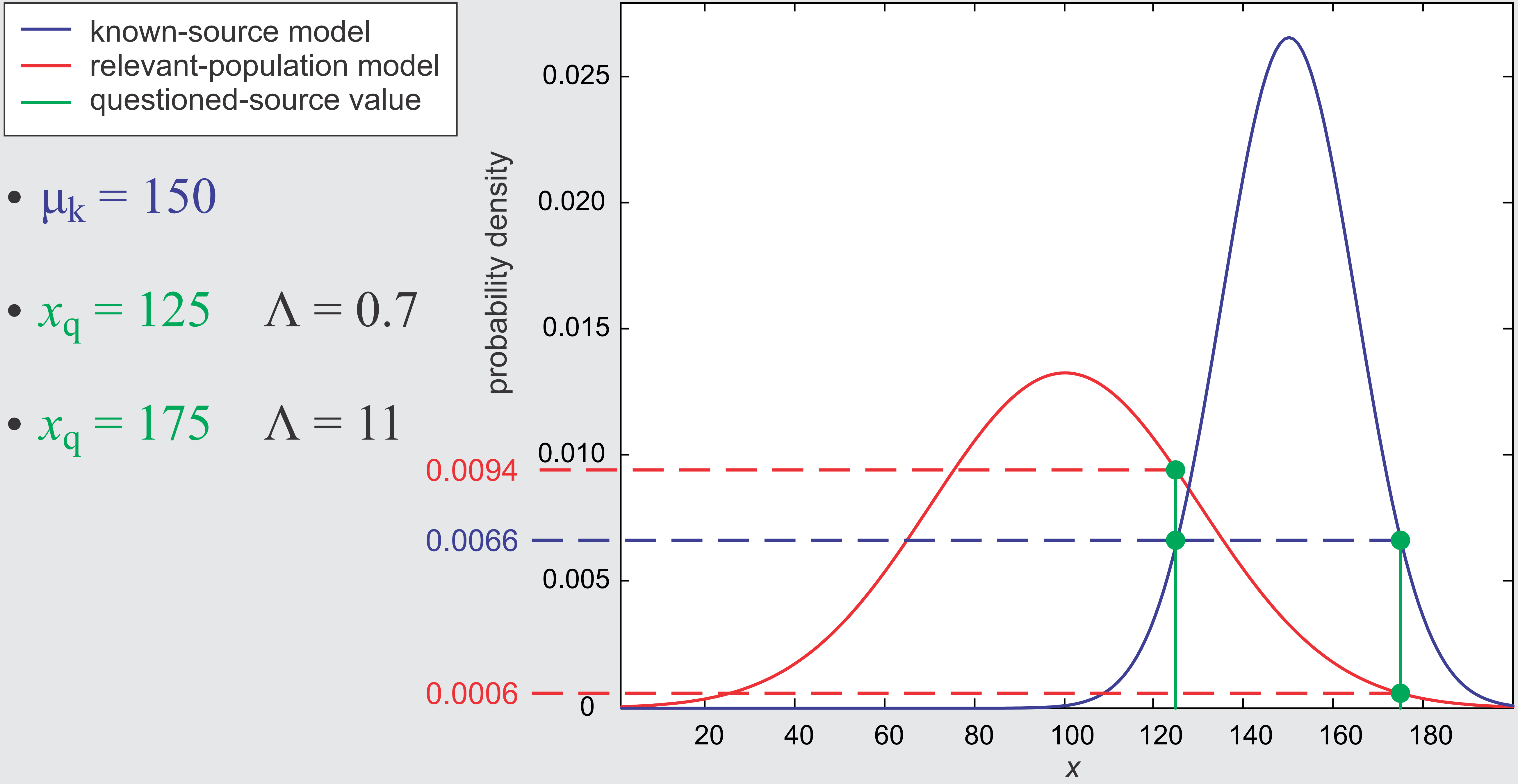
$$\Lambda = \frac{f\left(\delta\left(x_{\mathrm{q}}, x_{\mathrm{k}}\right) \middle| M_{\delta,\mathrm{s}}\right)}{f\left(\delta\left(x_{\mathrm{q}}, x_{\mathrm{k}}\right) \middle| M_{\delta,\mathrm{d}}\right)}$$

$$\delta\left(x_{\mathrm{q}}, x_{\mathrm{k}}\right) = \left|x_{\mathrm{q}} - x_{\mathrm{k}}\right|$$

- degree of similarity is the inverse of degree of difference

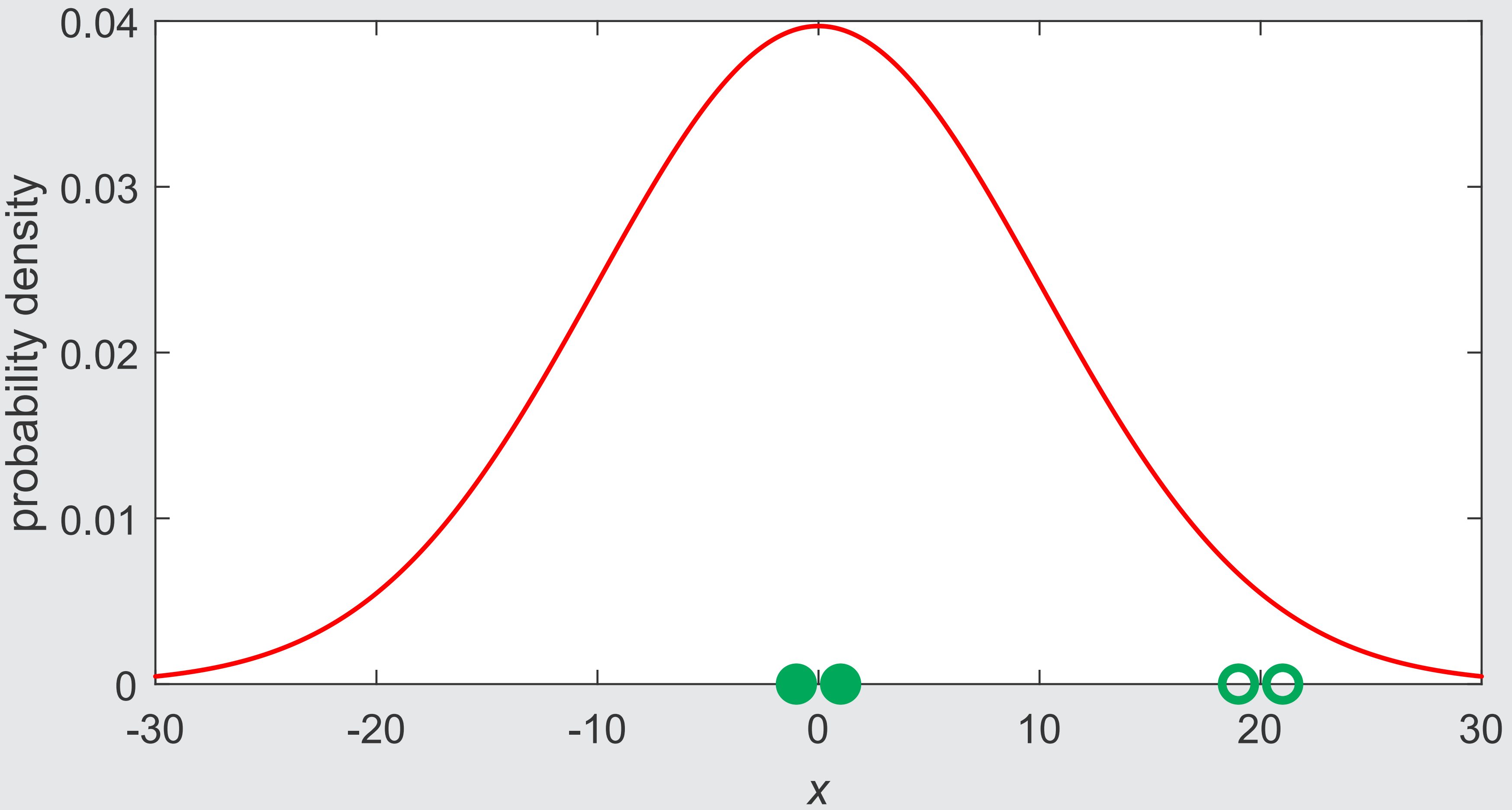
Specific-source likelihood ratio

- $\mu_k = 150$
- $x_q = 125 \quad \Lambda = 0.7$
- $x_q = 175 \quad \Lambda = 11$



Common-source likelihood ratio

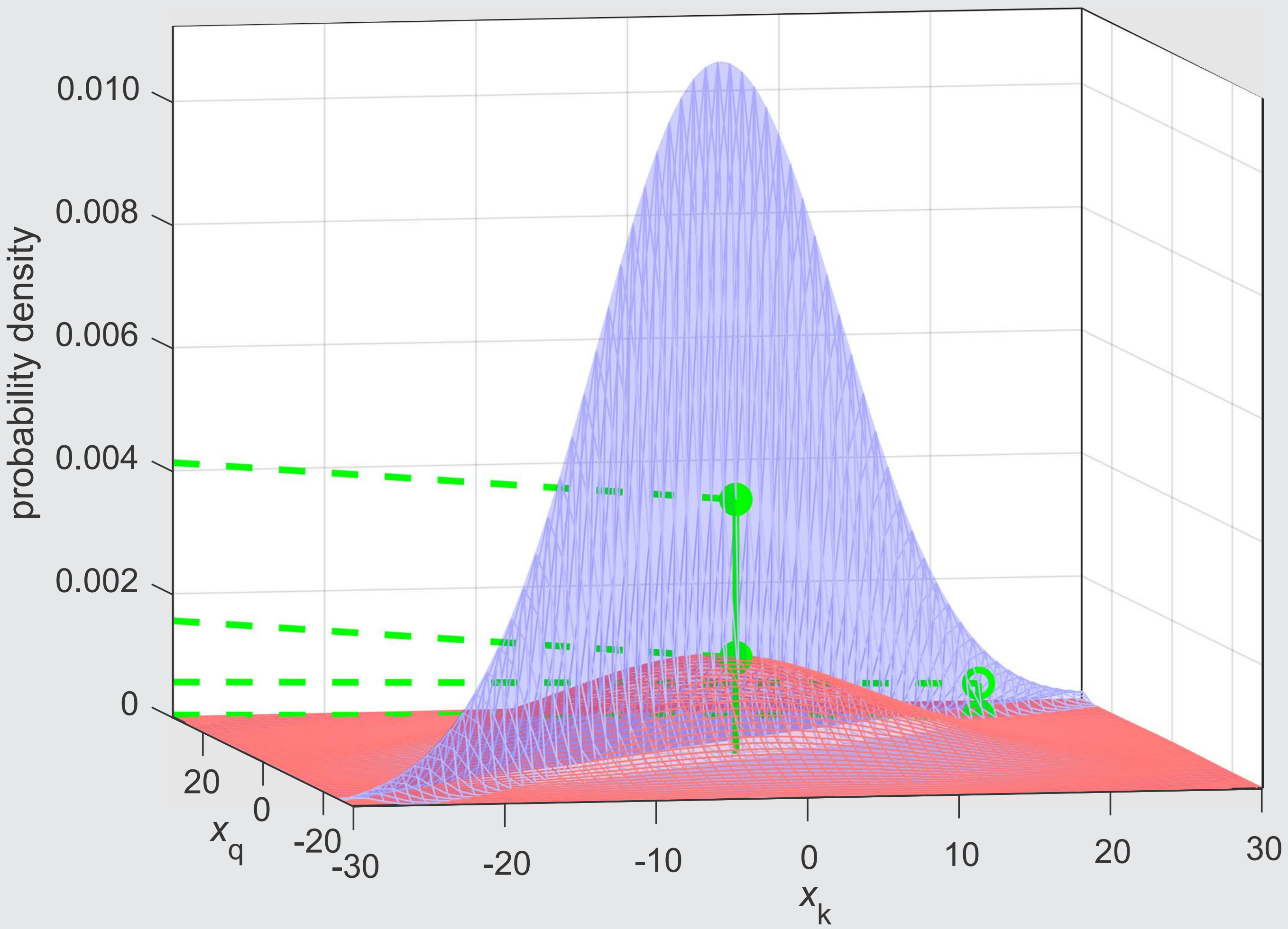
- $\mu_r = 0$
- $\sigma_b^2 = 100$
- $\sigma_w^2 = 1$



- $x_q = -1 \qquad x_k = 1$
- $x_q = 19 \qquad x_k = 21$

Common-source likelihood ratio

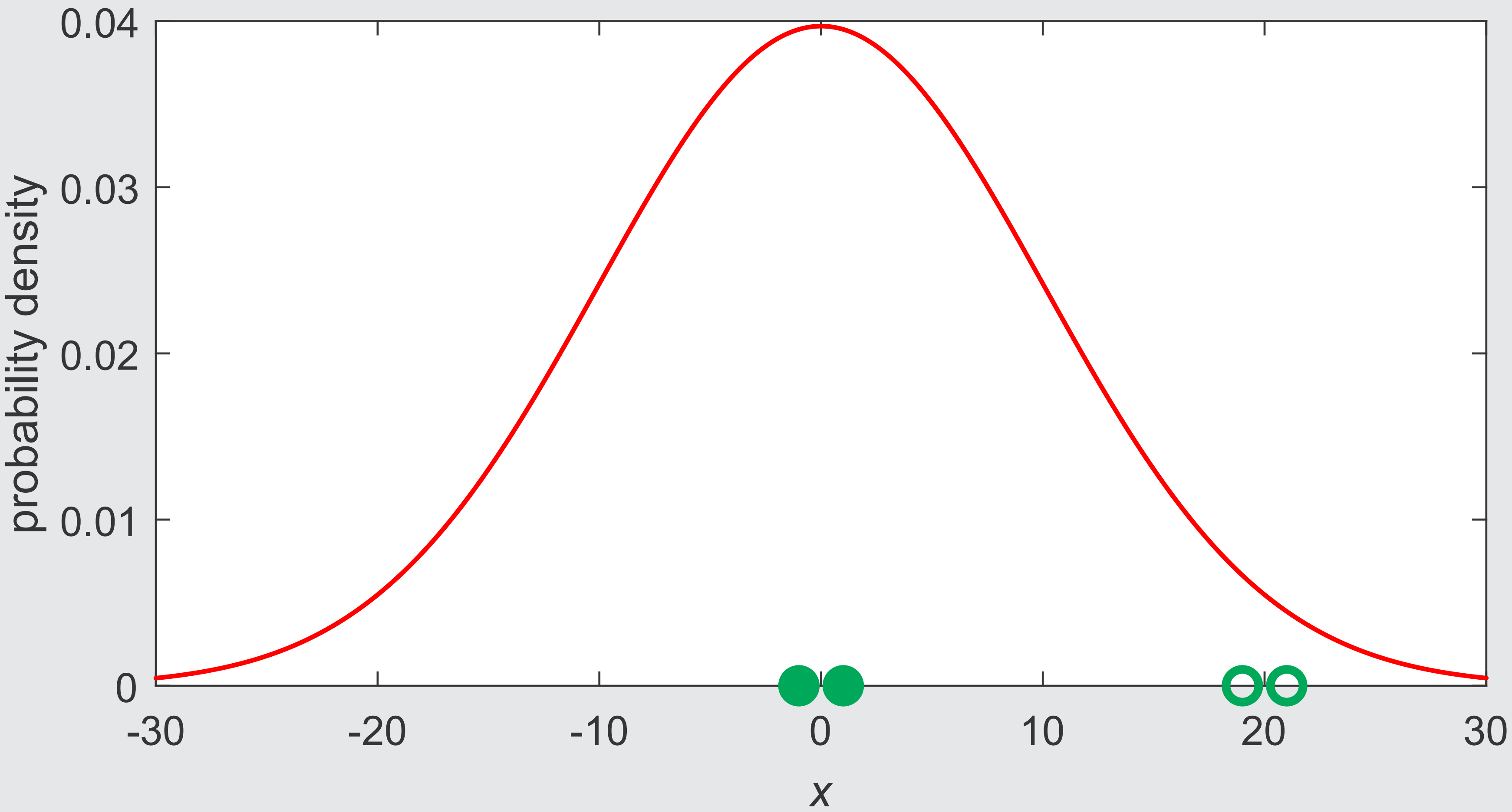
$$\Lambda = \frac{f\left(\begin{bmatrix} x_\mathrm{q} \\ x_\mathrm{k} \end{bmatrix} \middle| \begin{bmatrix} \mu_\mathrm{r} \\ \mu_\mathrm{r} \end{bmatrix}, \begin{bmatrix} \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 & \sigma_\mathrm{b}^2 \\ \sigma_\mathrm{b}^2 & \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 \end{bmatrix}\right)}{f\left(\begin{bmatrix} x_\mathrm{q} \\ x_\mathrm{k} \end{bmatrix} \middle| \begin{bmatrix} \mu_\mathrm{r} \\ \mu_\mathrm{r} \end{bmatrix}, \begin{bmatrix} \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 & 0 \\ 0 & \sigma_\mathrm{w}^2 + \sigma_\mathrm{b}^2 \end{bmatrix}\right)}$$



