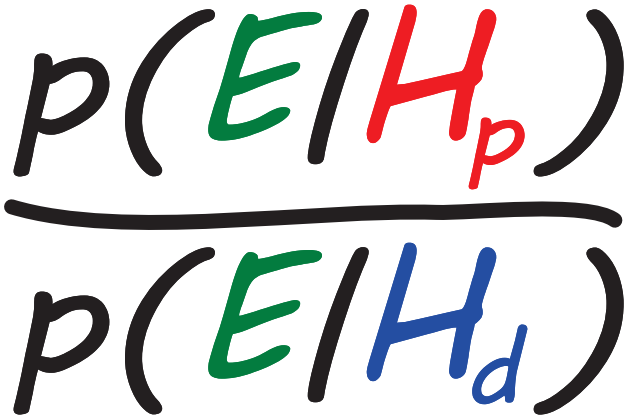
Measuring the validity and reliability of forensic analysis systems

Geoffrey Stewart Morrison



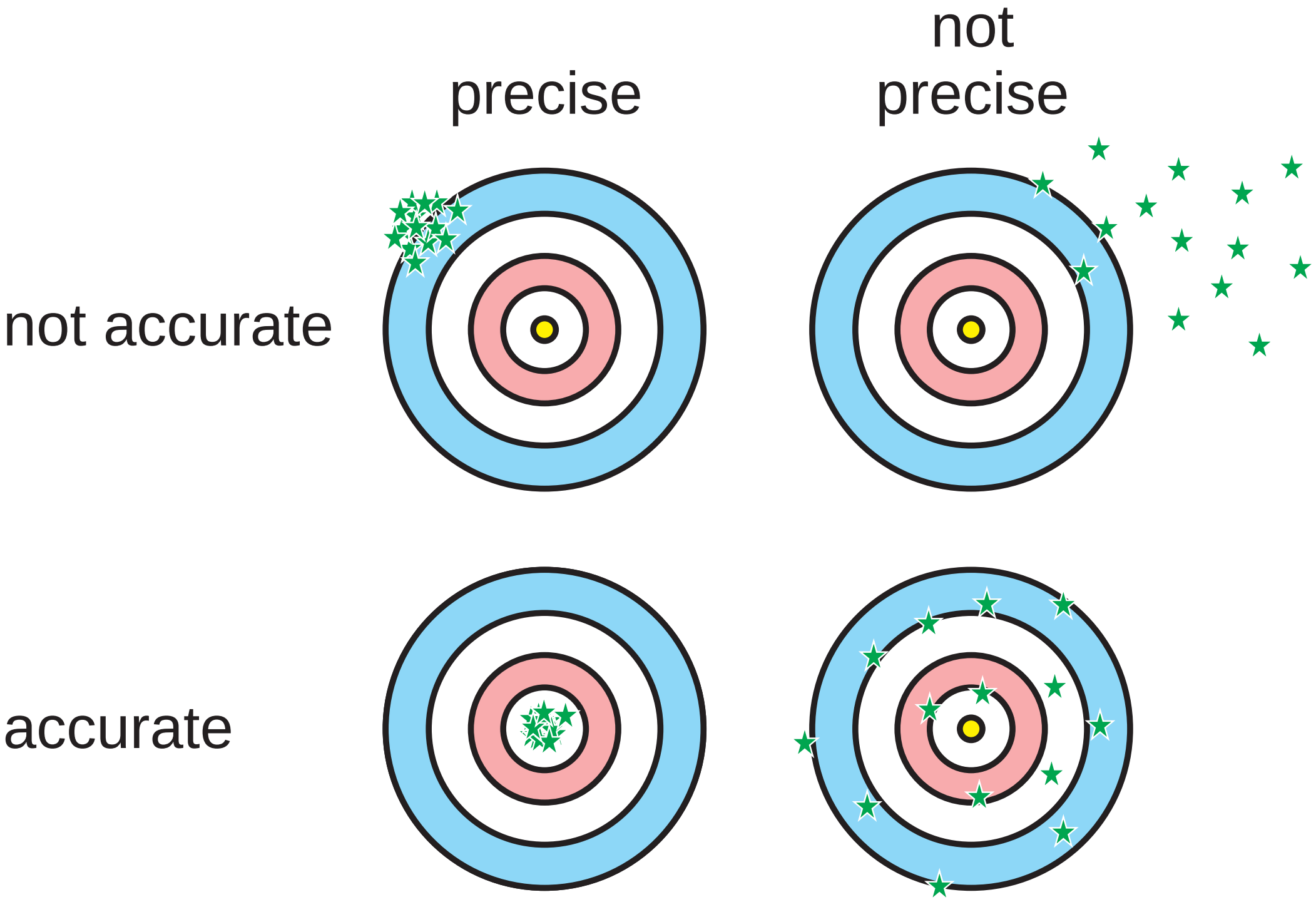
Concerns

- Logically correct framework for evaluation of forensic evidence
  - ENFSI Guideline for Evaluative Reporting 2015; NCFS Views on statistical statements 2016
- But what is the warrant for the opinion expressed? Where do the numbers come from?
  - Risinger at ICFIS 2011
- Demonstrate validity and reliability
  - *Daubert* 1993; NRC Report 2009; FSR Codes of Practice 2014; PCAST Report 2016
- Transparency
  - *R v T* 2010
- Reduce potential for cognitive bias
  - NIST/NIJ Human Factors in Latent Fingerprint Analysis 2012
- Communicate strength of forensic evidence to triers of fact

Paradigm

- Use of the likelihood-ratio framework for the evaluation of forensic evidence
  - logically correct
- Use of relevant data (data representative of the relevant population), quantitative measurements, and statistical models
  - transparent and replicable
  - relatively robust to cognitive bias
- Empirical testing of validity and reliability under conditions reflecting those of the case under investigation, using test data drawn from the relevant population
  - only way to know how well it works

Validity and Reliability
(Accuracy and Precision)


