Special Session on Forensic Voice Comparison and Forensic Acoustics @ 2nd Pan-American/Iberian Meeting on Acoustics, Cancún, México, 15–19 November, 2010
http://cancun2010.forensic-voice-comparison.net

3aSC6

Measuring Reliability in Forensic Voice Comparison

Geoffrey Stewart Morrison
Julien Epps
Philip Rose
Tharmarajah Thiruvaran
Cuiling Zhang

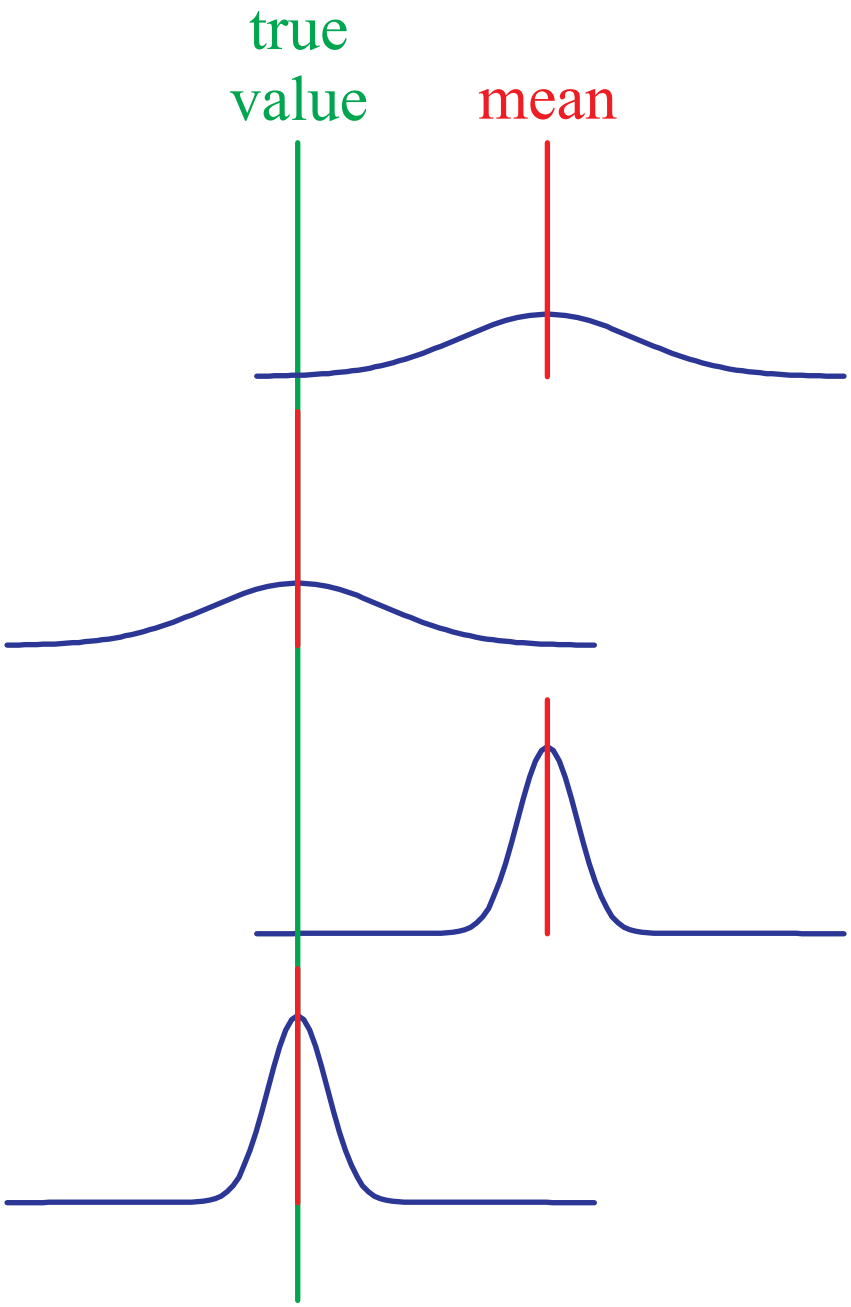
FORENSIC VOICE COMPARISON LABORATORY
SCHOOL OF ELECTRICAL ENGINEERING & TELECOMMUNICATIONS
UNSW
THE UNIVERSITY OF NEW SOUTH WALES
SYDNEY • AUSTRALIA