- $x_\mathrm{q} = -1 \quad x_\mathrm{k} = 1 \qquad \Lambda = (41 \times 10^{-4}) / (16 \times 10^{-4}) = 2.6$
- $x_\mathrm{q} = 19 \quad x_\mathrm{k} = 21 \qquad \Lambda = (56 \times 10^{-5}) / (3.0 \times 10^{-5}) = 19$

Similarity-score-based likelihood ratio

- $\mu_r = 0$
- $\sigma_b^2 = 100$
- $\sigma_w^2 = 1$



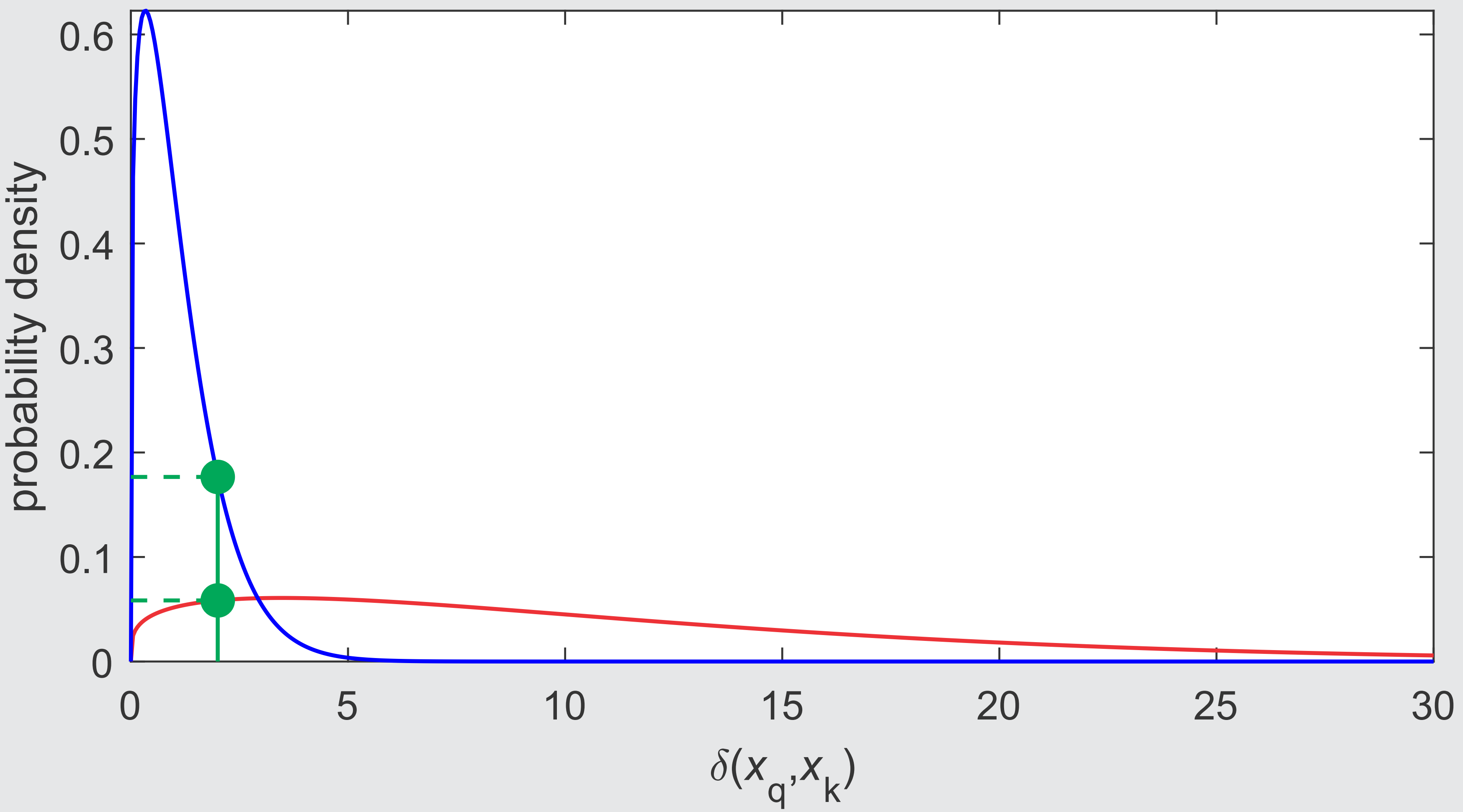
- $x_q = -1 \quad x_k = 1$
- $x_q = 19 \quad x_k = 21$

Similarity-score-based likelihood ratio

$$\Lambda = \frac{f\left(\delta\left(x_\mathrm{q}, x_\mathrm{k}\right) \middle| M_{\delta,\mathrm{s}}\right)}{f\left(\delta\left(x_\mathrm{q}, x_\mathrm{k}\right) \middle| M_{\delta,\mathrm{d}}\right)}$$

$$\delta\left(x_\mathrm{q}, x_\mathrm{k}\right) = \left|x_\mathrm{q} - x_\mathrm{k}\right|$$

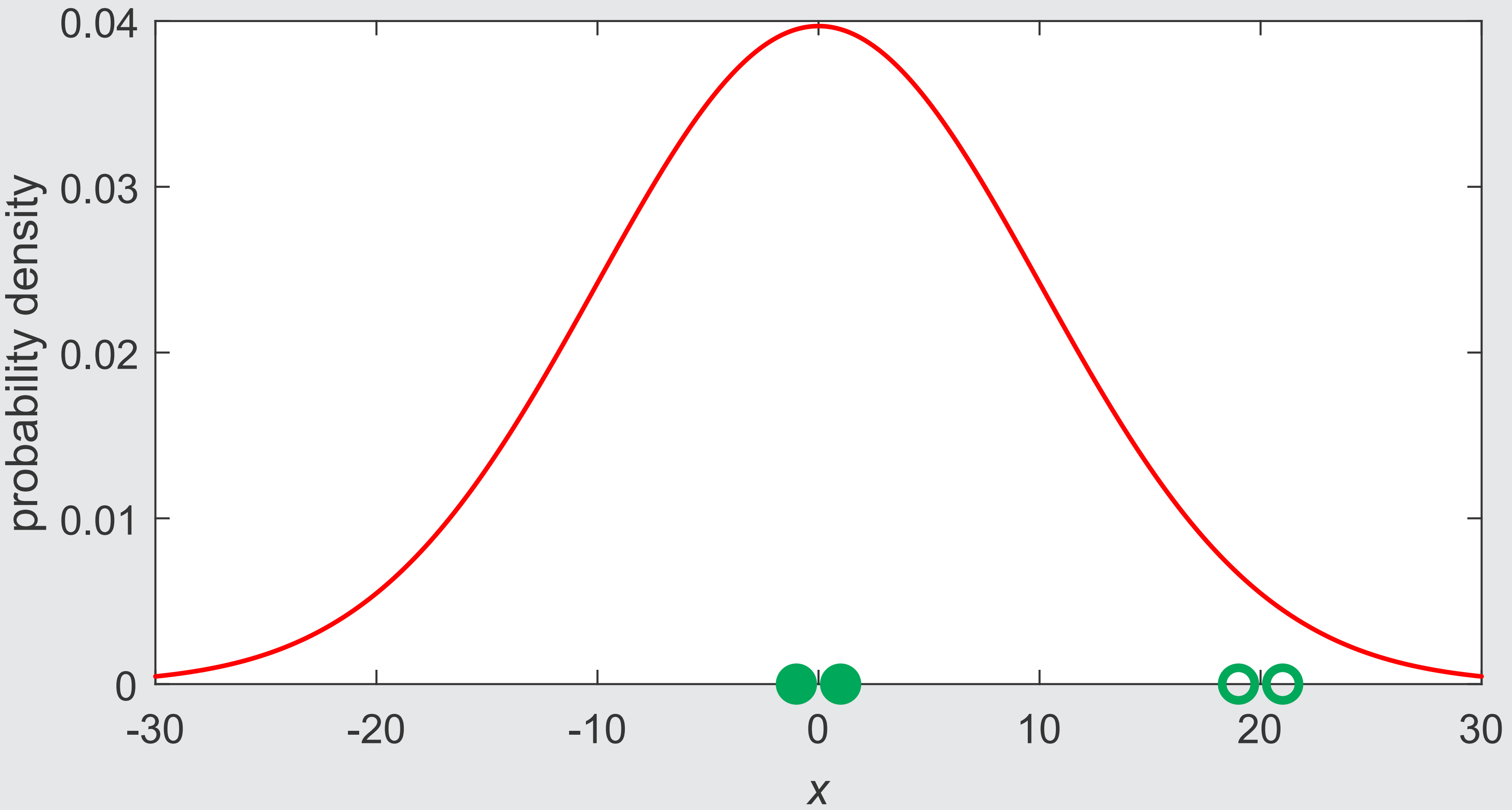
- Weibull distributions



- $x_\mathrm{q} = -1$ $x_\mathrm{k} = 1$ $\delta(-1, 1) = 2$ $\Lambda = 0.18 / 0.058 = 3.1$
- $x_\mathrm{q} = 19$ $x_\mathrm{k} = 21$ $\delta(19, 21) = 2$ $\Lambda = 0.18 / 0.058 = 3.1$

Common-source versus similarity-score-based likelihood ratio

- $\mu_r = 0$
- $\sigma_b^2 = 100$
- $\sigma_w^2 = 1$



- $x_q = -1$ $x_k = 1$ $\Lambda_{\text{common-source}} = 2.6$ $\Lambda_{\text{similarity-score}} = 3.1$
- $x_q = 19$ $x_k = 21$ $\Lambda_{\text{common-source}} = 19$ $\Lambda_{\text{similarity-score}} = 3.1$

Part II
Calibration and Validation of Likelihood-Ratio Systems

Contents – Part II

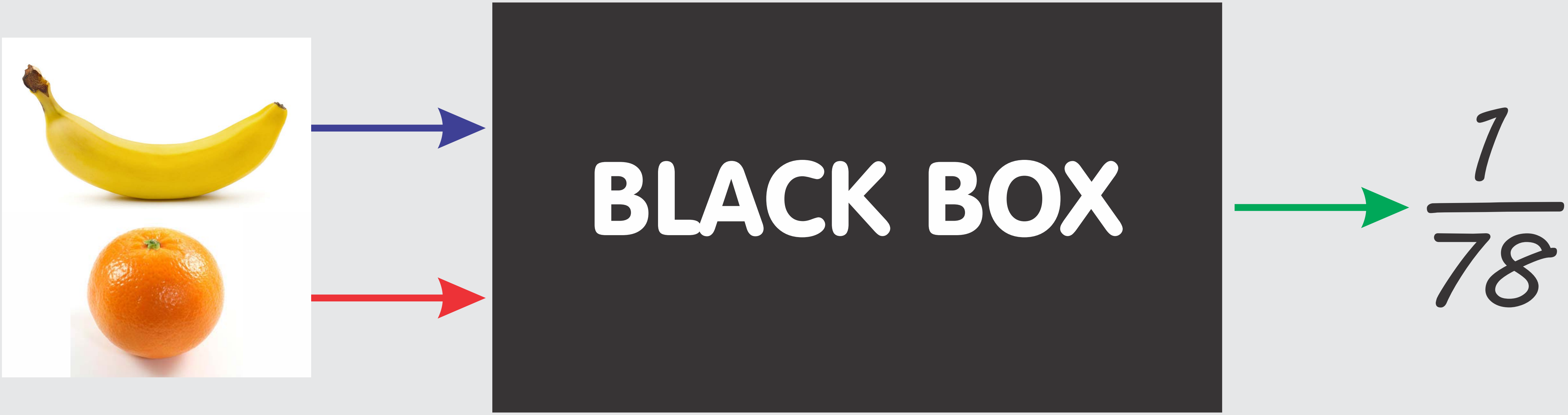
- Preliminaries
  - Black boxes
  - Logarithms
- Calibration
  - Calibration in weather forecasting
  - Calibration principles
  - Well-calibrated likelihood ratios
  - Calibration models
- Validation
  - Validation protocols
  - Validation metric (log-likelihood-ratio cost, C_{llr})
  - Validation graphic (Tippett plot)
- Consensus on Validation
  - Key points

Preliminaries

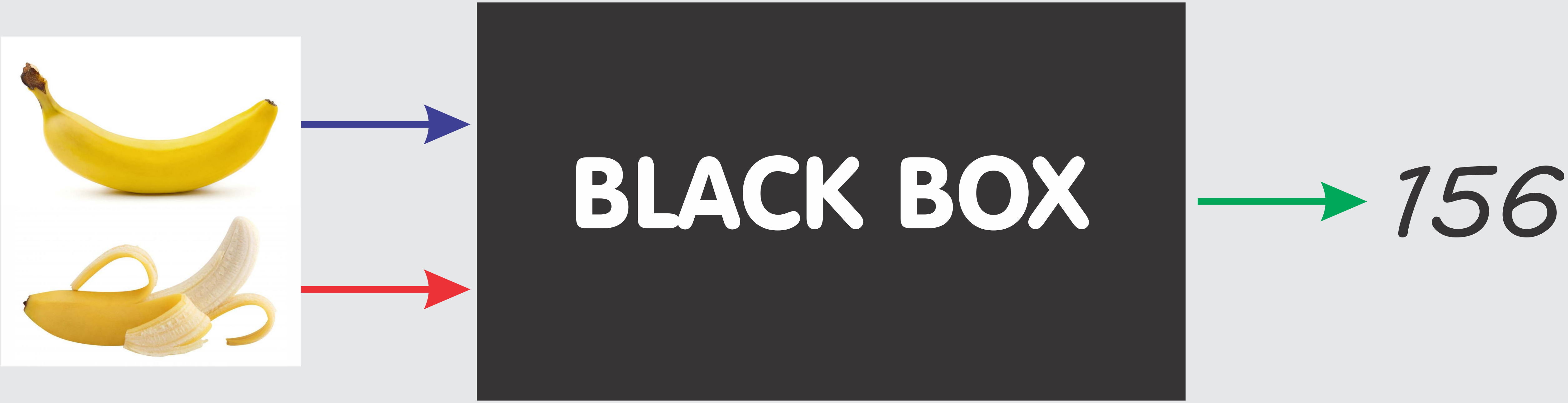
Preliminaries – black boxes

- 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

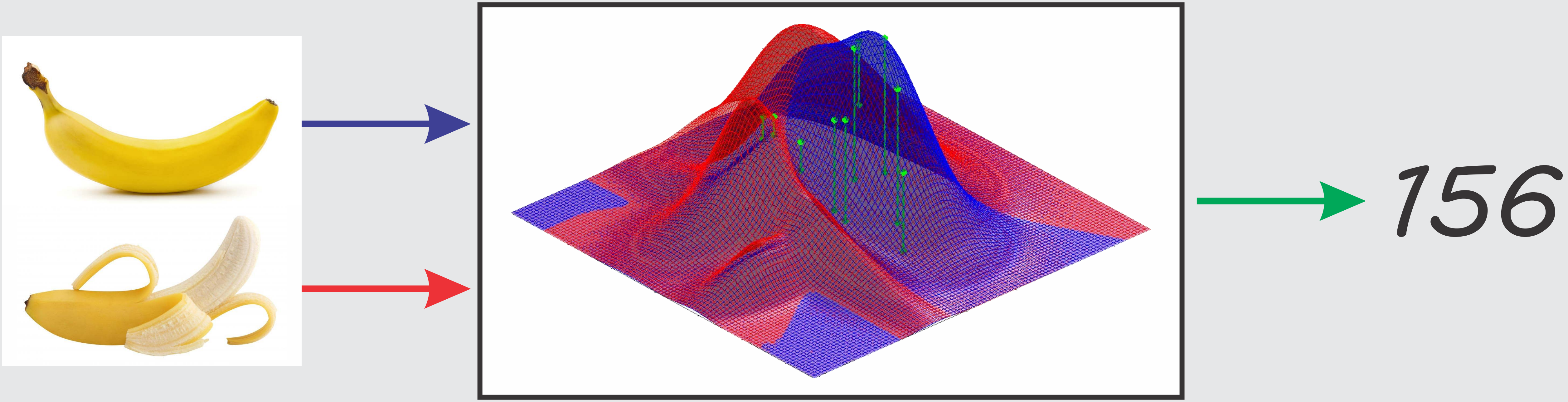
Preliminaries – black boxes



Preliminaries – black boxes



Preliminaries – black boxes



Preliminaries – black boxes

Preliminaries – black boxes

156

Preliminaries – 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**
10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}	10^{3}
$\log_{10}(LR)$						
−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**
$\mathbf{2^{-3}}$	$\mathbf{2^{-2}}$	$\mathbf{2^{-1}}$	$\mathbf{2^{0}}$	$\mathbf{2^{1}}$	$\mathbf{2^{2}}$	$\mathbf{2^{3}}$