precise
not precise
not accurate
accurate

Measuring Validity

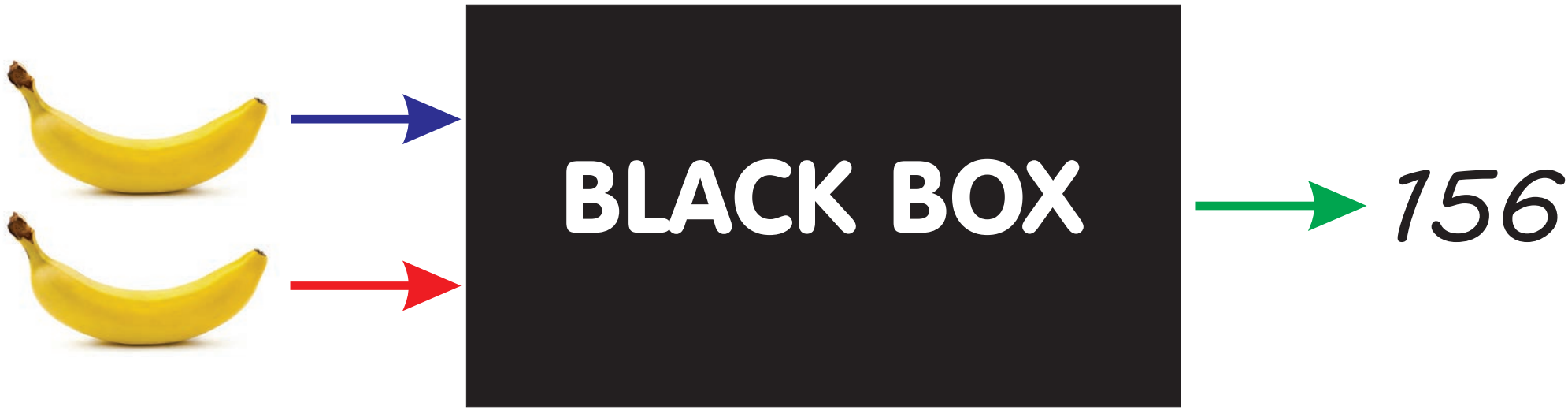
Measuring Validity

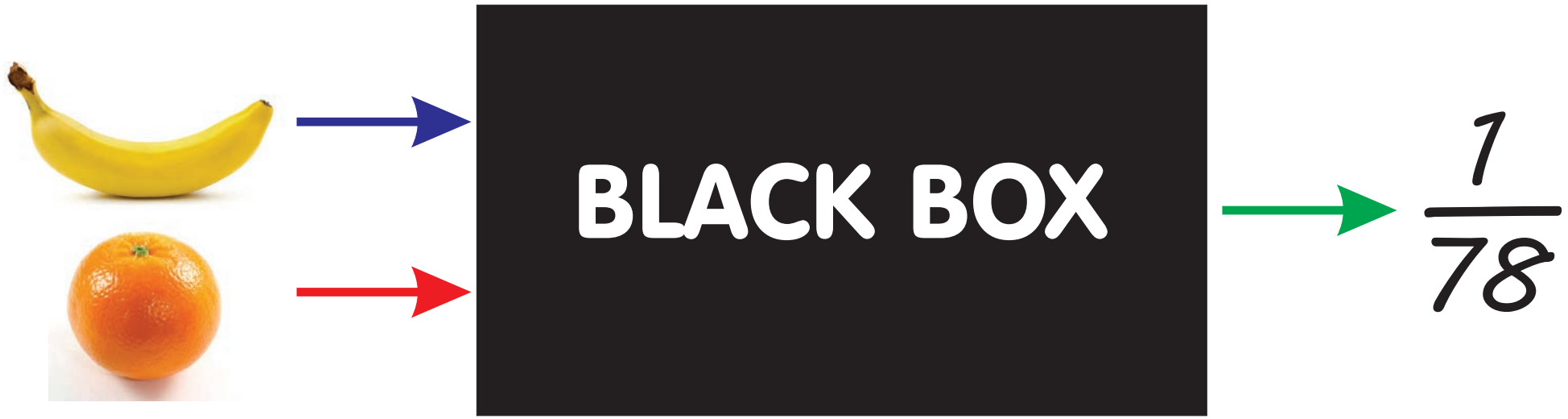
- Test set consisting of a large number of pairs of samples, some known to have the same origin and some known to have different origins

- **Test set must represent the relevant population and reflect the conditions of the case at trial**

- Use forensic-comparison system to calculate LR for each pair

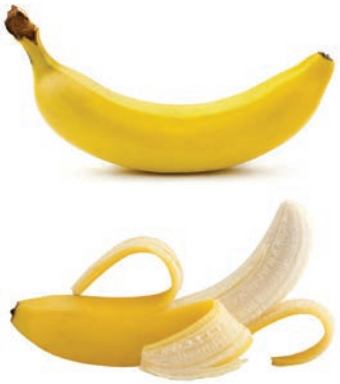
- Compare output with knowledge about input


BLACK BOX
156

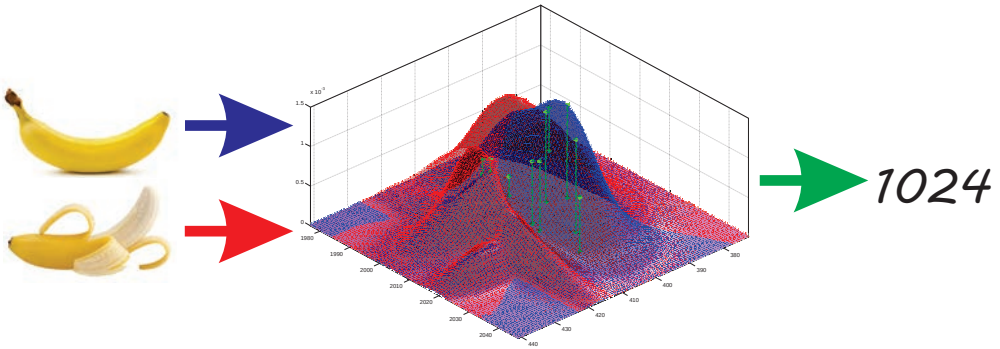

BLACK BOX
1
78

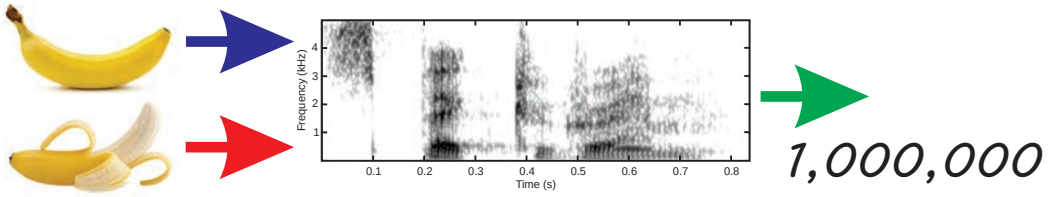
BLACK BOX

To be, or not to be, that is the question



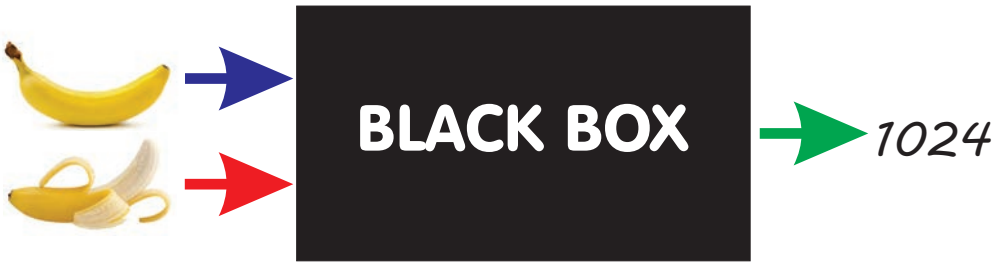
To be, or not to be, that is the question

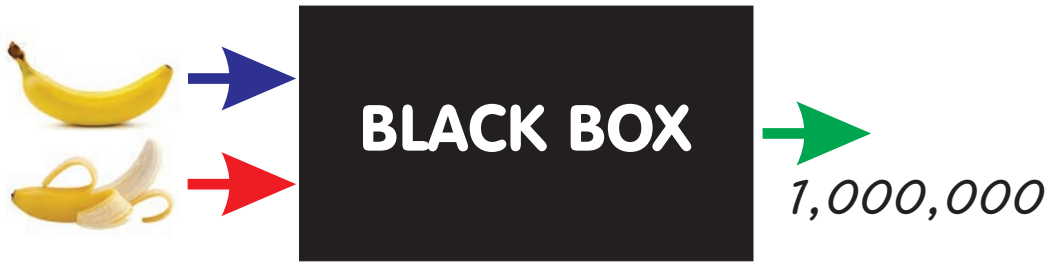

1024

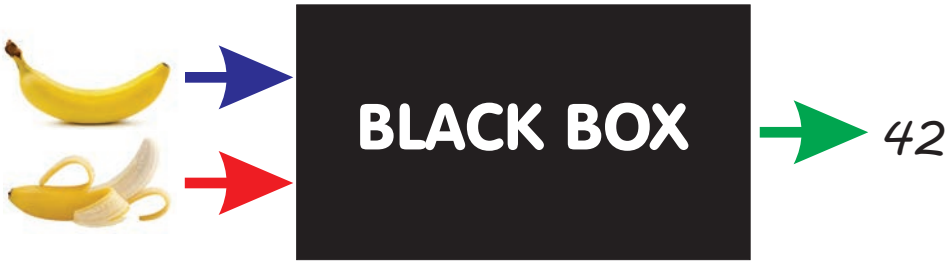

Frequency (kHz)
Time (s)
1,000,000

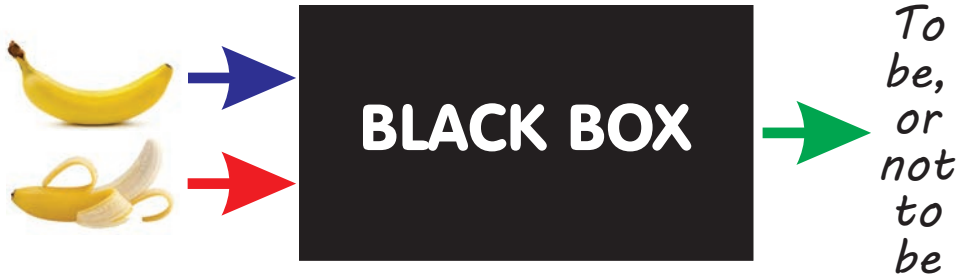

42


To be, or not to be


BLACK BOX
1024


BLACK BOX
1,000,000


BLACK BOX
42


BLACK BOX
To be, or not to be

Measuring Validity

- Correct-classification / classification-error rate is not appropriate
  - based on posterior probabilities
  - hard threshold rather than gradient

fact	decision	
	same	different
same	correct acceptance	false rejection
different	false acceptance	correct rejection