SCHOOL OF LANGUAGE STUDIES
ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

中国刑事警察学院
China Criminal Police University

Australian Government
Australian Research Council

Validity and Reliability (Accuracy and Precision)



poor accuracy
poor precision

good accuracy
poor precision

poor accuracy
good precision

good accuracy
good precision

Validity and Reliability in Forensic Science

- 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)

Testing the Validity of a Forensic-Comparison System

Measuring Validity

- Test set consisting of a large number of pairs known to be same origin and a large number of pairs known to be different origin
- Use forensic-comparison system to calculate LR for each pair
- Compare output with knowledge about input

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	correct acceptance	incorrect rejection
different	incorrect acceptance	correct rejection

Measuring Validity

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

- 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}

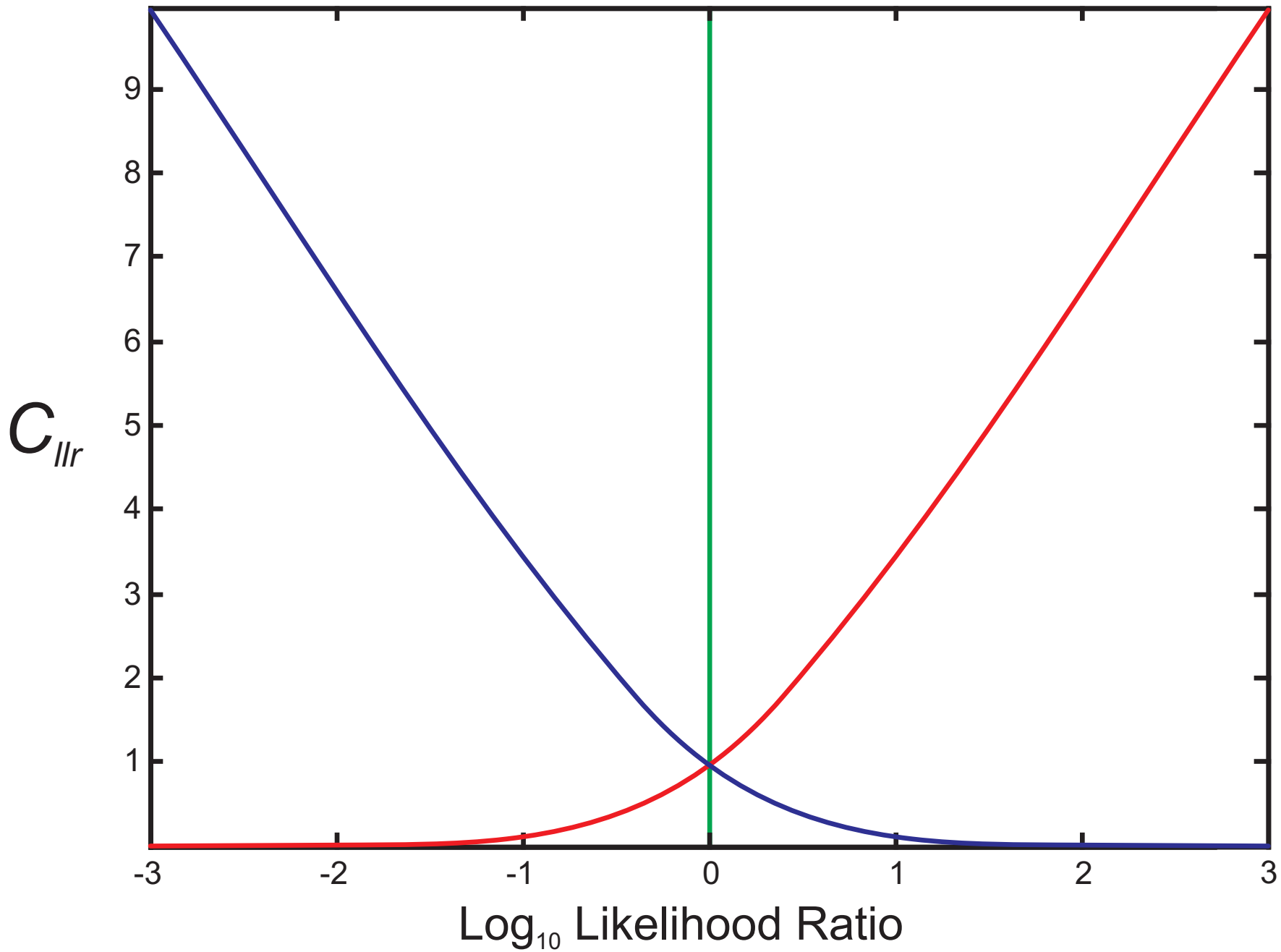
Measuring Validity

- 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)						