$\log_2(LR)$

−3	**−2**	**−1**	**0**	**+1**	**+2**	**+3**

Preliminaries – logarithms

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

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

Calibration

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%

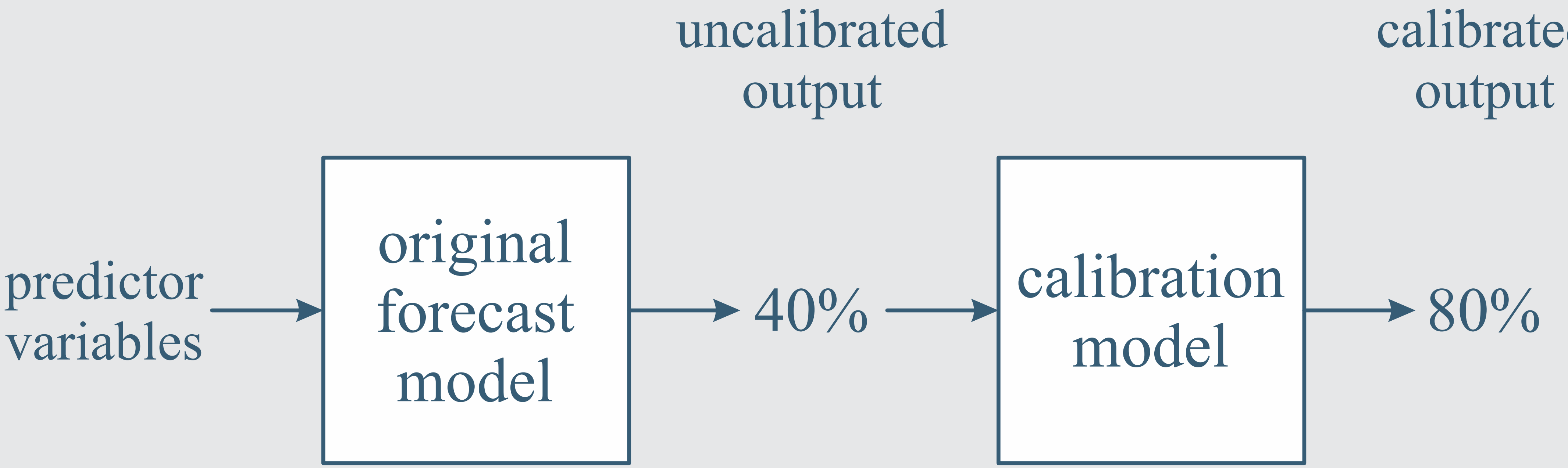
Not well calibrated:

- Prediction: 40%
- Actual: 80%

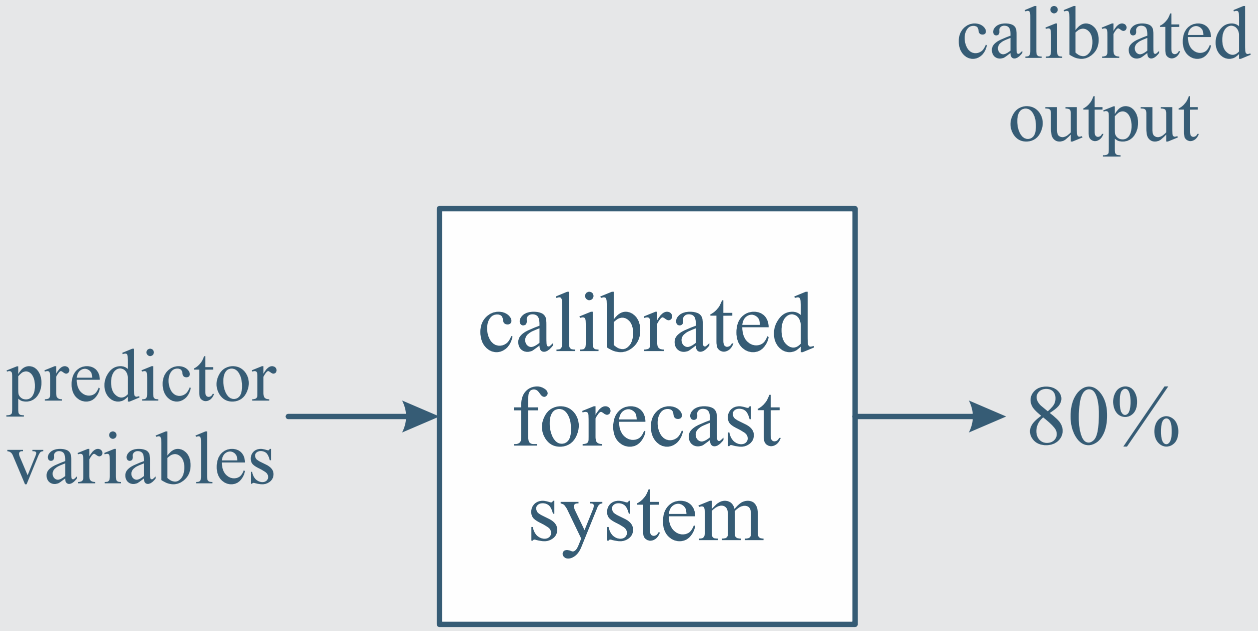
Calibration in weather forecasting

- 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

- 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

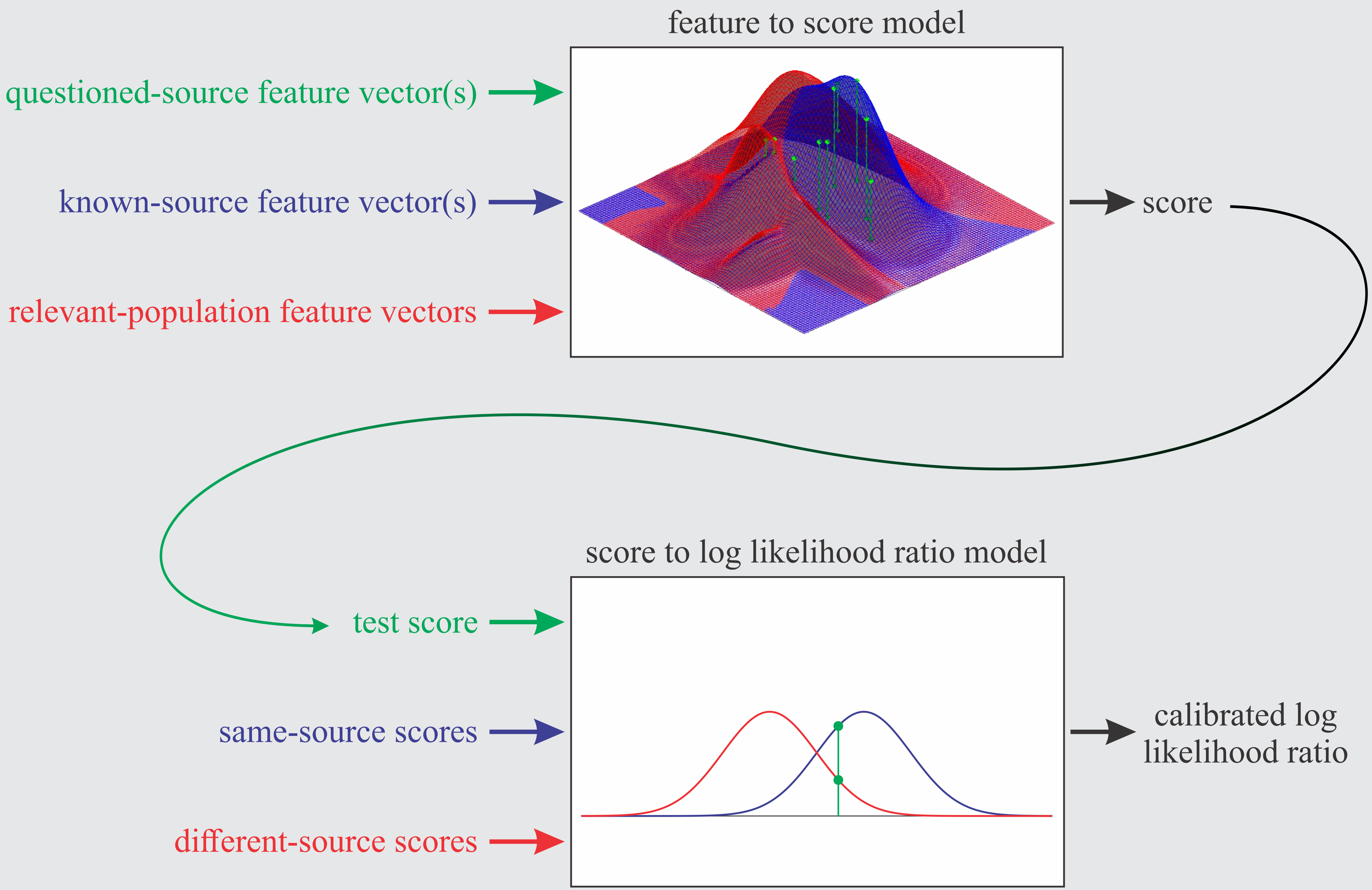
- Solution:
  - Treat the output of the first (complex) model as an uncalibrated log likelihood ratio (a score)
  - Use a parsimonious model to convert the score to a calibrated log likelihood ratio

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

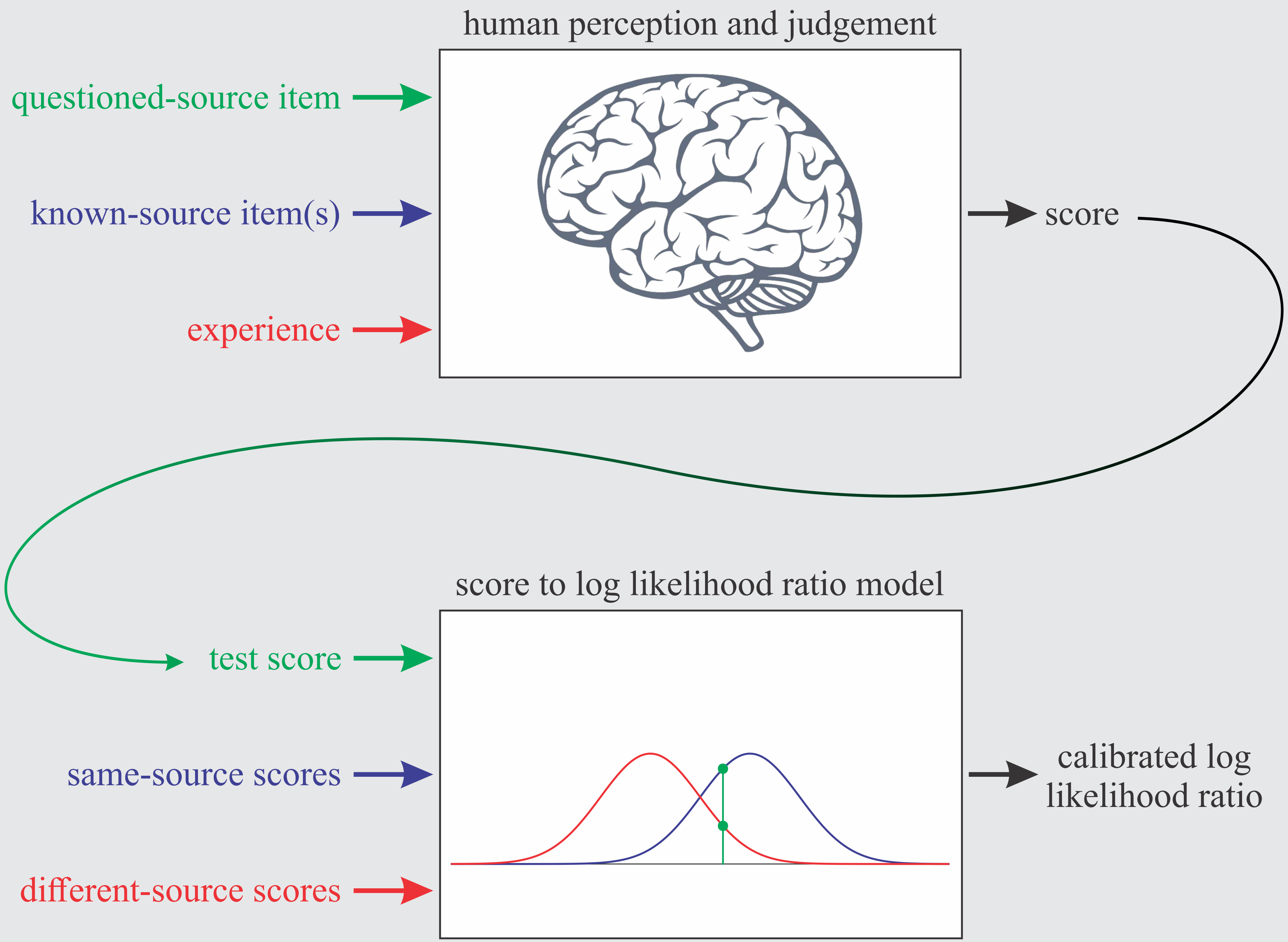
- Important condition:

  - 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



Well-calibrated likelihood ratios

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

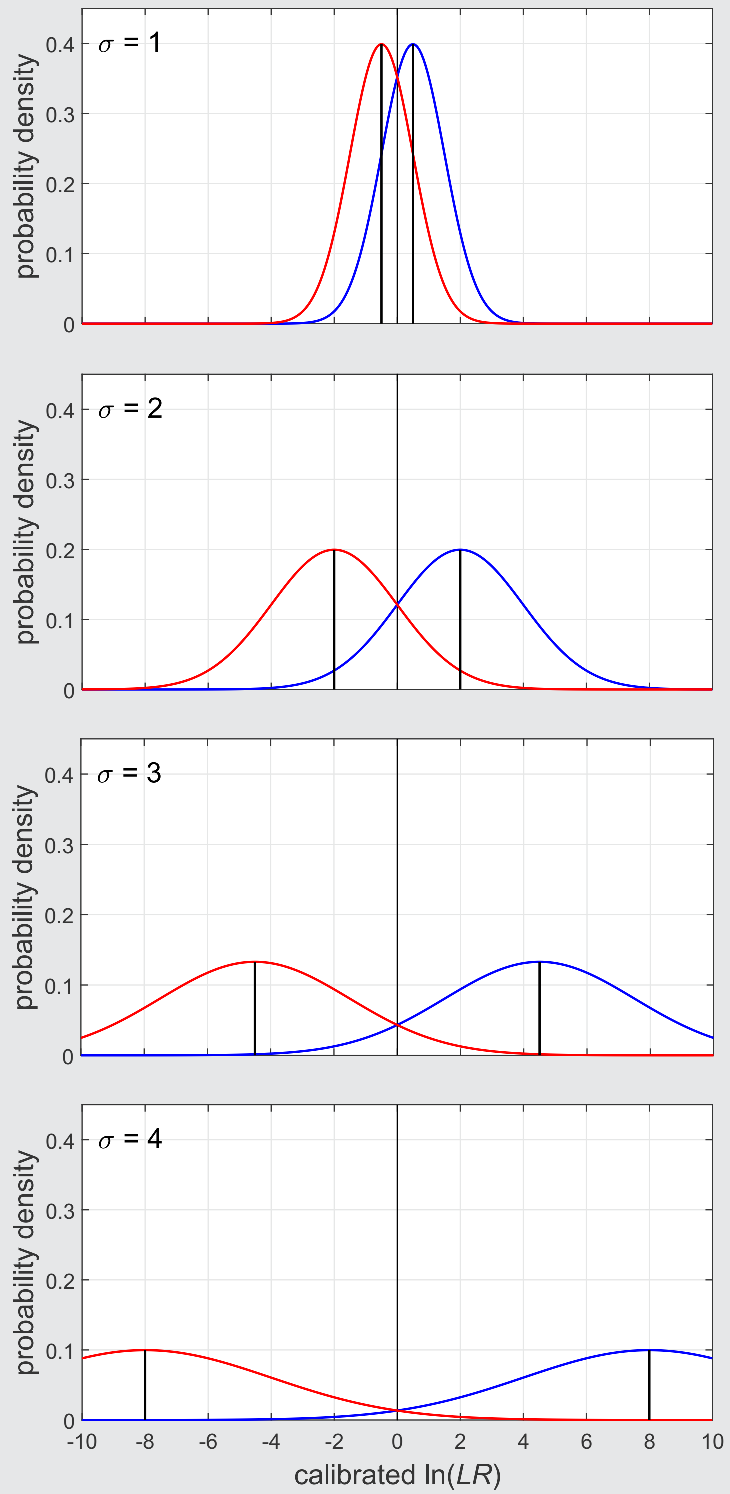
$$LR = \frac{f(\,LR \mid H_{\mathrm{s}}\,)}{f(\,LR \mid H_{\mathrm{d}}\,)}$$

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

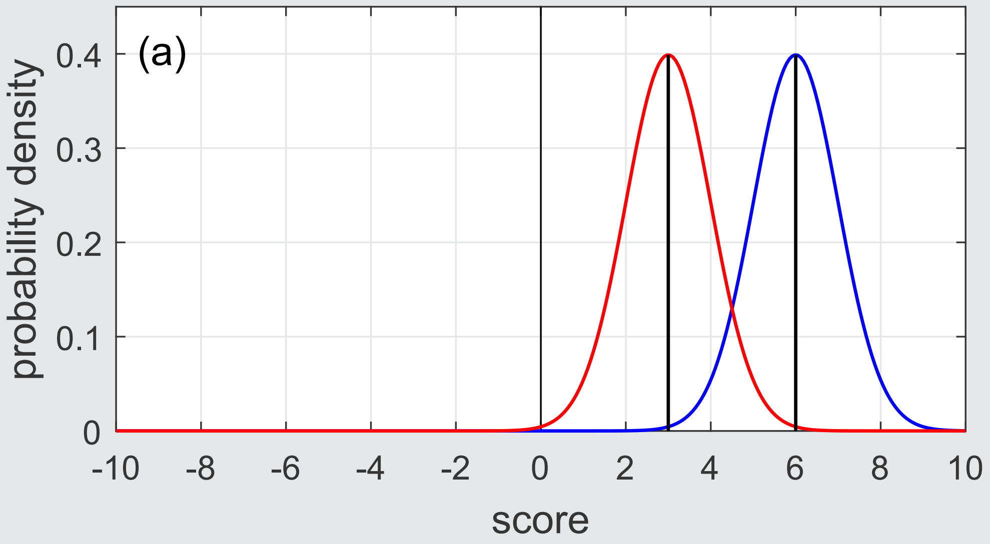
(a)

Uncalibrated scores

$\mu_{\mathrm{d}} = 3$

$\mu_{\mathrm{s}} = 6$

$\sigma = 1$



Calibration models

(a)

Uncalibrated scores

$\mu_d = 3$

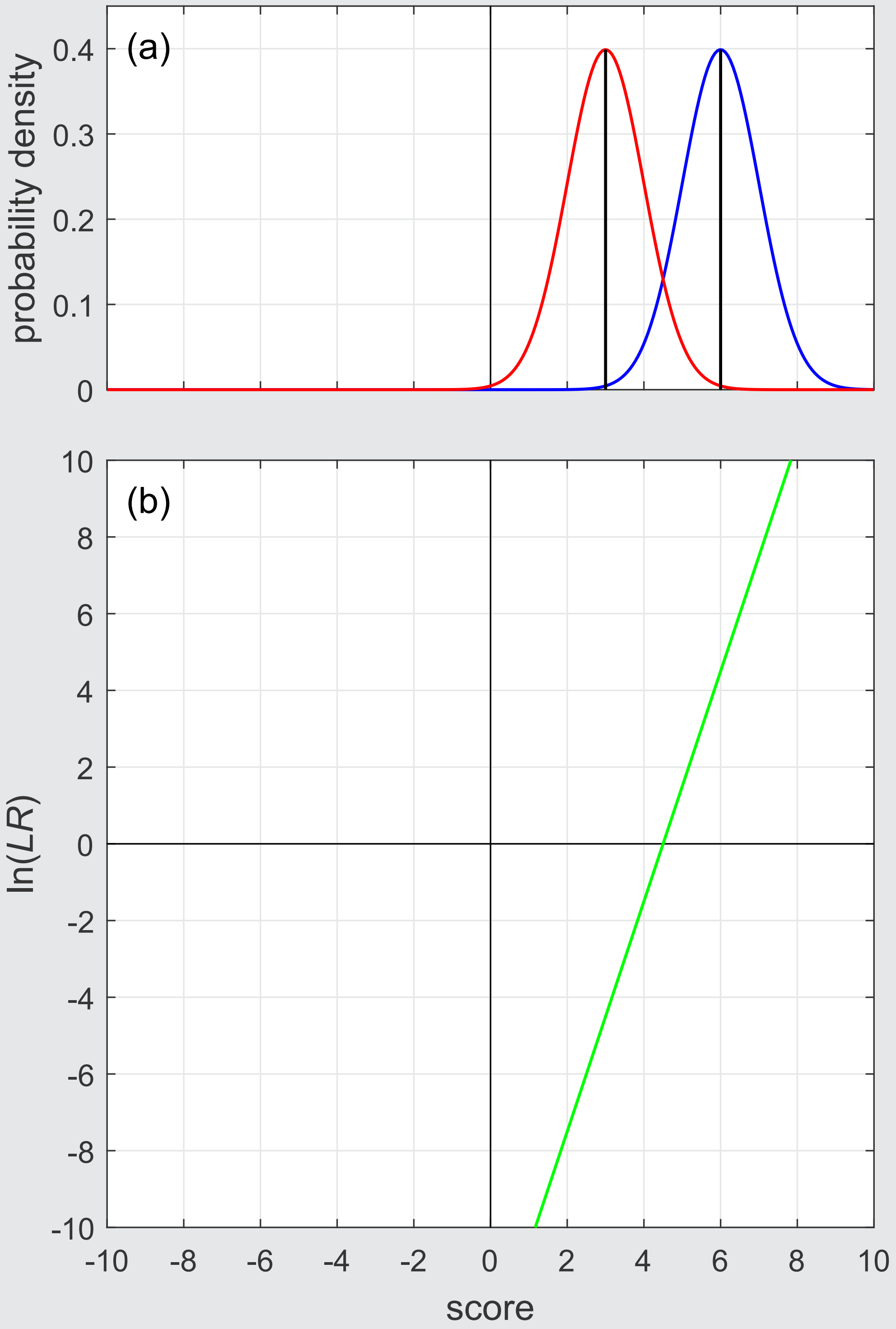
$\mu_s = 6$

$\sigma = 1$

(b)

Score to ln(*LR*)

mapping function



Calibration models

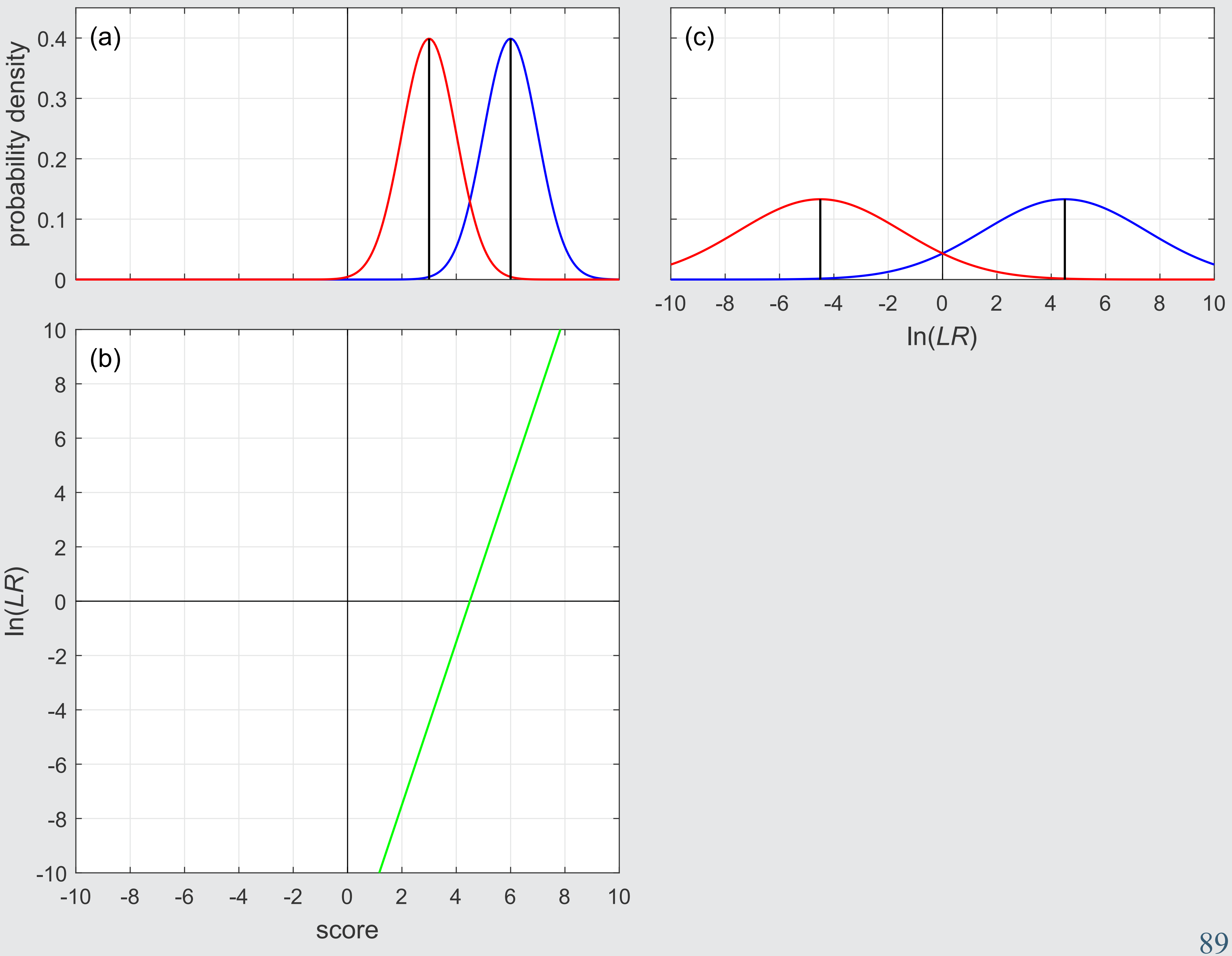
(c)

Calibrated ln(LR)

$\mu_d = -4.5$

$\mu_s = +4.5$

$\sigma = 3$



Calibration models

(c)

Calibrated ln(*LR*)