Measuring Validity

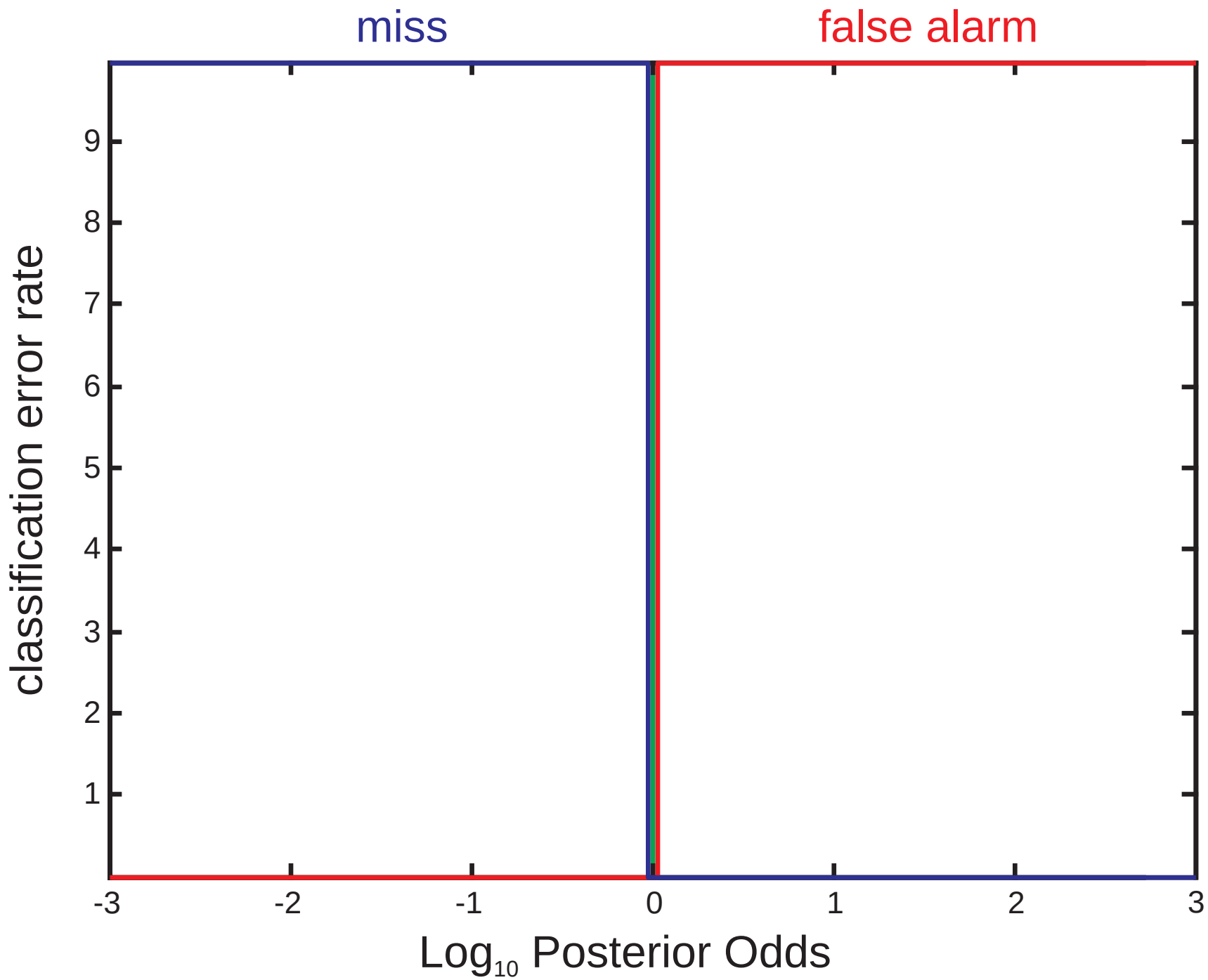
- Correct-classification / classification-error rate is not appropriate
  - based on posterior probabilities
  - hard threshold rather than gradient

fact	decision: same	decision: different
same		miss
different	false alarm	

Measuring Validity

- Correct-classification / classification-error rate is not appropriate
  - based on posterior probabilities
  - hard threshold rather than gradient

fact	decision	
	same	different
same	0	1
different	1	0


miss
false alarm
classification error rate
9
8
7
6
5
4
3
2
1
-3
-2
-1
0
1
2
3
Log10 Posterior Odds

Measuring Validity

- Goodness is **extent** to which LRs from same-origin pairs > 1, and LRs from different-origin pairs < 1

- Goodness is **extent** to which log(LR)s from same-origin pairs > 0, and log(LR)s from different-origin pairs < 0

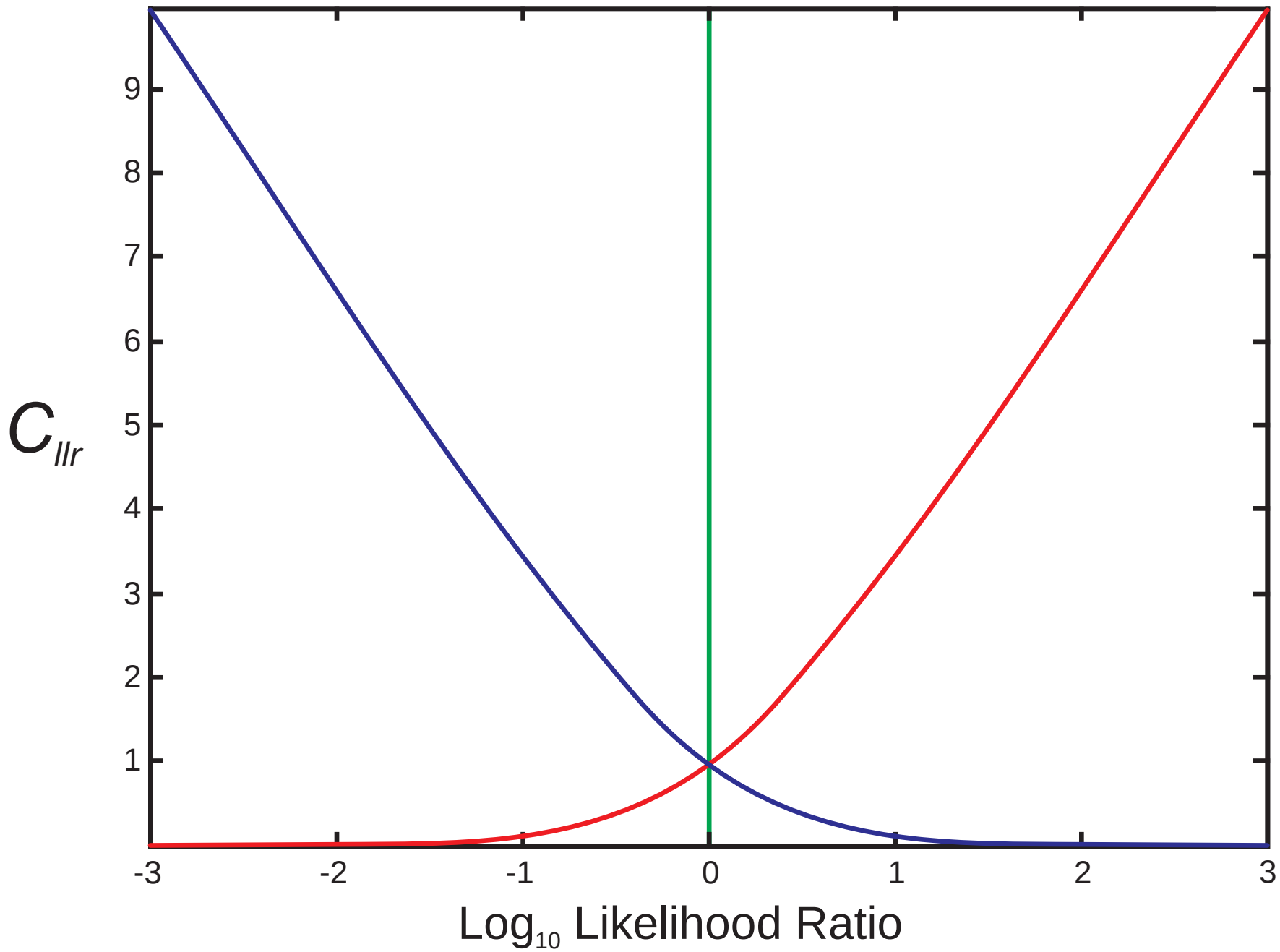
LR						
1/1000	**1/100**	**1/10**	**1**	**10**	**100**	**1000**
-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**
$\log_{10}$(LR)						

Measuring Validity

- A metric which captures the gradient goodness of a set of likelihood ratios derived from test data is the log-likelihood-ratio cost, C_{llr}

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{so}} \sum_{i=1}^{N_{so}} \log_2 \left(1 + \frac{1}{LR_{so_i}} \right) + \frac{1}{N_{do}} \sum_{j=1}^{N_{do}} \log_2 \left(1 + LR_{do_j} \right) \right)$$

Brümmer N, du Preez J (2006). **Application independent evaluation of speaker detection**, *Computer Speech & Language*, 20, 230–275. doi:10.1016/j.csl.2005.08.001