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{ss}} \sum_{i=1}^{N_{ss}} \log_2 \left(1 + \frac{1}{LR_{ss_i}} \right) + \frac{1}{N_{ds}} \sum_{j=1}^{N_{ds}} \log_2 \left(1 + LR_{ds_j} \right) \right)$$

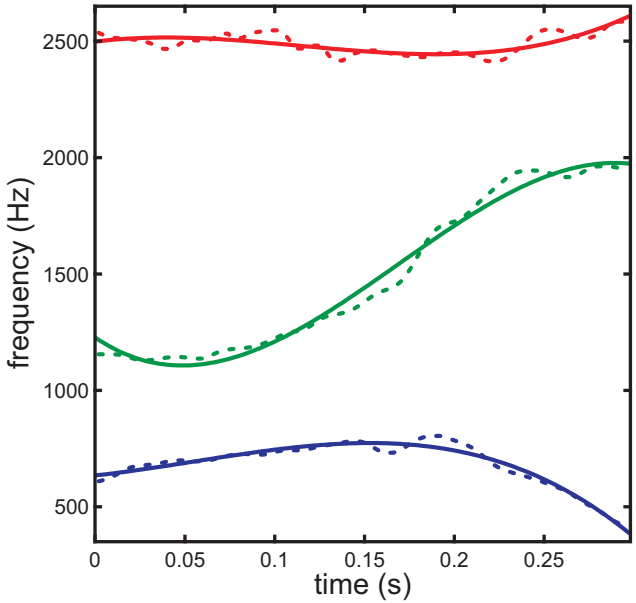

C_{llr}
9
8
7
6
5
4
3
2
1
-3
-2
-1
0
1
2
3
Log_{10} Likelihood Ratio

Example of Testing the Validity of Forensic-Comparison Systems

System and Data

(Morrison, 2011)

- **Acoustic-phonetic systems:**
  - **dual-target:** "initial target" and "final target" in /aɪ/ tokens
  - **trajectory:** coefficient values of cubic polynomial fitted to formant trajectories of /aɪ/ tokens
  - Aitken & Lucy (2004) MVKD
  - logistic-regression calibration



- **Database:**
  - 25 male Australian English speakers
  - two non-contemporaneous recordings (24 tokens / recording)
  - cross-validation