$\mu_d = -4.5$

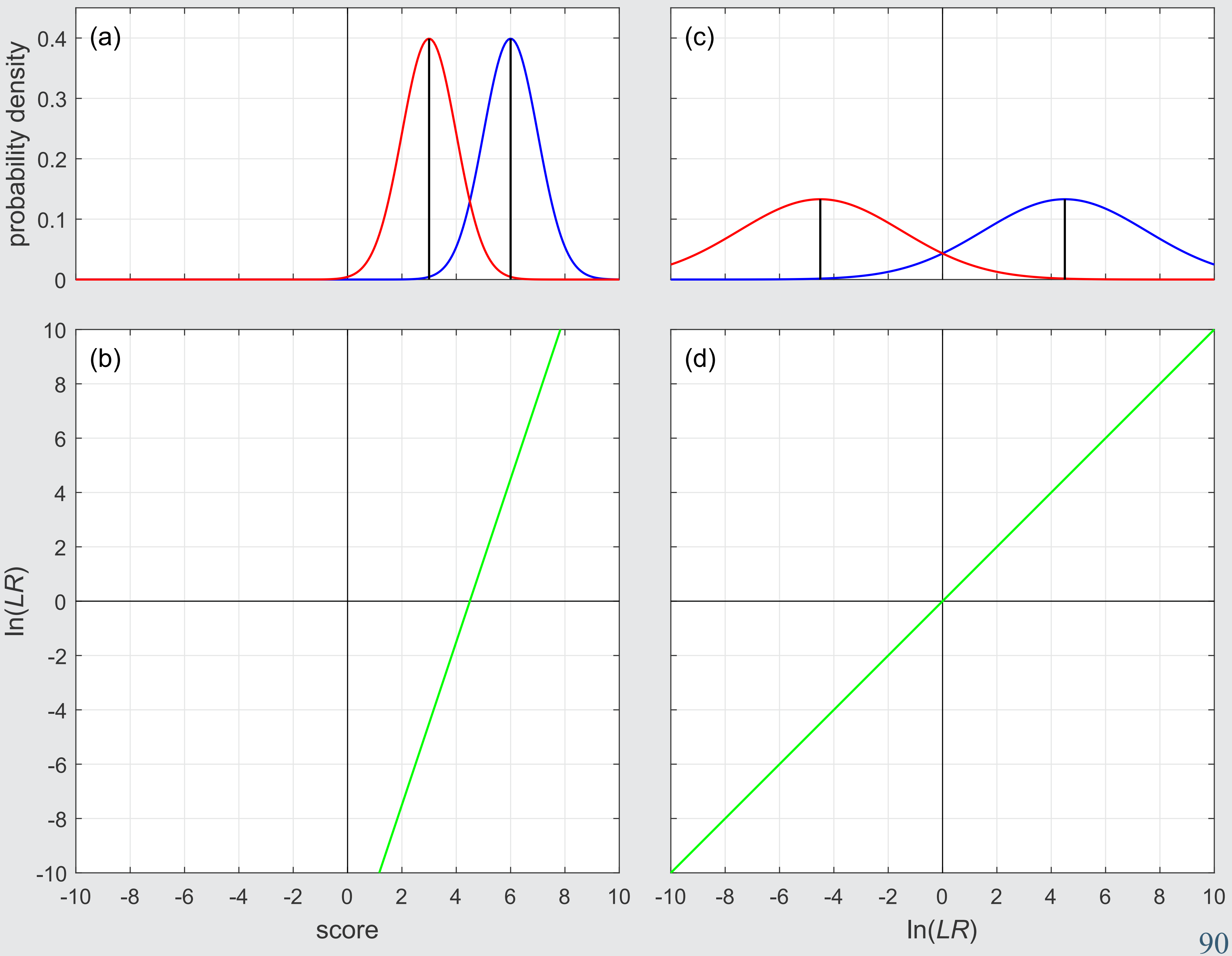
$\mu_s = +4.5$

$\sigma = 3$

(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

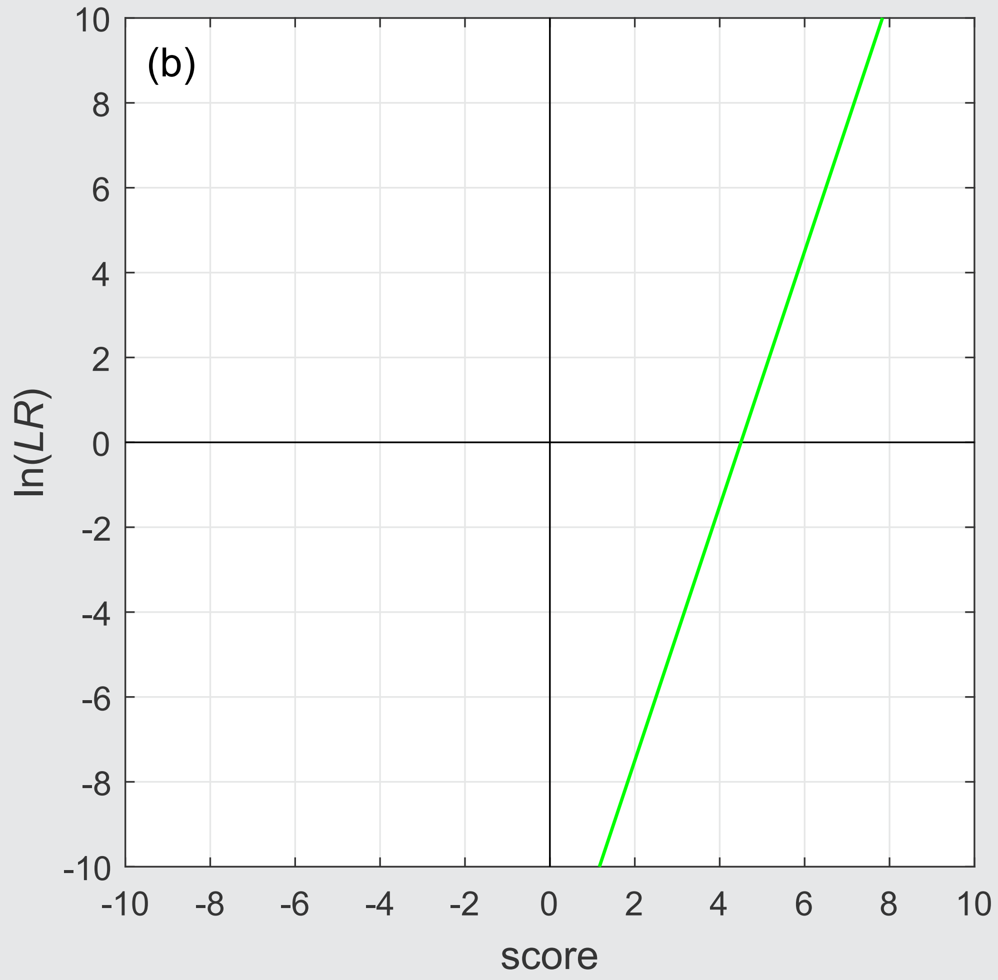
- Score [x] to ln(LR) [y] mapping function:

$$y = a + bx$$

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

$$a = -b\frac{6 + 3}{2} \qquad b = \frac{6 - 3}{1^2}$$

$$a = -3 \times 4.5 \qquad b = 3$$



Calibration models

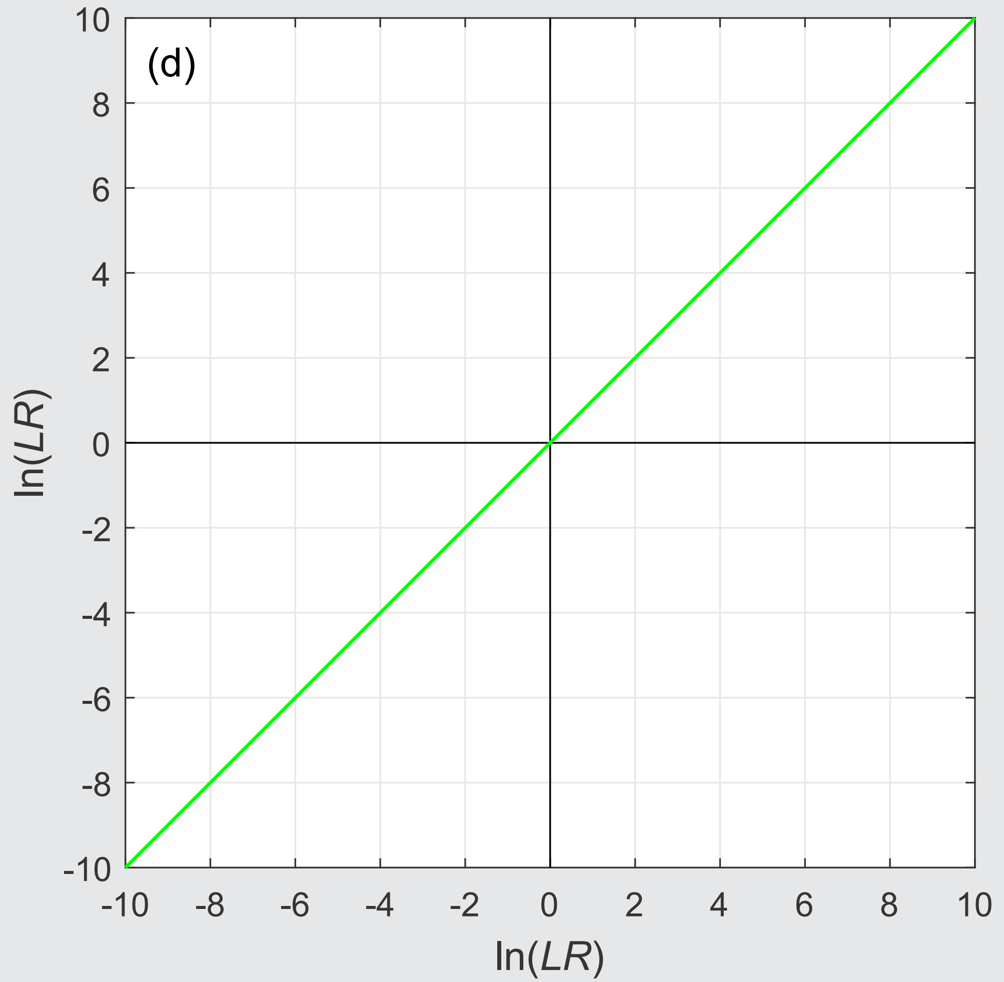
- ln(*LR*) [*x*] to ln(*LR*) [*y*] mapping function:

$$y = a + bx$$

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

$$a = -b\frac{4.5 + (-4.5)}{2} \qquad b = \frac{4.5 - (-4.5)}{3^2}$$

$$a = 0 \qquad b = 1$$



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

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

- Classification-error rate

output

input		same source	different source
	same source	correct	incorrect
	different source	incorrect	correct

Validation metric

- Classification-error rate

  - names

		output	
		same source	different source
input	same source	hit	miss
	different source	false alarm	correct rejection

Validation metric

- Classification-error rate
    - penalty values

		output: same source	output: different source
input	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: $y_i = \text{d}$

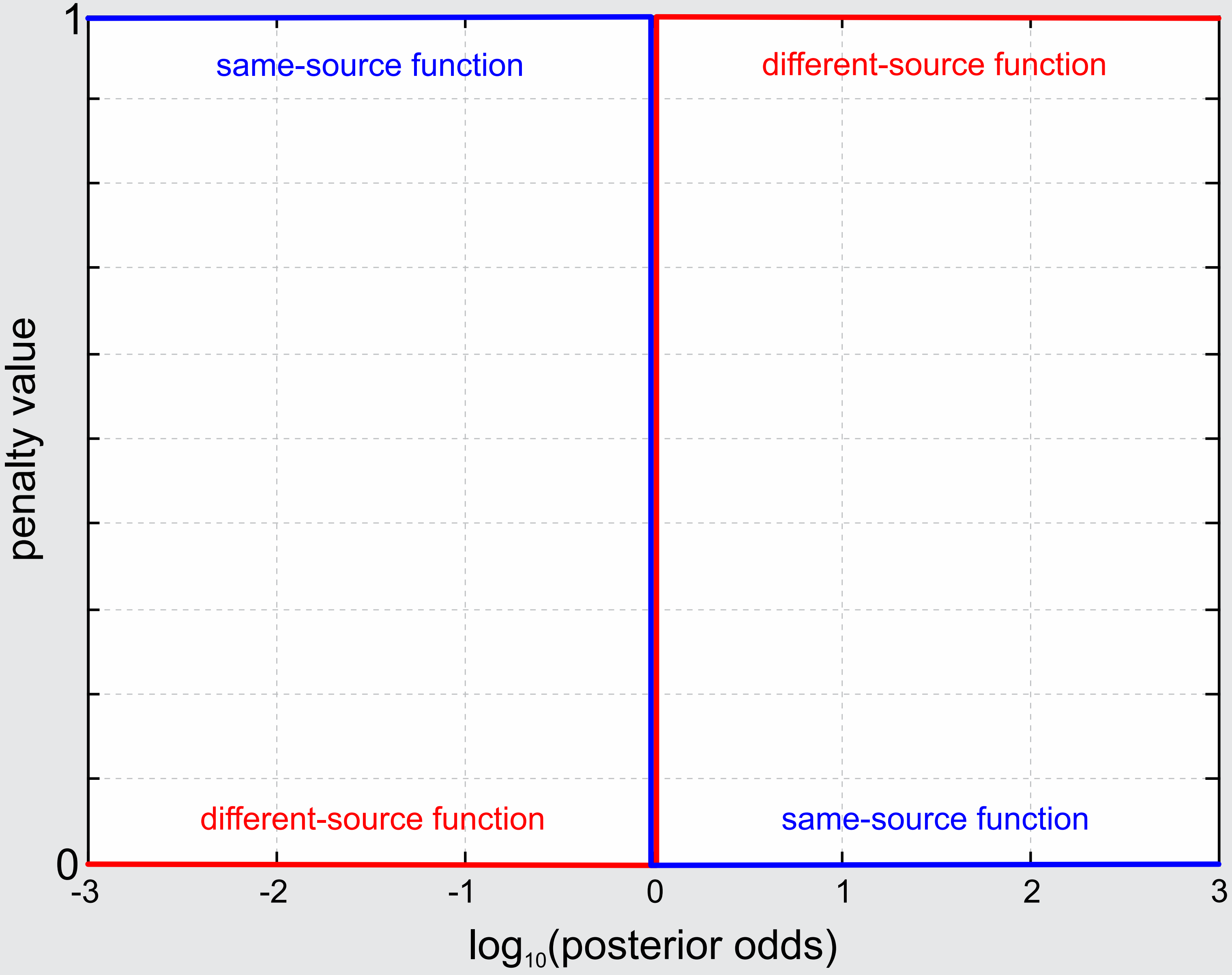
false alarm: $y_j = \text{s}$

Validation metric

- 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

- Penalty functions for calculating classification-error rate

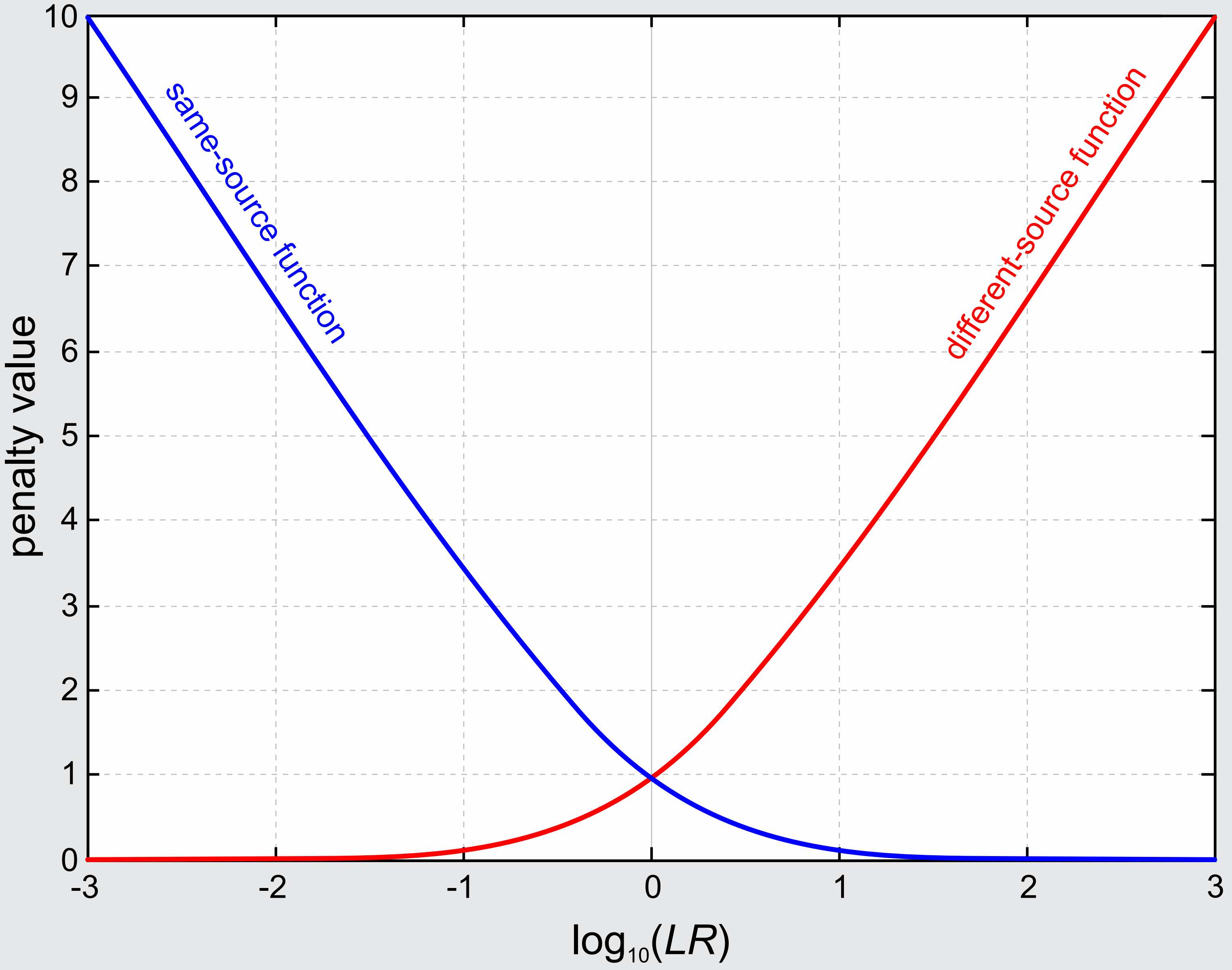


Validation metric

- 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}

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

Validation metric

- The **better the performance** of the system, the **smaller the C_{llr} value**
    - $C_{\mathrm{llr}} > 0$
    - 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 $C_{\mathrm{llr}} = 1$