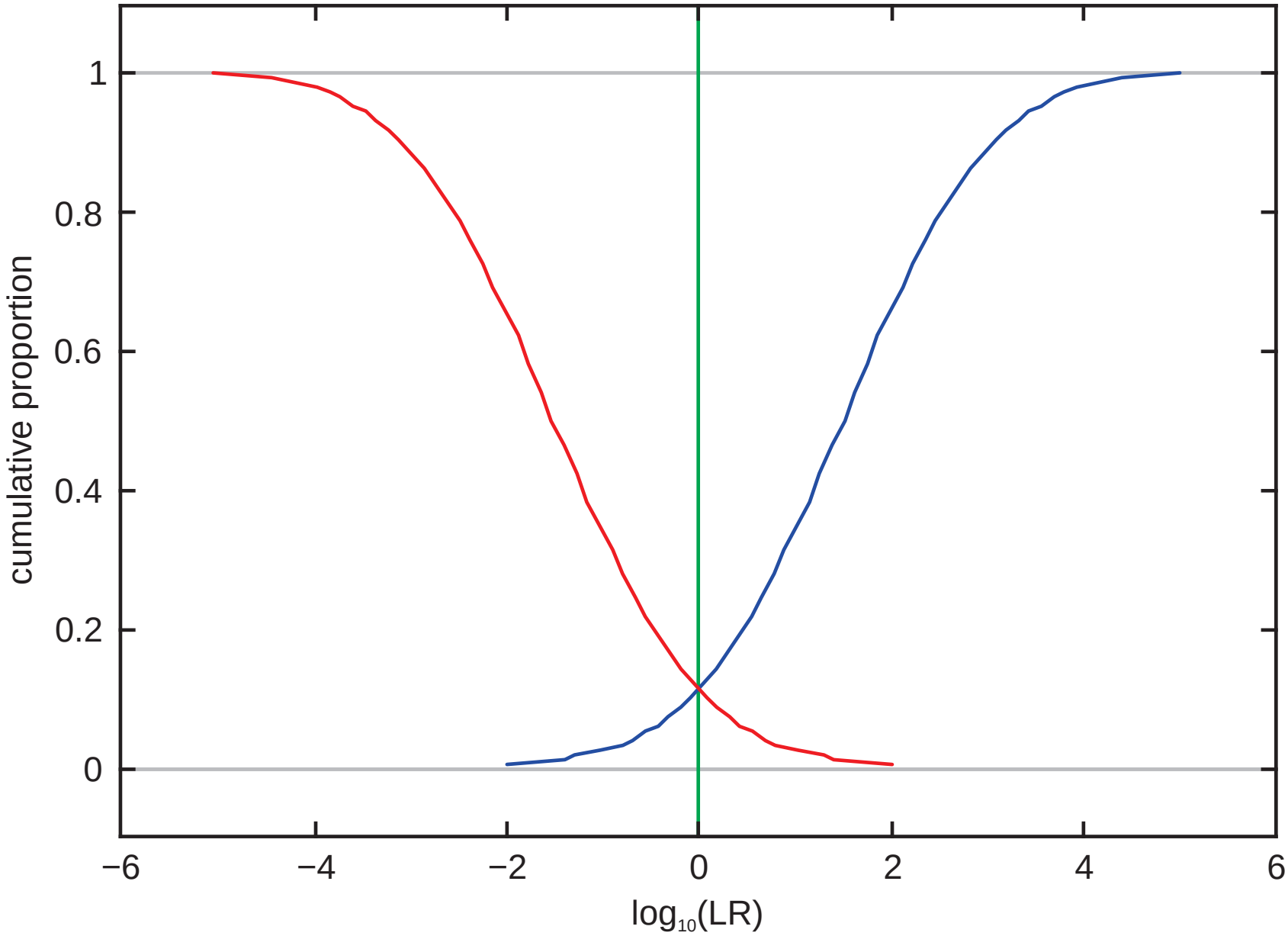

C_{llr}
9
8
7
6
5
4
3
2
1
-3
-2
-1
0
1
2
3
Log_{10} Likelihood Ratio

Measuring Validity

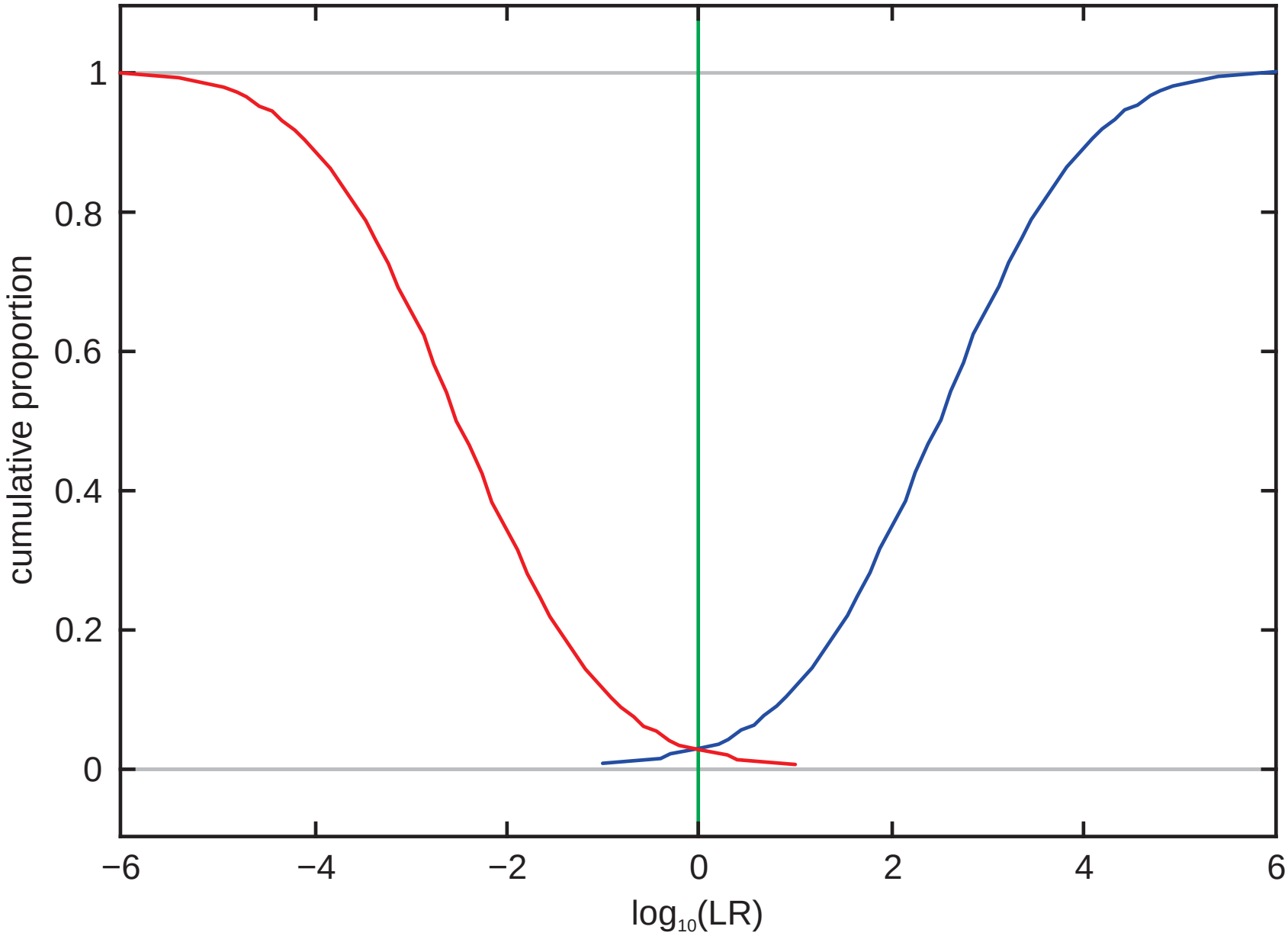
- System A: $C_{llr} = 0.548$
- System B: $C_{llr} = 0.101$
- System C: $C_{llr} = 1.018$