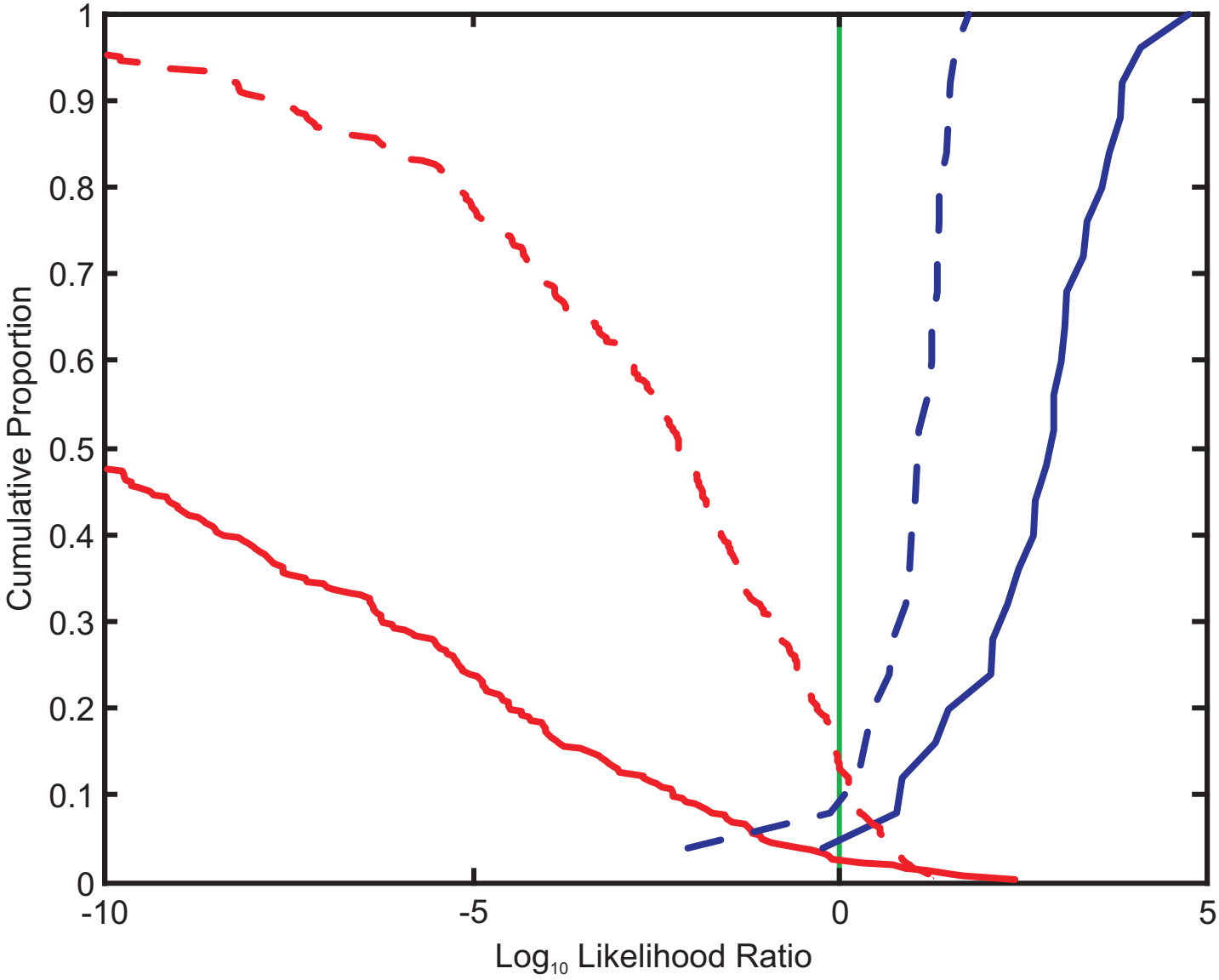
Results

- **dual-target**

$$C_{llr} = 0.43$$

- **trajectory**

$$C_{llr} = 0.10$$



Testing the Reliability of a Forensic-Comparison System

Measuring Reliability

- Imagine that we have four recordings (A, B, C, D) of each speaker
- There are two non-overlapping pairs for each same-speaker comparison and four non-overlapping pairs for each different-speaker comparison
- These are statistically independent and can be used to estimate a 95% credible interval (CI)

Measuring Reliability

- Two non-overlapping pairs for each same-speaker comparison

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

Measuring Reliability

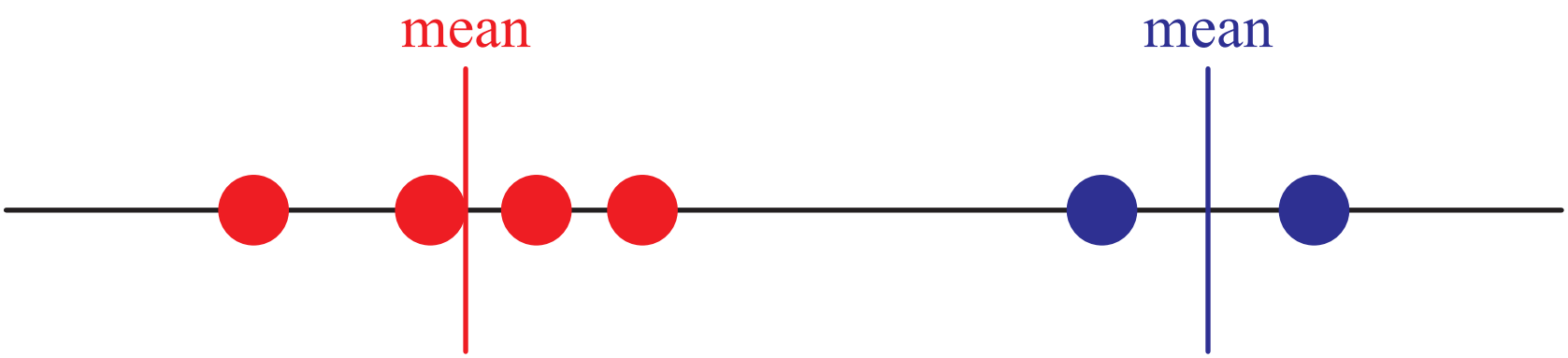
- Four non-overlapping pairs for each different-speaker comparison