Validation metric

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

Validation metric

- Perfectly calibrated ln(*LR*) distributions

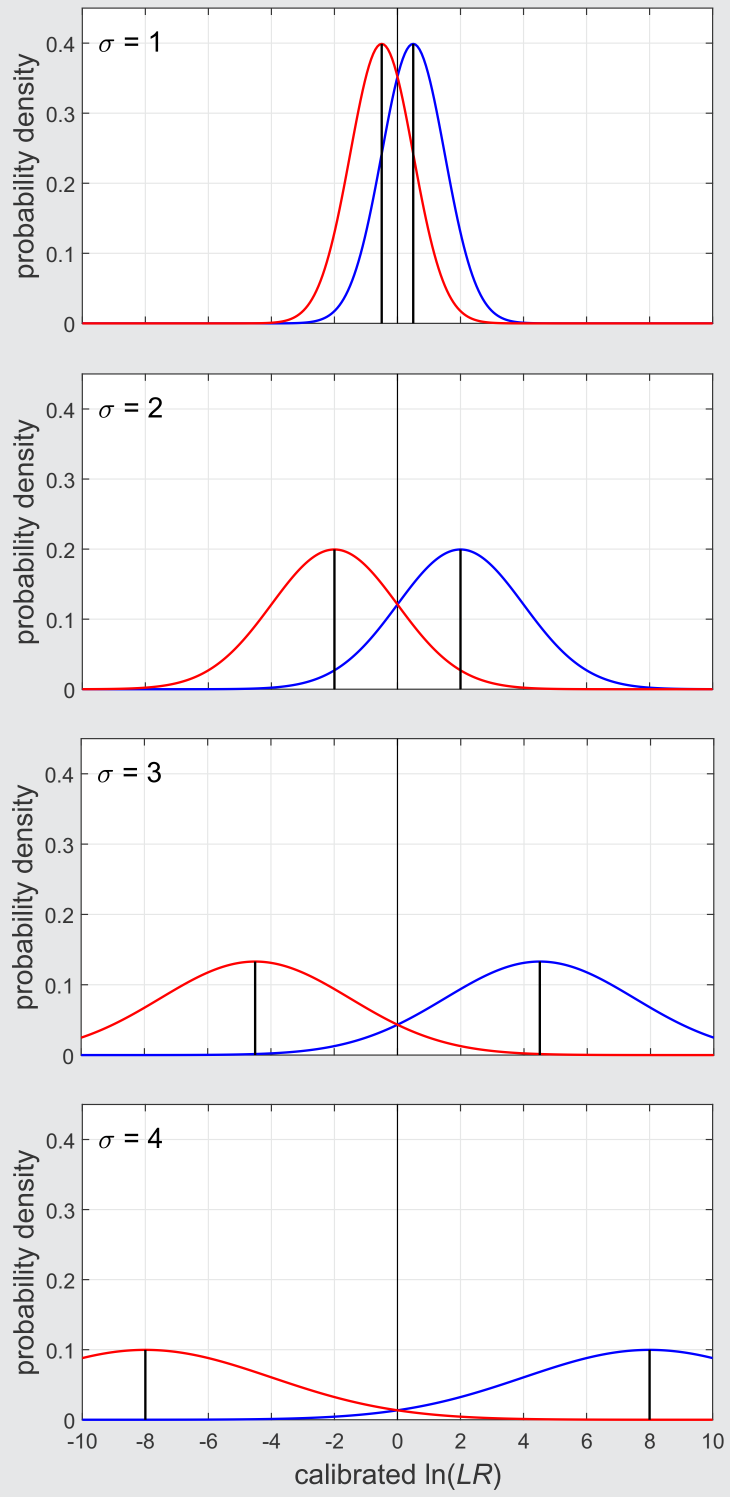
    - C_{llr} values

0.84

0.51

0.24

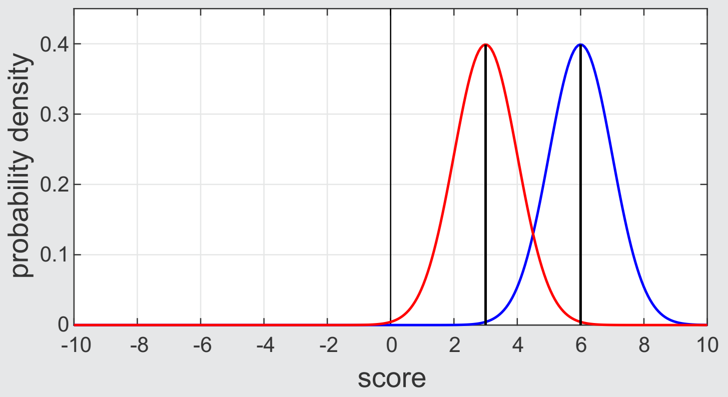
0.09



Validation metric

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

5.2

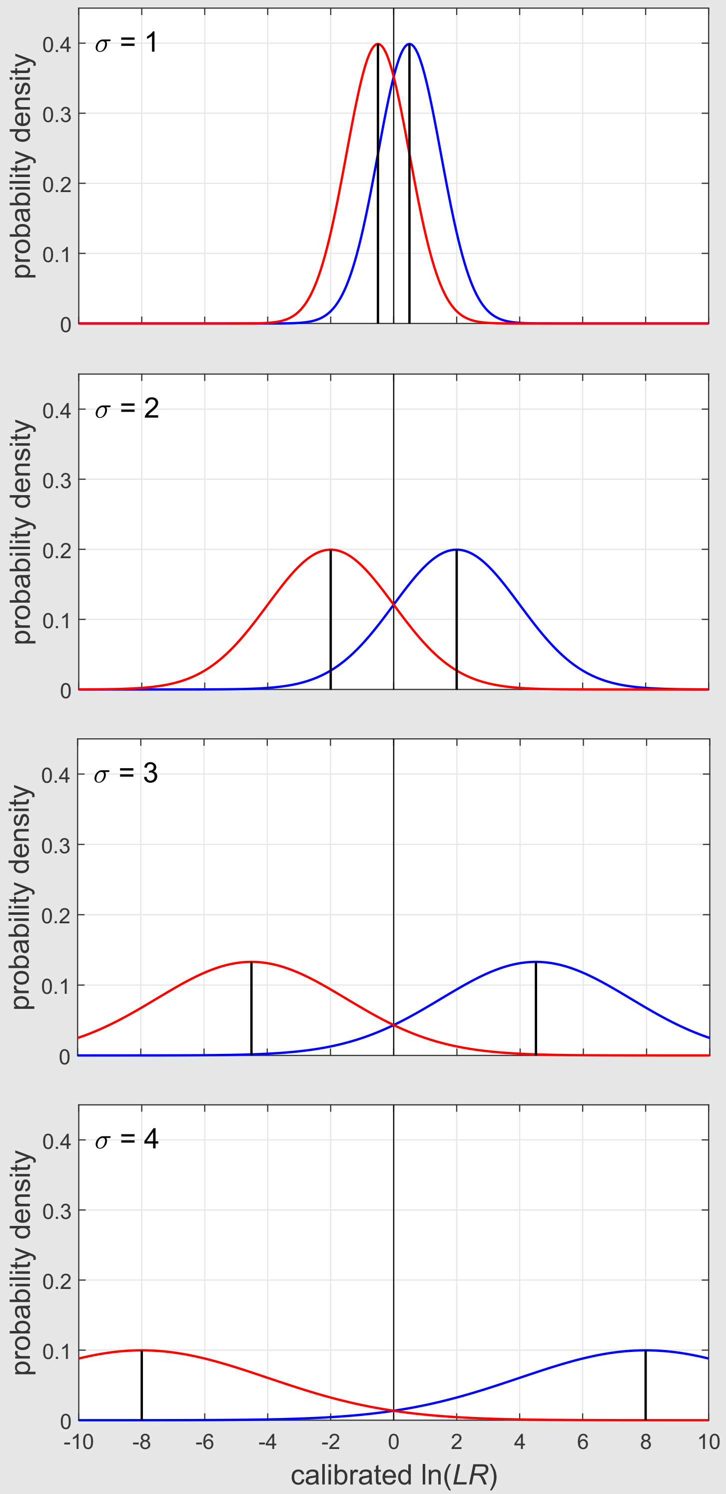


0.84

0.51

0.24

0.09



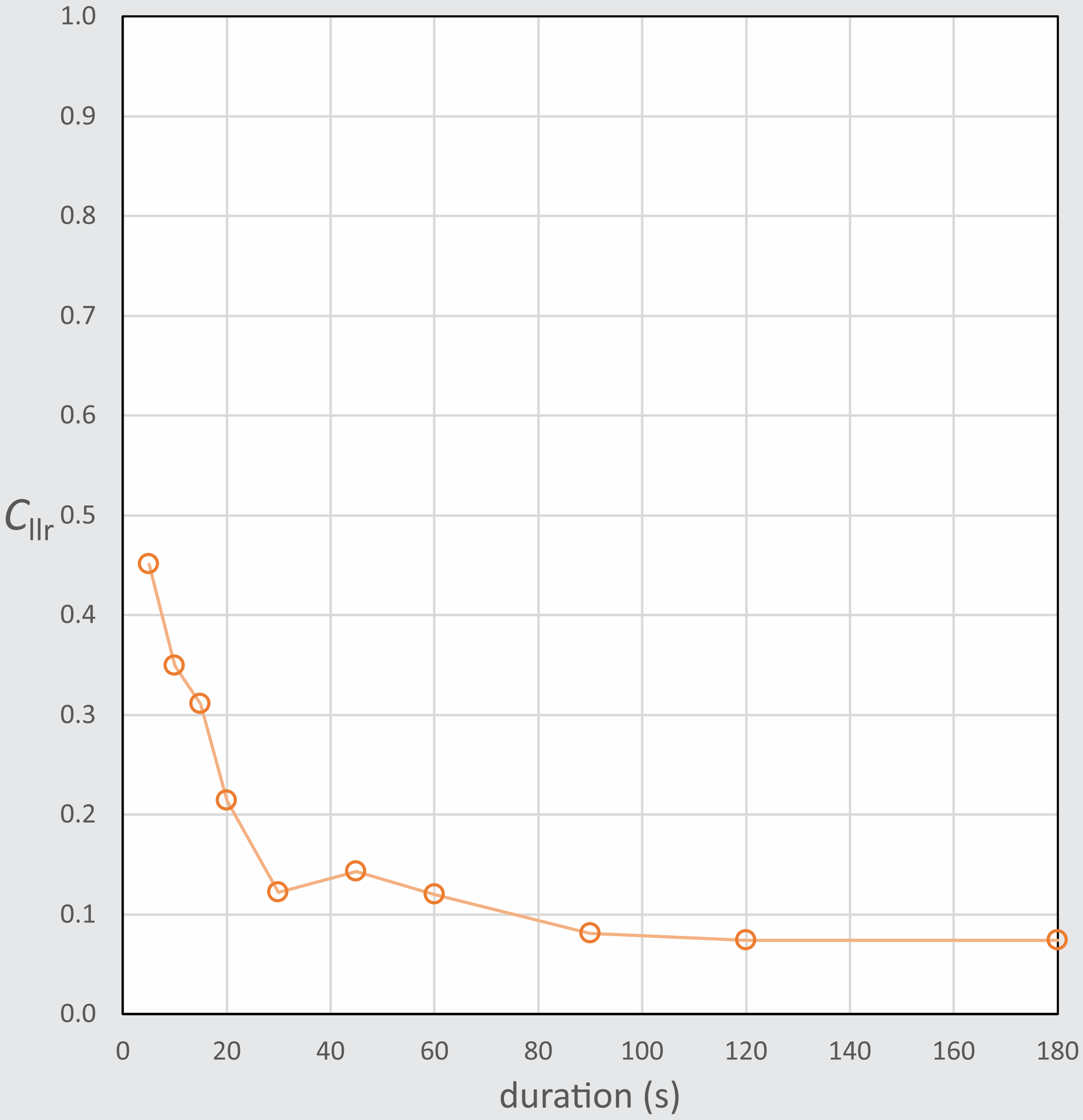
Validation metric

- 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 plot

- 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

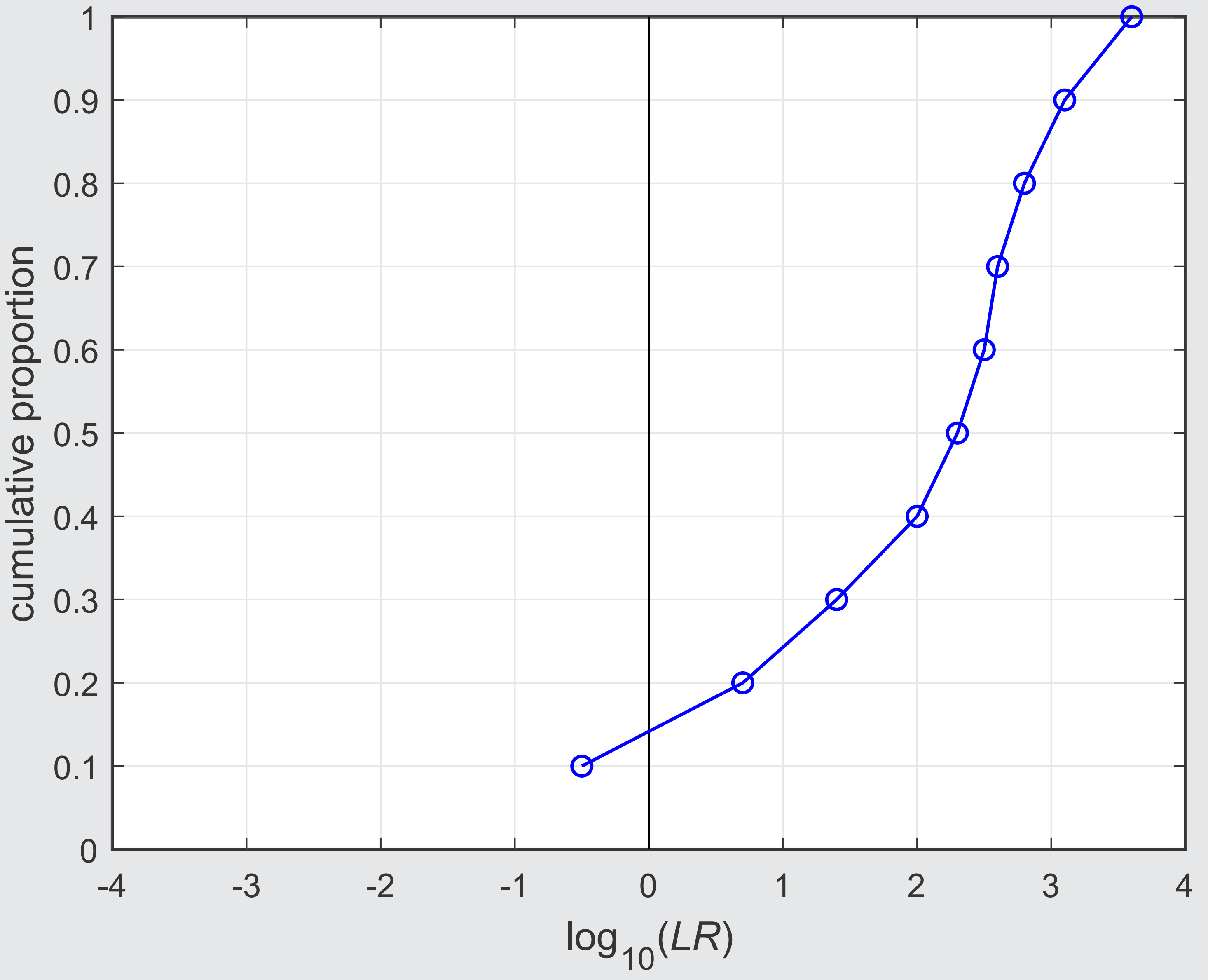
Validation plot

- **Tippett plot**:
  - rank the log(LR) values resulting from same-source pairs from smallest to largest
  - plot the proportion of values that are $\leq$ 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 plot

- **Tippett plot**:



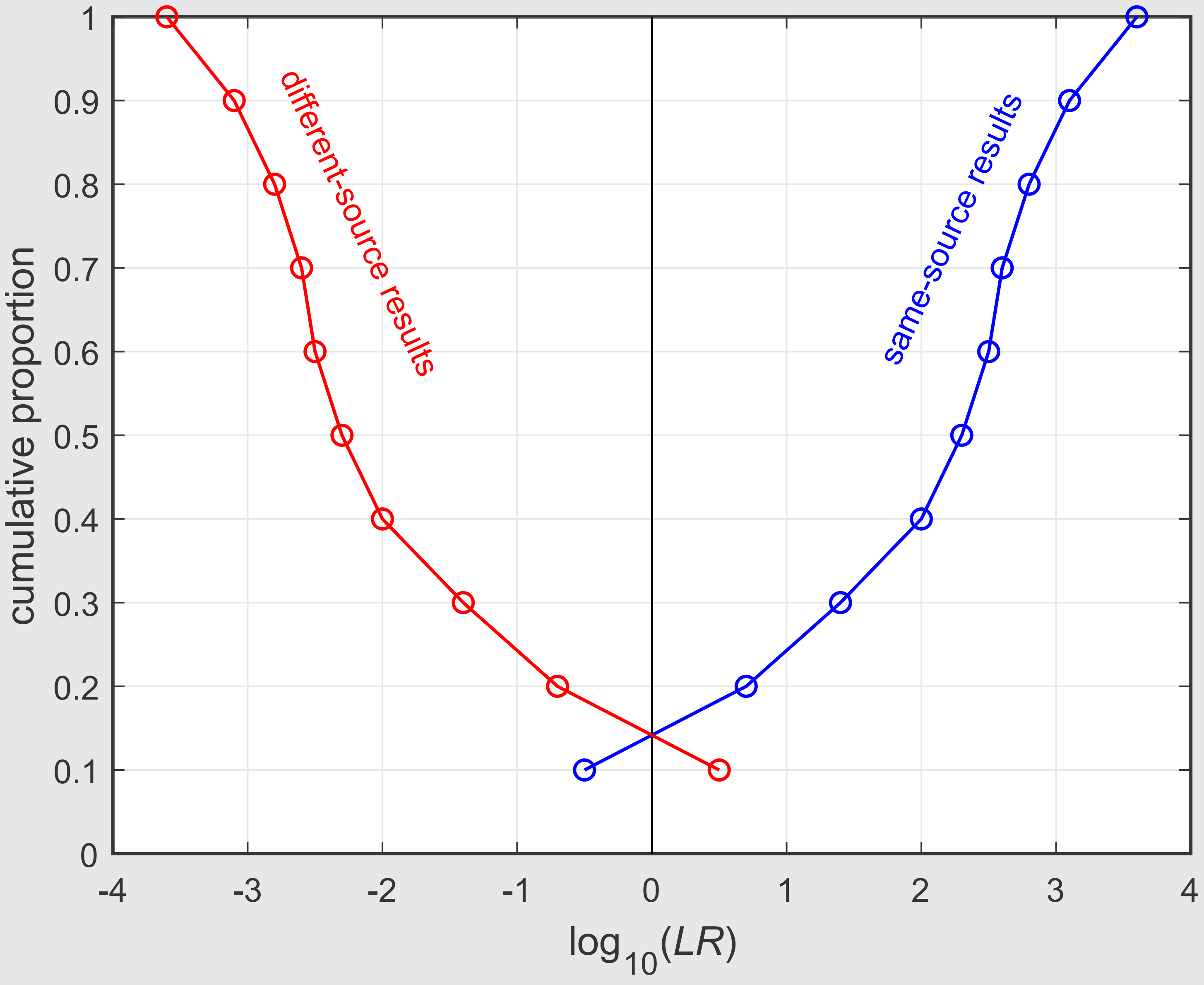
Validation plot

- **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 plot

- **Tippett plot:**



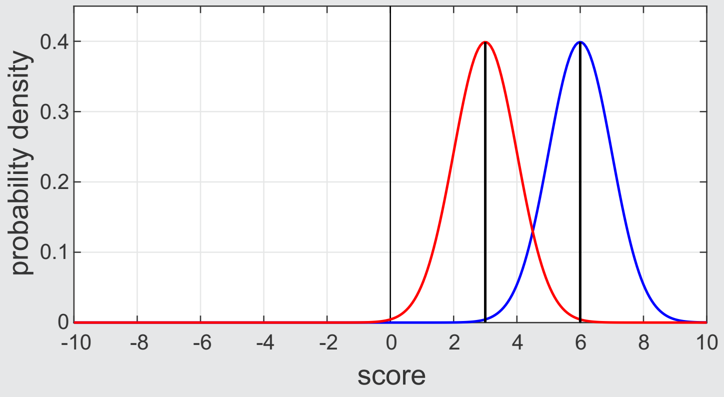
Validation plot

- 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 plot

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

5.2

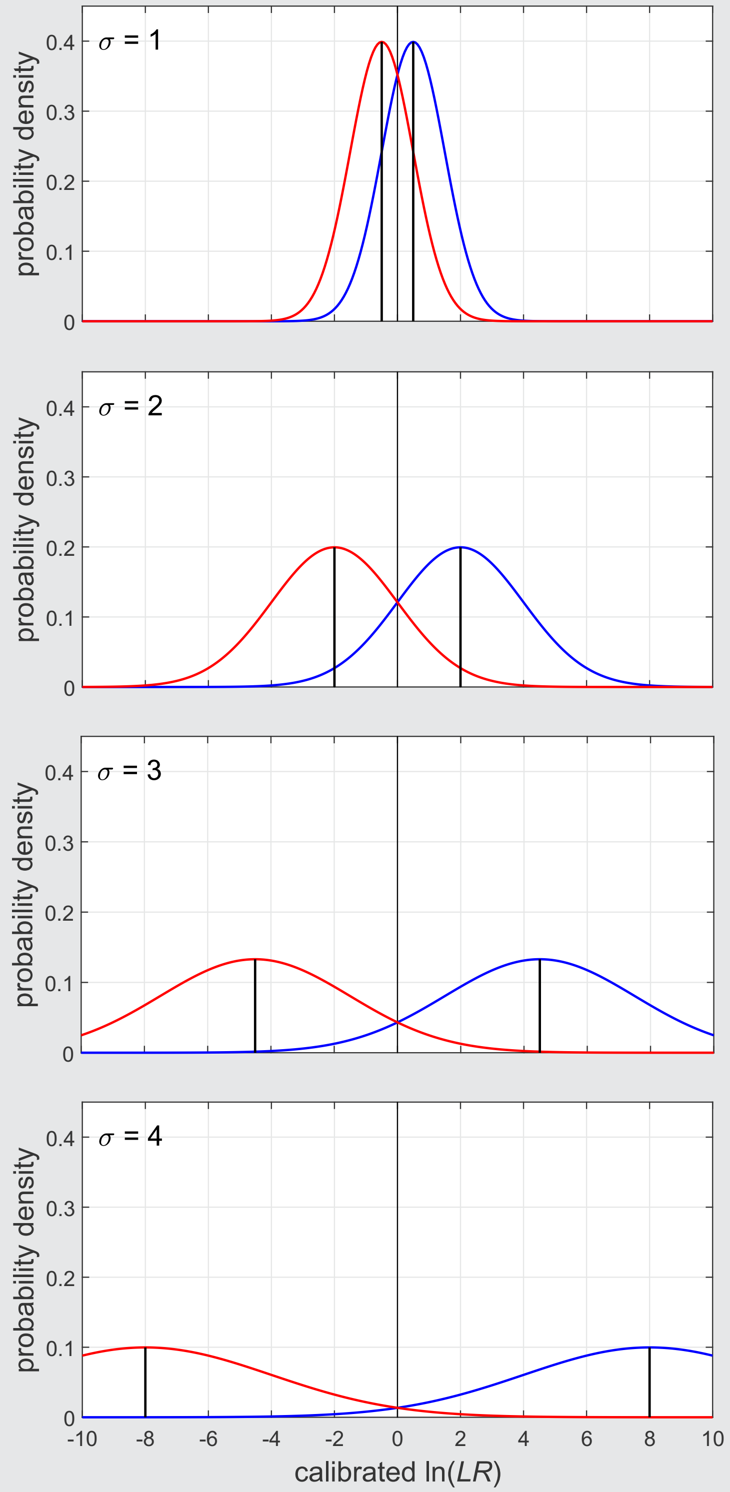


0.84

0.51

0.24

0.09

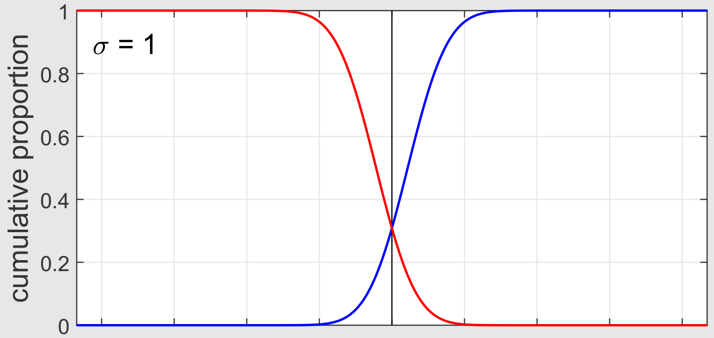


Validation plot

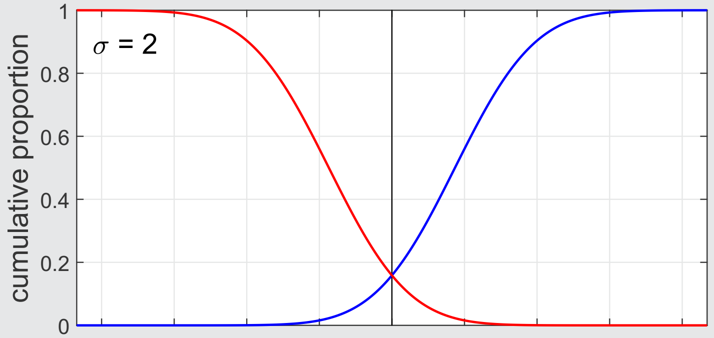
- Tippett plots

    - C_{llr} values

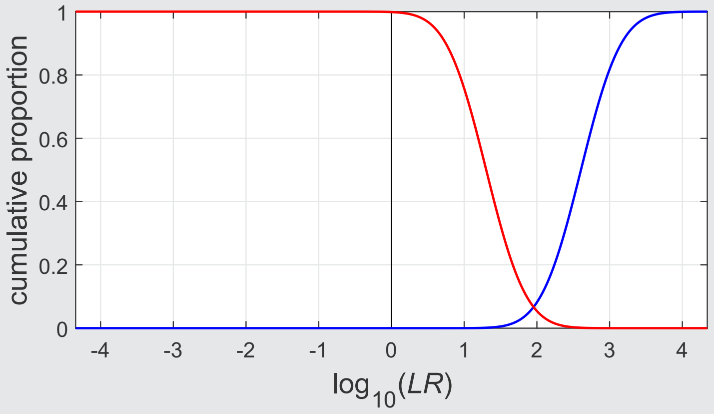
0.84



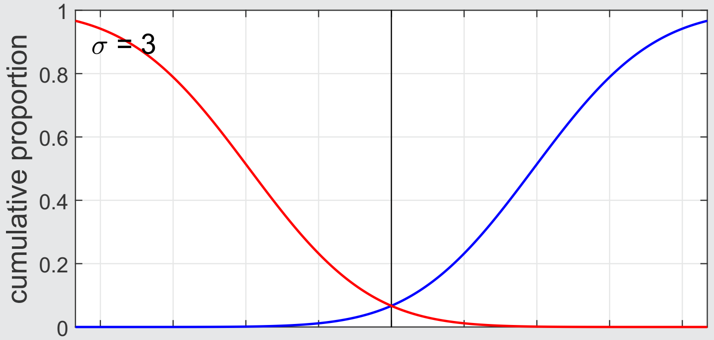
0.51



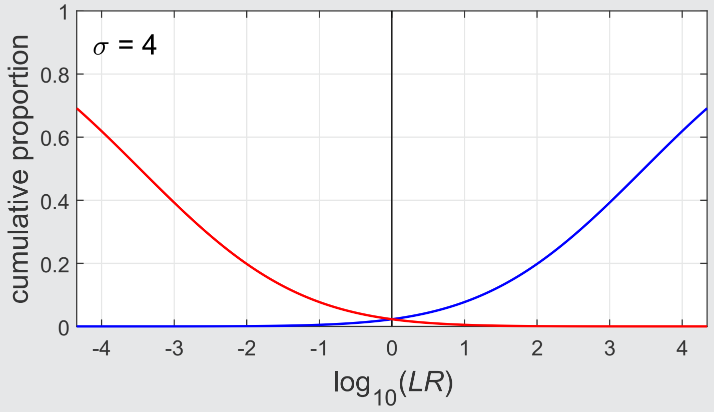
5.2



0.24



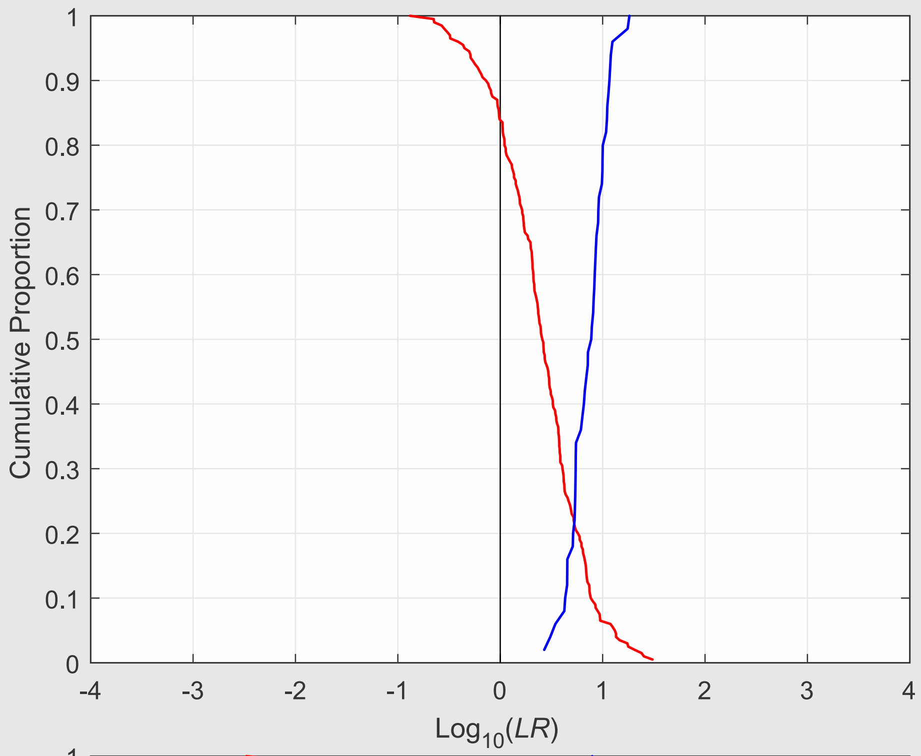
0.09



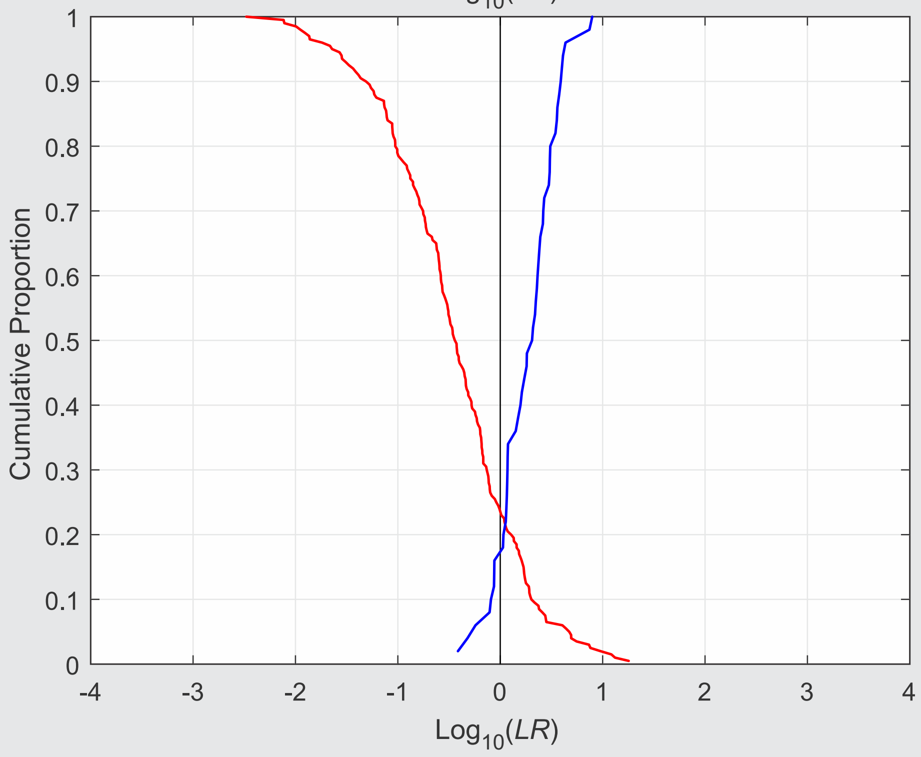
Validation plot

- Example Tippett plots
    - C_{llr} values

1.07



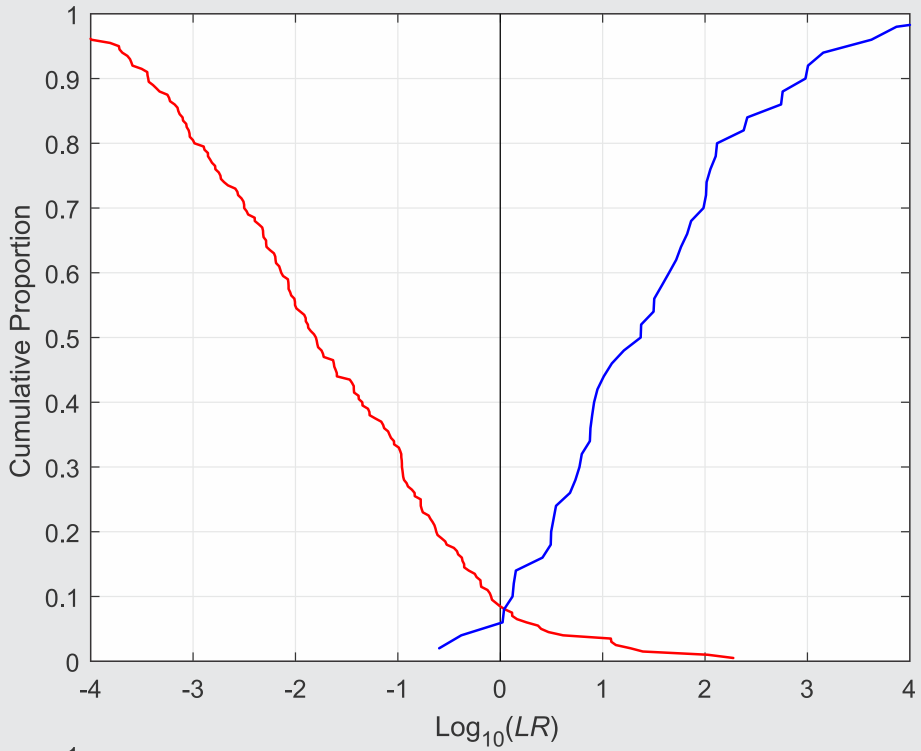
0.70



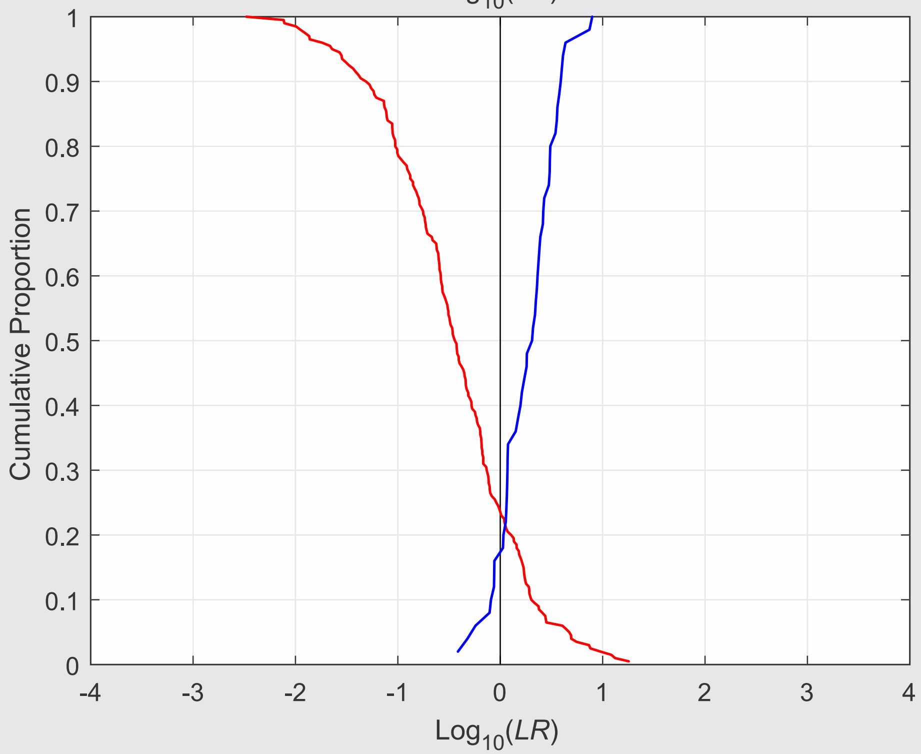
Validation plot

- Example Tippett plots
    - C_{llr} values

0.31



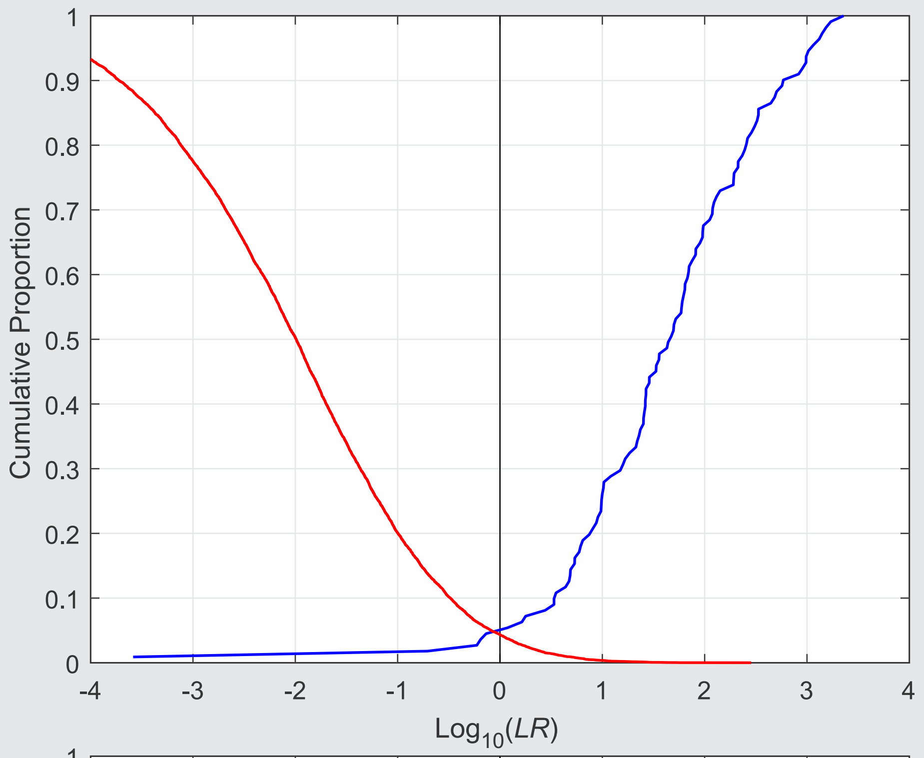
0.70



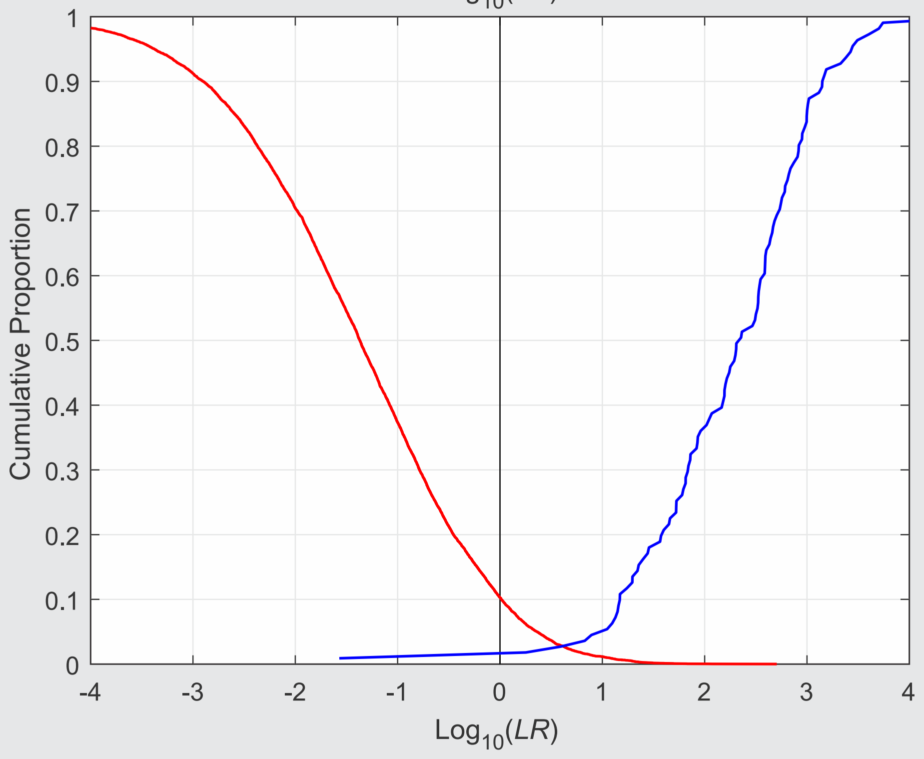
Validation plot

- 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



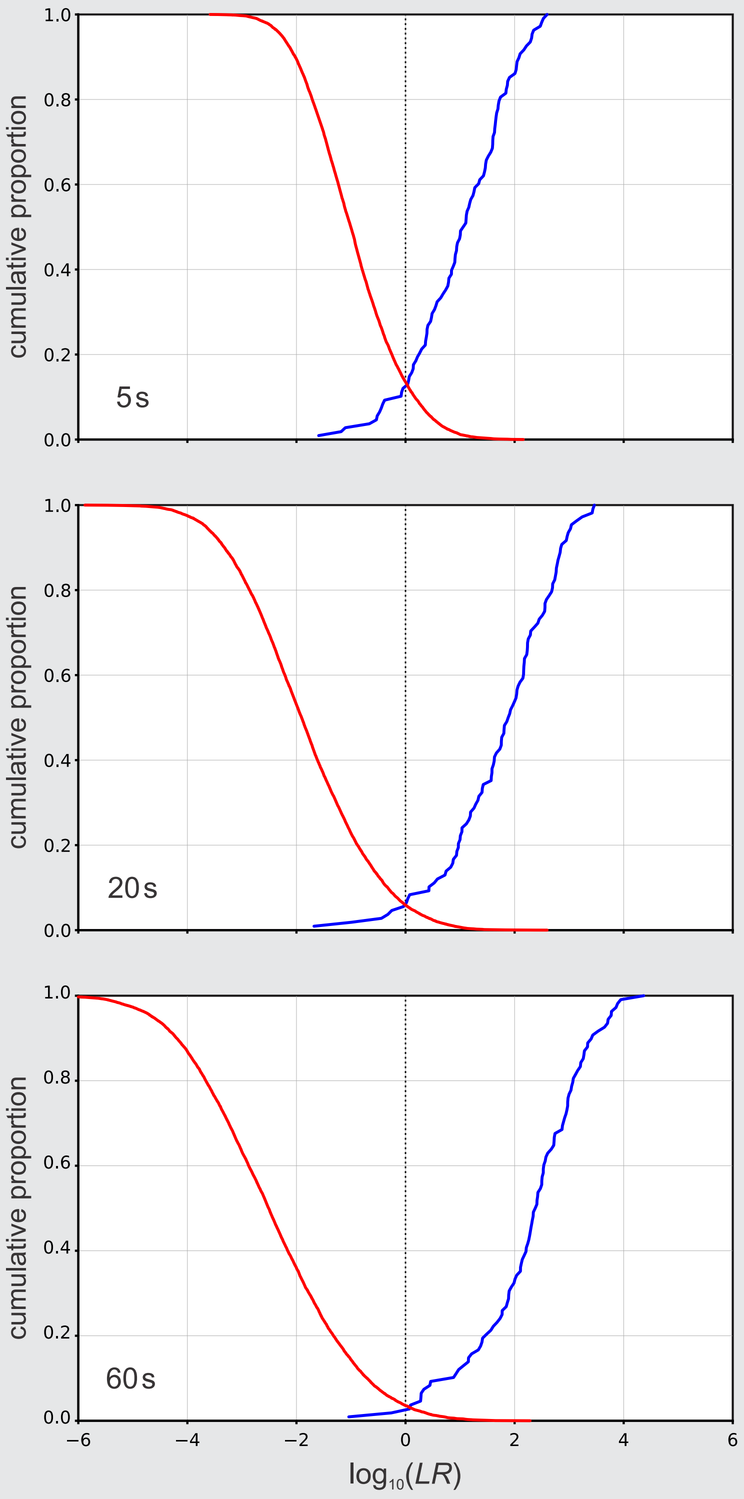
Validation plot

- 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



Consensus on Validation

Consensus on validation

- Morrison G.S., Enzinger E., Hughes V., Jessen M., Meuwly D., Neumann C., Planting S., Thompson W.C., van der Vloed D., Ypma R.J.F., Zhang C., Anonymous A., Anonymous B. (2021). **Consensus on validation of forensic voice comparison**. *Science & Justice*, 61, 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

Consensus on validation

- Key points:

2.12.1. The forensic practitioner **should communicate** to the court what **propositions** the forensic practitioner has adopted for the case, including what they have adopted as the **relevant population**.

2.12.2. The forensic practitioner **should communicate** to the court what the forensic practitioner understands the **conditions of the questioned-source and known-source items** to be.

Consensus on validation

- Key points:

  2.12.3. The forensic-comparison system **should be well calibrated**.

Consensus on validation

- Key points:

2.12.4. **Validation data should be representative of the relevant population** for the case, and **reflective of the conditions** of the questioned-source and known-source items in the case.

2.12.5. The forensic practitioner's **decision** as to whether the validation data are sufficiently representative of the relevant population for the case, and sufficiently reflective of the conditions of the questioned-source and known-source items in the case, will be a **subjective judgement**.

Consensus on validation

- Key points:

2.12.6. **Validation results should be presented as a Tippett plot and a C_{llr} value**. These **should be examined for signs of miscalibration**.

2.12.7. **The validation threshold (acceptance criterion) for C_{llr} should be 1**. As long as C_{llr} is less than 1, the system is providing useful information.

Consensus on validation

- Key points:

  2.12.8. To decide whether the **likelihood-ratio value** calculated for the comparison of the questioned-source and known-source items is **supported by the validation results**, it **should be compared with the values shown in the Tippett plot**

Thank You