Tippett Plots

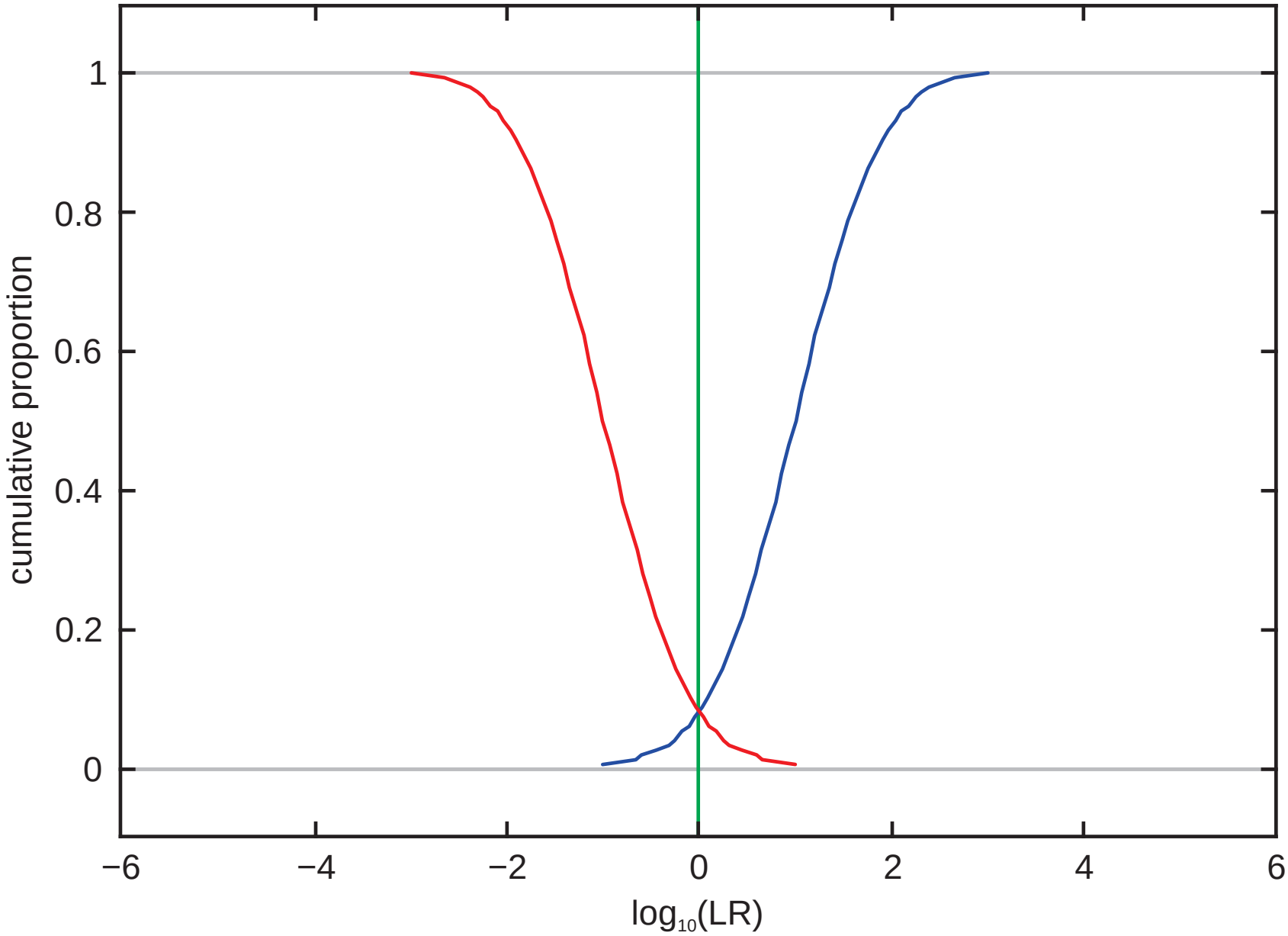
Tippett Plots



Tippett Plots

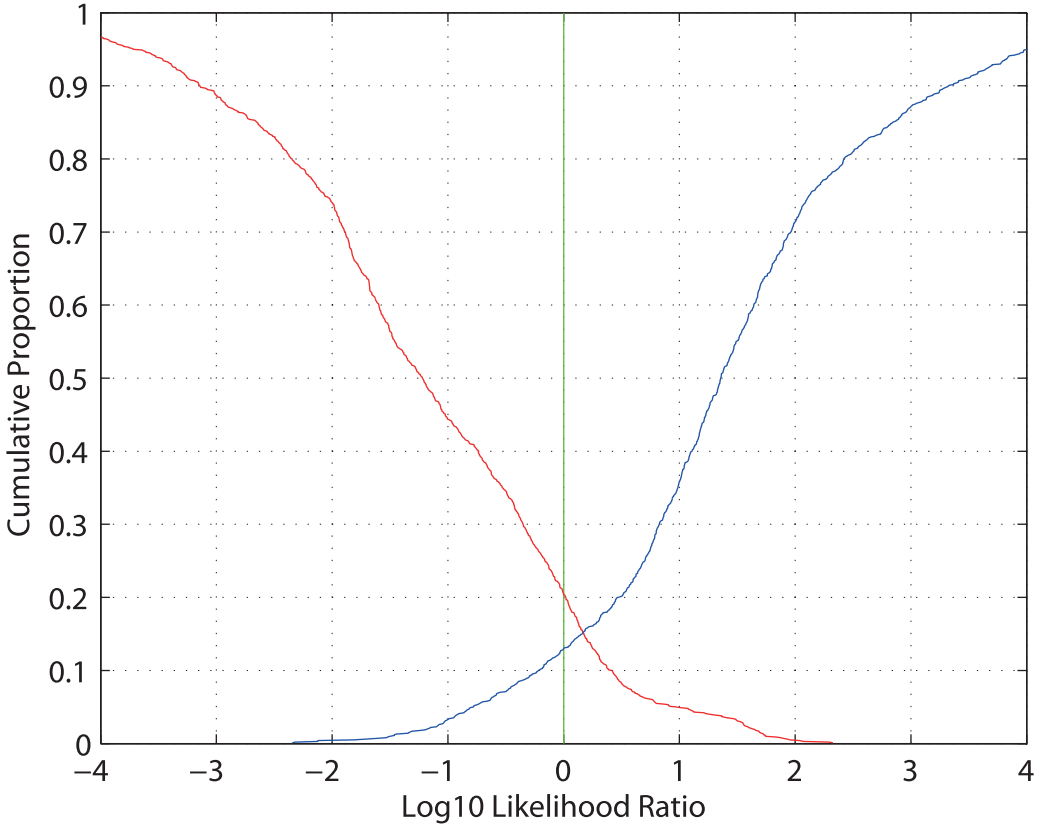


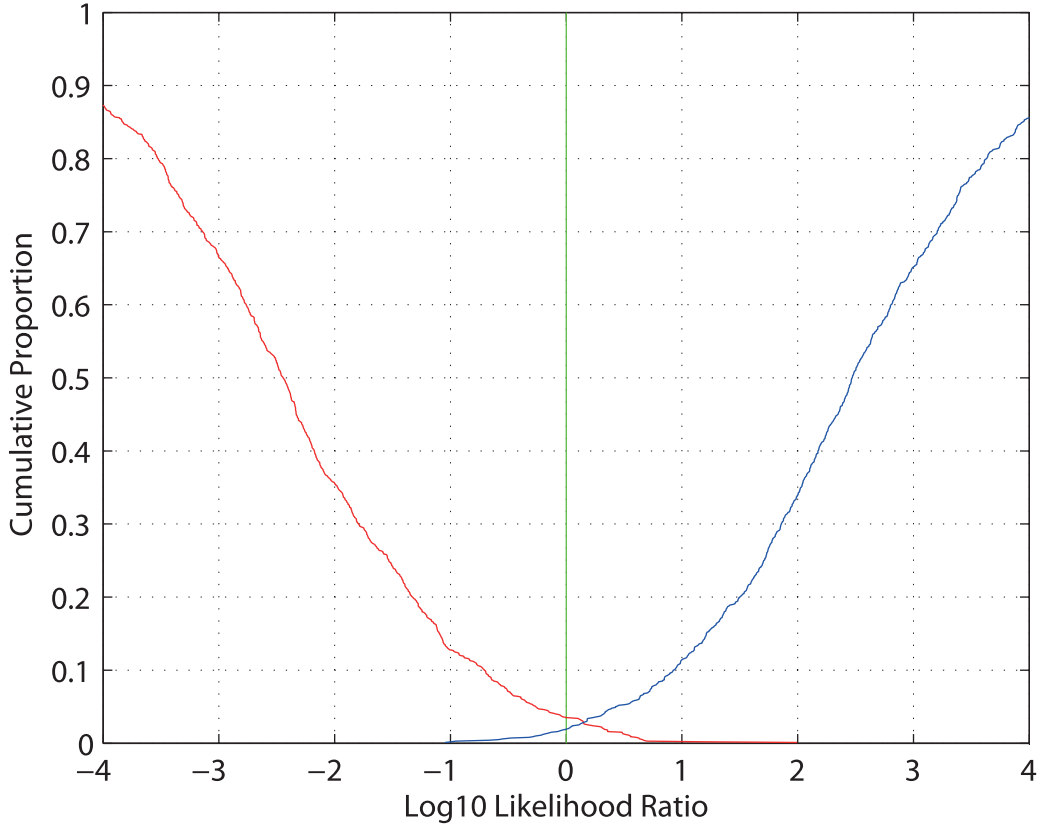
Tippett Plots



Tippett Plots

- System A: $C_{llr} = 0.548$
- System B: $C_{llr} = 0.101$





Measuring Reliability

Sources of imprecision

- intrinsic variability at the source level
  - within-source between-sample variability
- variability in the transfer process
- variability in the measurement technique
- variability in sampling of the relevant population
- variability in the estimation of statistical model parameters

Morrison, G. S. (2016). **Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate**. *Science & Justice*. doi:10.1016/j.scijus.2016.05.002