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

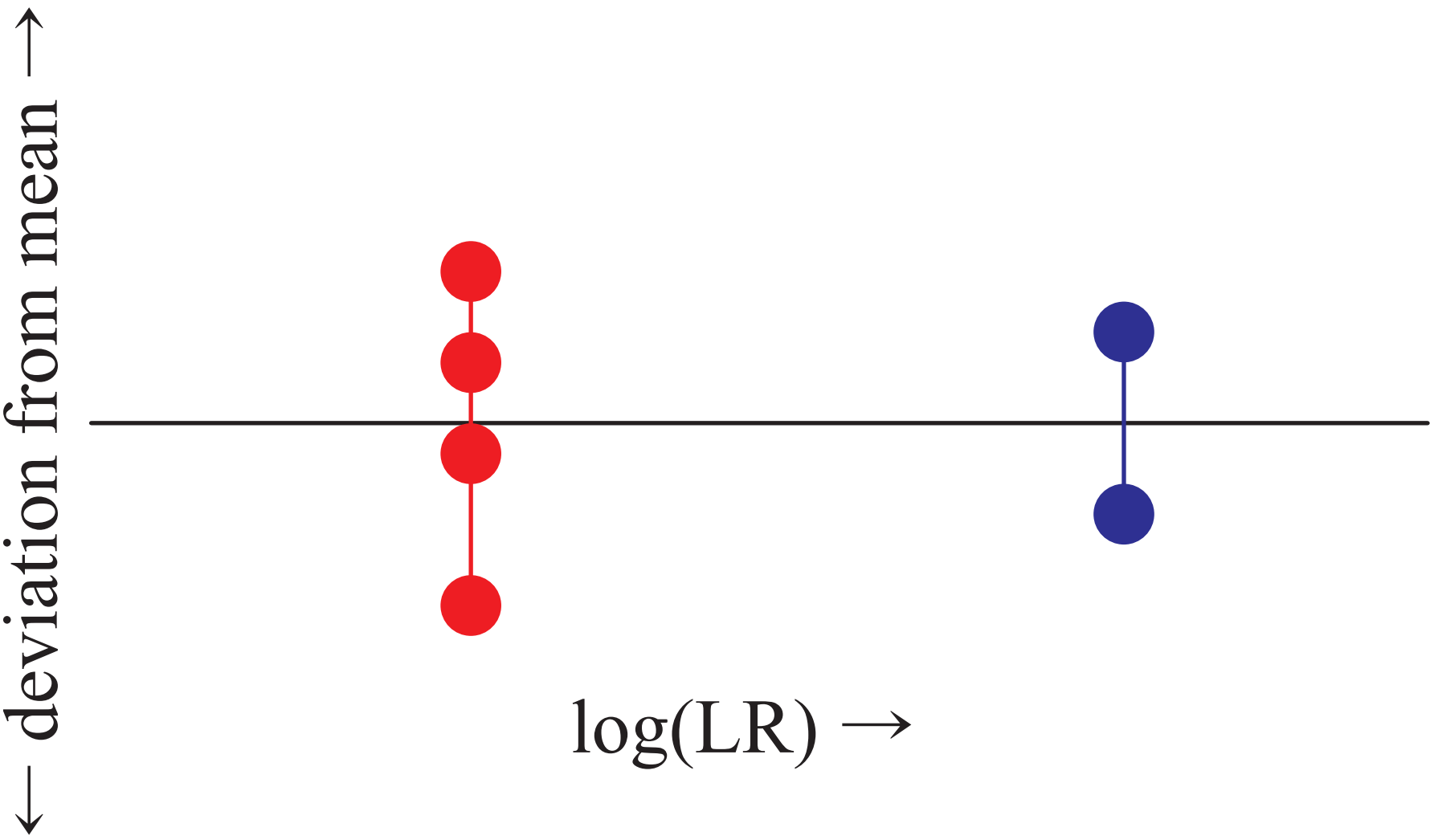
Measuring Reliability