Measuring Reliability

- Imagine that in the test set we have three recordings (*A*, *B*, *C*) of each speaker

- *A* has the same conditions (speaking style, transmission channel, duration, etc.) as the offender recording

- *B* and *C* have the same conditions as the suspect recording

- Use LRs calculated on *A*-*B* and *A*-*C* pairs to estimate a 95% credible interval (CI)

Measuring Reliability

- Two pairs for each same-speaker comparison

suspect	recording	offender	recording
001	B	001	A
001	C	001	A
002	B	002	A
002	C	002	A
:	:	:	:

Measuring Reliability

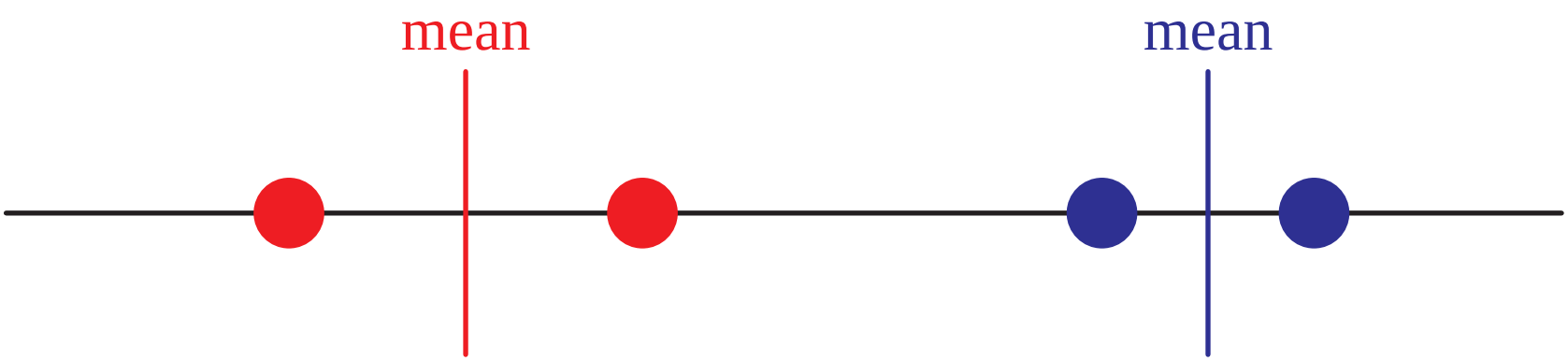
- Two pairs for each different-speaker comparison

suspect	recording	offender	recording
002	B	001	A
002	C	001	A
003	B	001	A
003	C	001	A
:	:	:	:
001	B	002	A
001	C	002	A
:	:	:	:

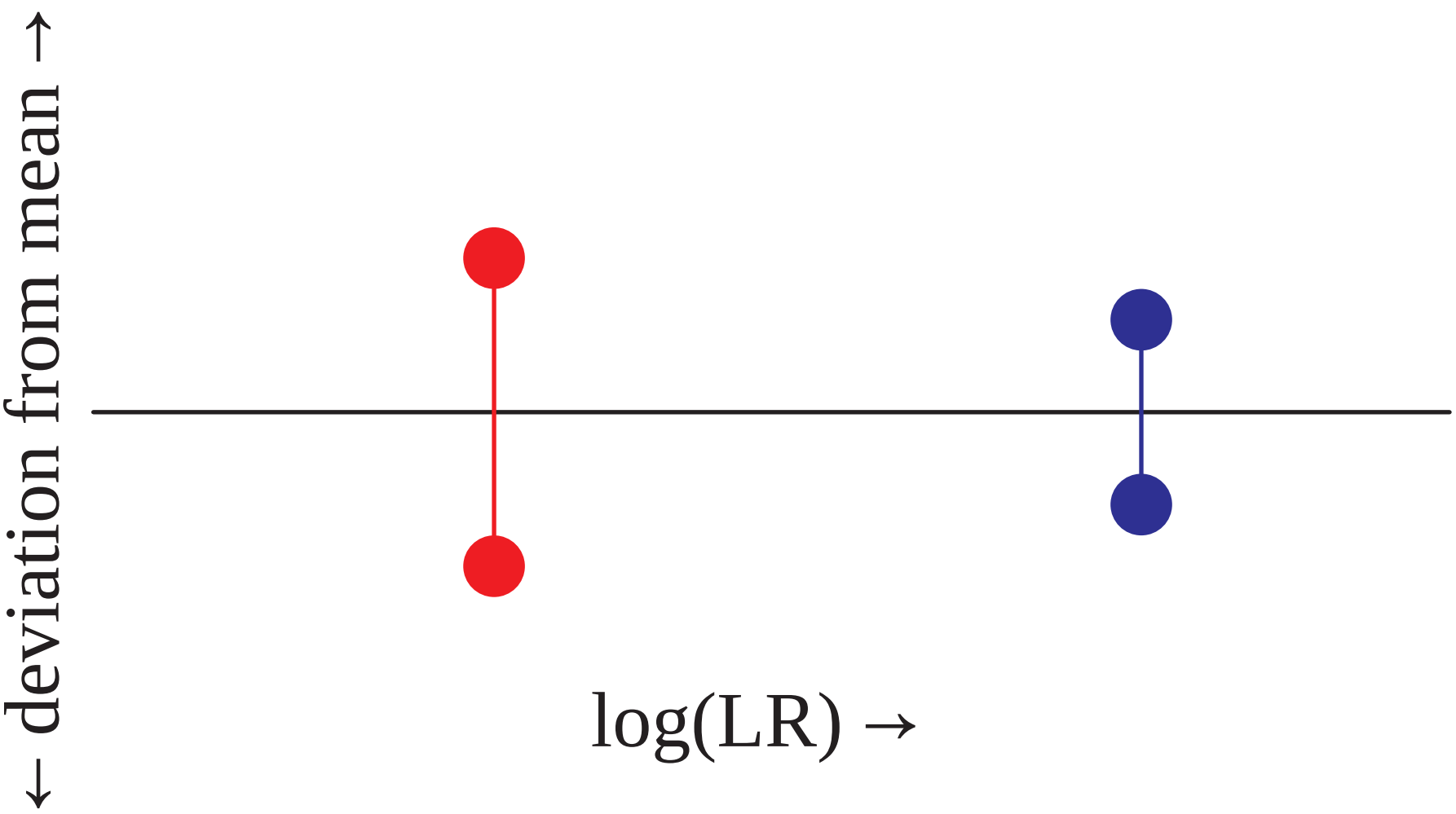
Measuring Reliability

log(LR) →

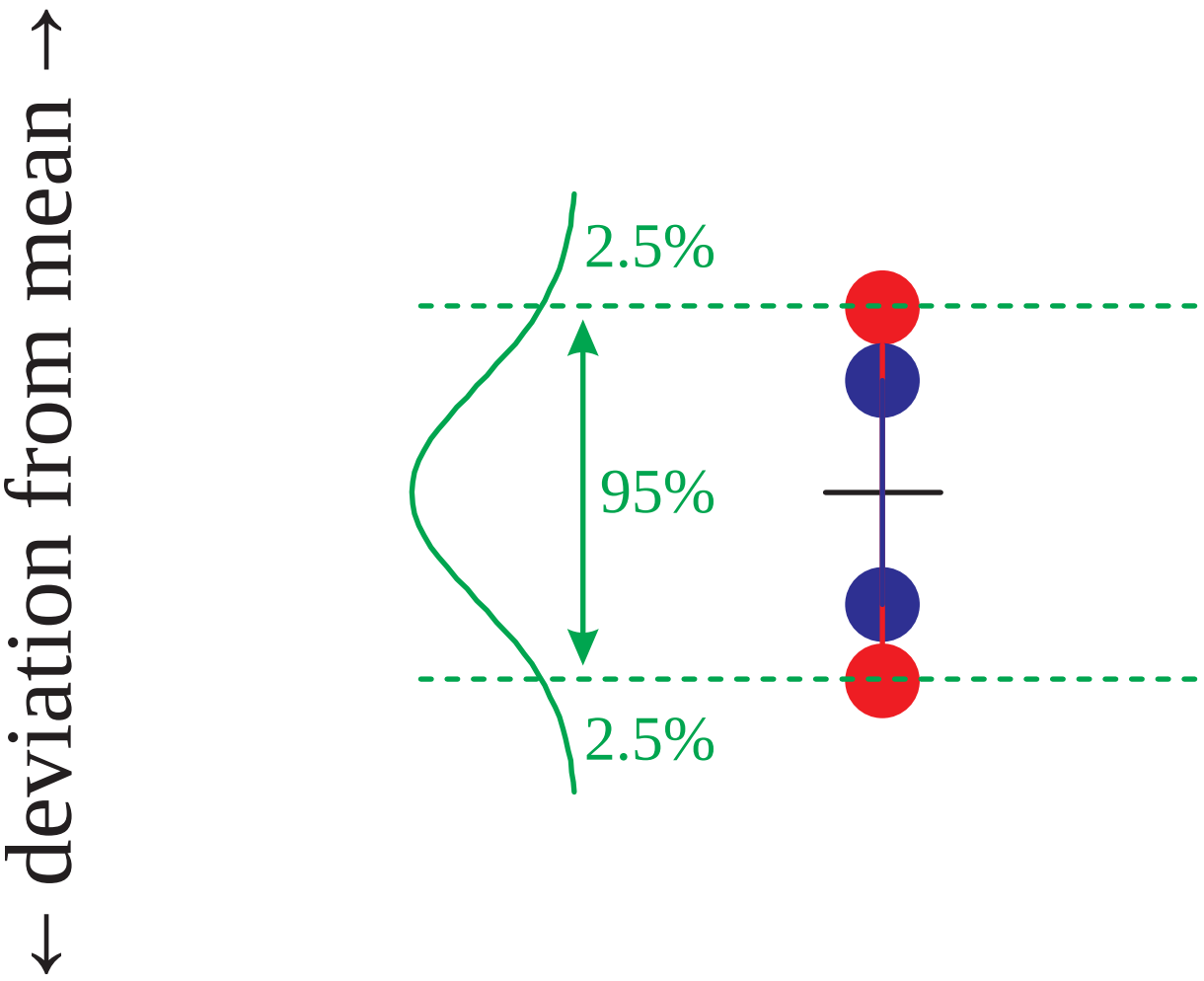
Measuring Reliability



Measuring Reliability



Measuring Reliability



Measuring Validity & Reliability

- System A: C_{llr} = 0.548 95% CI = ±0.498

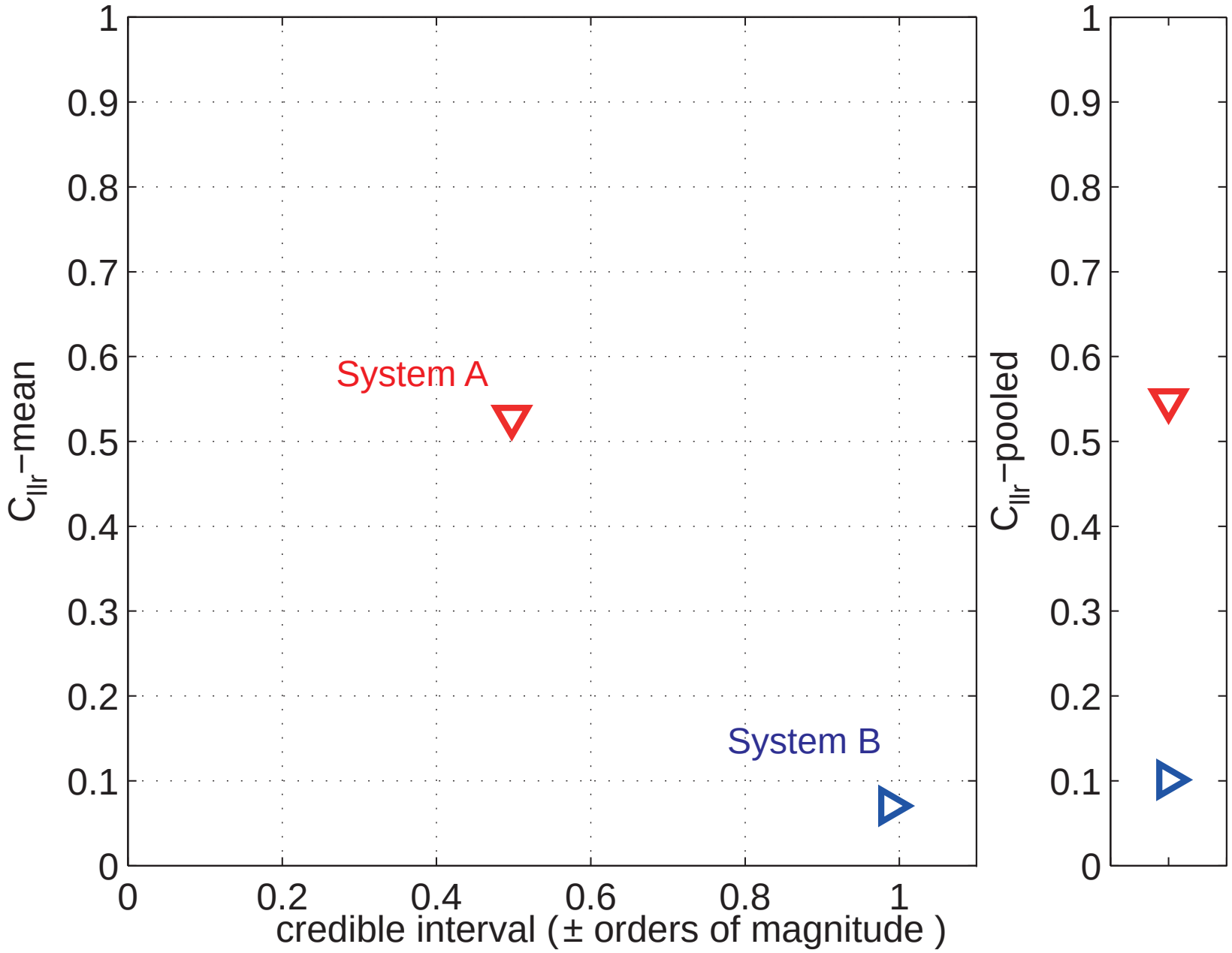
- System B: C_{llr} = 0.101 95% CI = ±0.988

Measuring Validity & Reliability

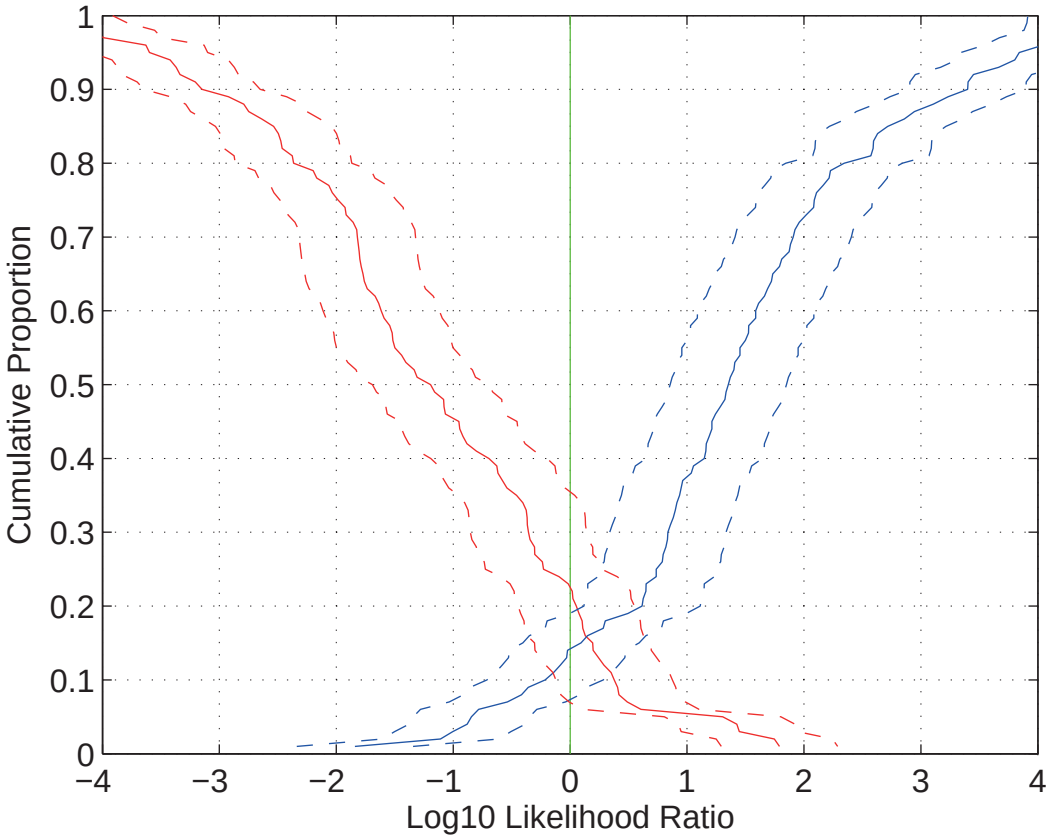
- System A: $C_{llr} = 0.548$ $C_{llr}^{\text{mean}} = 0.529$ 95% CI = ±0.498

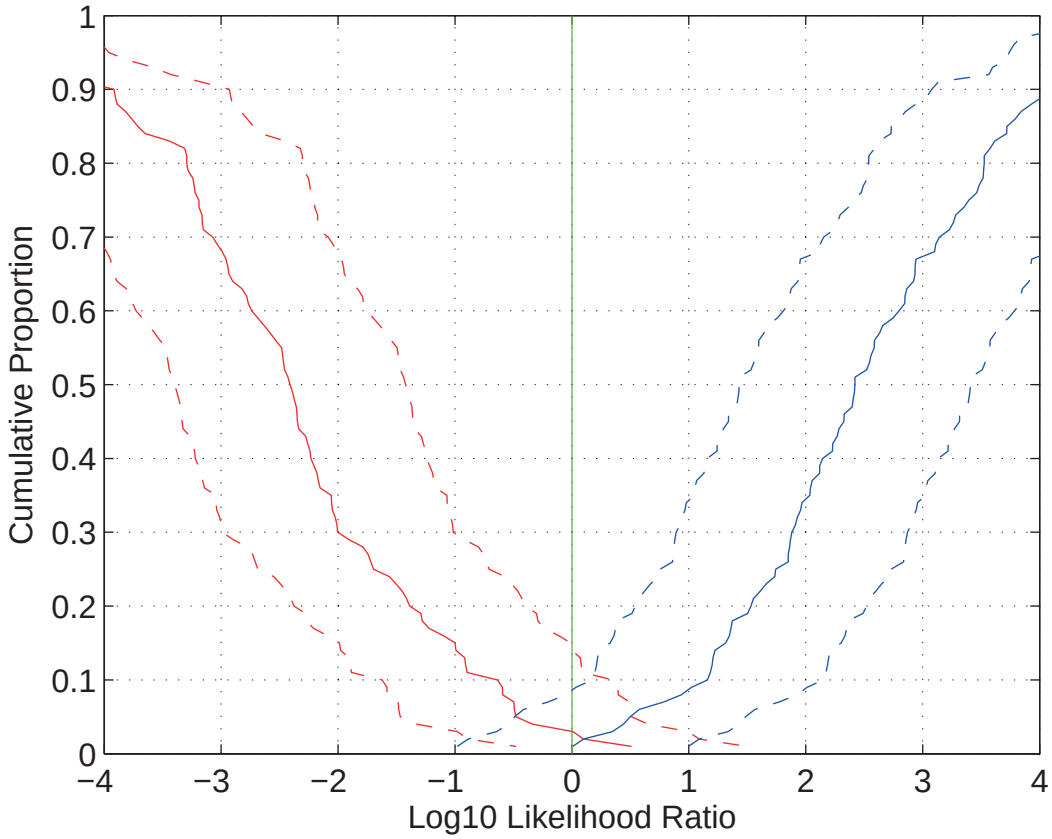
- System B: $C_{llr} = 0.101$ $C_{llr}^{\text{mean}} = 0.071$ 95% CI = ±0.988

Measuring Validity & Reliability



Tippett Plots





Summation

If the background and test data were consistent with the conditions in a case at trial, and the comparison of the known- and questioned-voice samples resulted in a likelihood ratio of, 100 ($\log_{10}(LR)$ of +2), and the 95% CI estimate was ±1 orders of magnitude (±1 in $\log_{10}(LR)$), then the forensic scientist could make a statement of the following sort:

Based on my evaluation of the evidence, I have calculated that one would be 100 times more likely to obtain the acoustic properties of the questioned-voice sample had been produced by the accused than had it been produced by some other speaker selected at random from the population.

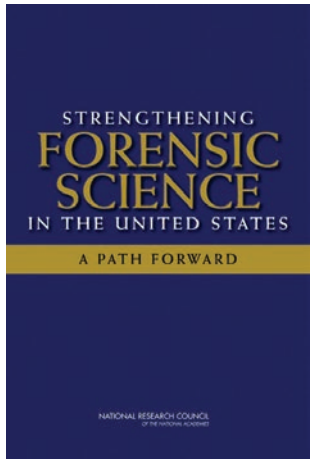
What this means is that whatever you believed about the relative probability of the same-speaker hypothesis versus the different-speaker hypothesis before this evidence was presented, you should now believe that the probability of the same-speaker hypothesis relative to the different-speaker hypothesis is 100 greater than you believed it to be before.

Based on my calculations, I am 95% certain that the acoustic differences are at least 10 times more likely and not more than 100 times more likely if the questioned-voice sample had been produced by the accused than if it had been produced by someone other than the accused.

Empirical Validation

Empirical Validation

- The National Research Council report to Congress on *Strengthening Forensic Science in the United States* (2009) urged that procedures be adopted which include:
  - "quantifiable measures of the reliability and accuracy of forensic analyses" (p. 23)
  - "the reporting of a measurement with an interval that has a high probability of containing the true value" (p. 121)
  - "the conducting of validation studies of the performance of a forensic procedure" (p. 121)



Empirical Validation

- The Forensic Science Regulator of England & Wales' *Codes of Practice and Conduct* (2014) require:
  - "all technical methods and procedures used by a provider shall be validated." (§20.1.1)
  - "Even where a method is considered standard and is in widespread use, validation will still need to be demonstrated." (§20.1.3)
  - "validation shall be carried out using simulated casework material ... and ... where appropriate, with actual casework material" (§20.7.3)
  - "demonstrate that they can provide consistent, reproducible, valid and reliable results" (§20.9.1)

Forensic Science Regulator
Overseeing Quality

Empirical Validation

- US Supreme Court: *Daubert v Merrell Dow Pharmaceuticals* (1993)
  - "In a case involving scientific evidence, *evidentiary reliability* will be based upon *scientific validity*" [emphasis in original]
  - "assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and ... whether that reasoning or methodology properly can be applied to the facts in issue."
  - "a key question to be answered in determining whether a theory or technique is scientific knowledge that will assist the trier of fact will be whether it can be (and has been) tested. ... '[T]he statements constituting a scientific explanation must be capable of empirical test'."
  - "in the case of a particular scientific technique, the court ordinarily should consider the known or potential rate of error"

Empirical Validation

- England & Wales: *Criminal Practice Directions* (2014)
  - "'the court must be satisfied that there is a sufficiently reliable scientific basis for the evidence to be admitted.'" (33A.4)
  - "whether the opinion takes proper account of matters, such as the degree of precision or margin of uncertainty, affecting the accuracy or reliability of those results;" (33A.5c)
  - "potential flaws ... which detract from ... reliability, ...
    (a) ... not ... subjected to sufficient scrutiny (including, where appropriate, experimental or other testing), ...
    (c) ... flawed data;
    (d) ... not properly carried out or applied, or was not appropriate for use in the particular case;" (33A.6)

Empirical Validation

- The President's Council of Advisors on Science and Technology report *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods* (PCAST, 2016)
  - "Without appropriate estimates of accuracy, an examiner's statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact." (p 6)
  - "the expert should not make claims or implications that go beyond the empirical evidence and the applications of valid statistical principles to that evidence." (p 6)
  - "Where there are not adequate empirical studies and/or statistical models to provide meaningful information about the accuracy of a forensic feature-comparison method, DOJ attorneys and examiners should not offer testimony based on the method." (p 19)

Experience

Experience

- For an expert to say “I think this is true because I have been doing this job for *x* years” is, in my view, unscientific. On the other hand, for an expert to say “I think this is true and my judgement has been tested in controlled experiments” is fundamentally scientific.

Evett IW (1991) **Interpretation: a personal odyssey**. In C.G.G. Aitken, D.A. Stoney (Eds.), *The Use of Statistics in Forensic Science*. Ellis Horwood, Chichester, UK. pp. 9–22.

Experience

- Experience in applying spectrographic voice identification in law enforcement has led proponents of the method to express confidence its reliability. The basis for this confidence is not, however, accessible to objective assessment.

- Validation of this approach to voice identification becomes a matter of replicable experiments on the expert himself, considered as a voice identifying machine. ... validation requires experimental assessment of performance on relevant tasks. ... It may be objected that this minimal set of tests is unreasonably arduous. We do not believe that it is. As scientists we could accept no less in checking the reliability of a “black box” supposed to perform speaker identification.

Bolt RA, Cooper FS, David EE Jr., Denes PB, Pickett JM, Stevens KN (1970) **Speaker identification by speech spectrograms: a scientists’ view of its reliability for legal purposes**. *Journal of the Acoustical Society of America* 47, 597–612, http://dx.doi.org/10.1121/1.1911935.

Experience

- The President's Council of Advisors on Science and Technology report *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods* (PCAST, 2016)
  - "neither experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of 'judgment.' It is an empirical matter for which only empirical evidence is relevant. Similarly, an expert's expression of *confidence* based on personal professional experience or expressions of *consensus* among practitioners about the accuracy of their field is no substitute for error rates estimated from relevant studies. For forensic feature-comparison methods, establishing foundational validity based on empirical evidence is thus a *sine qua non*. Nothing can substitute for it." (p 6)

Thank You

http://geoff-morrison.net/

http://forensic-evaluation.net/