log(LR) →

Measuring Reliability

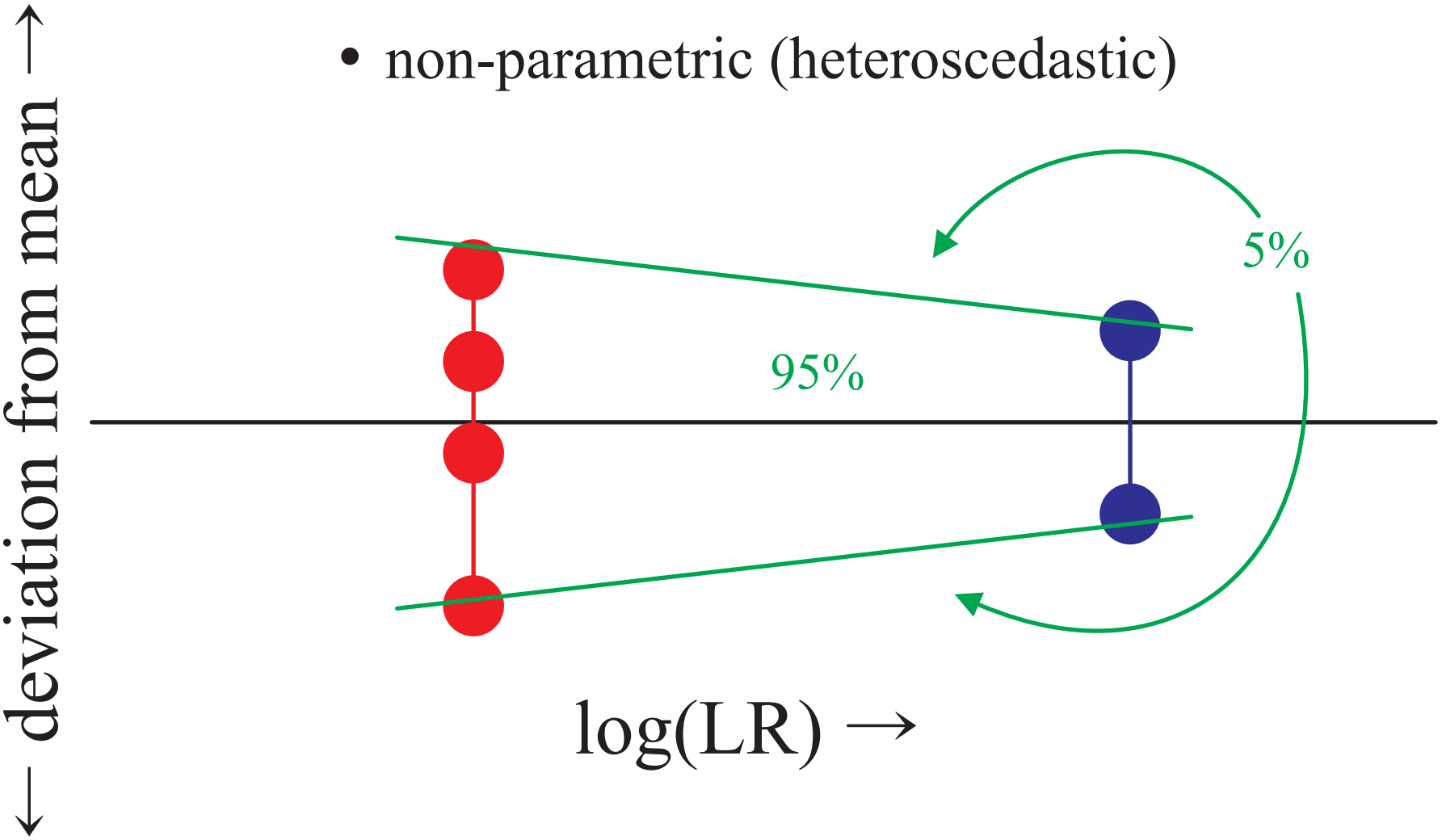


Measuring Reliability



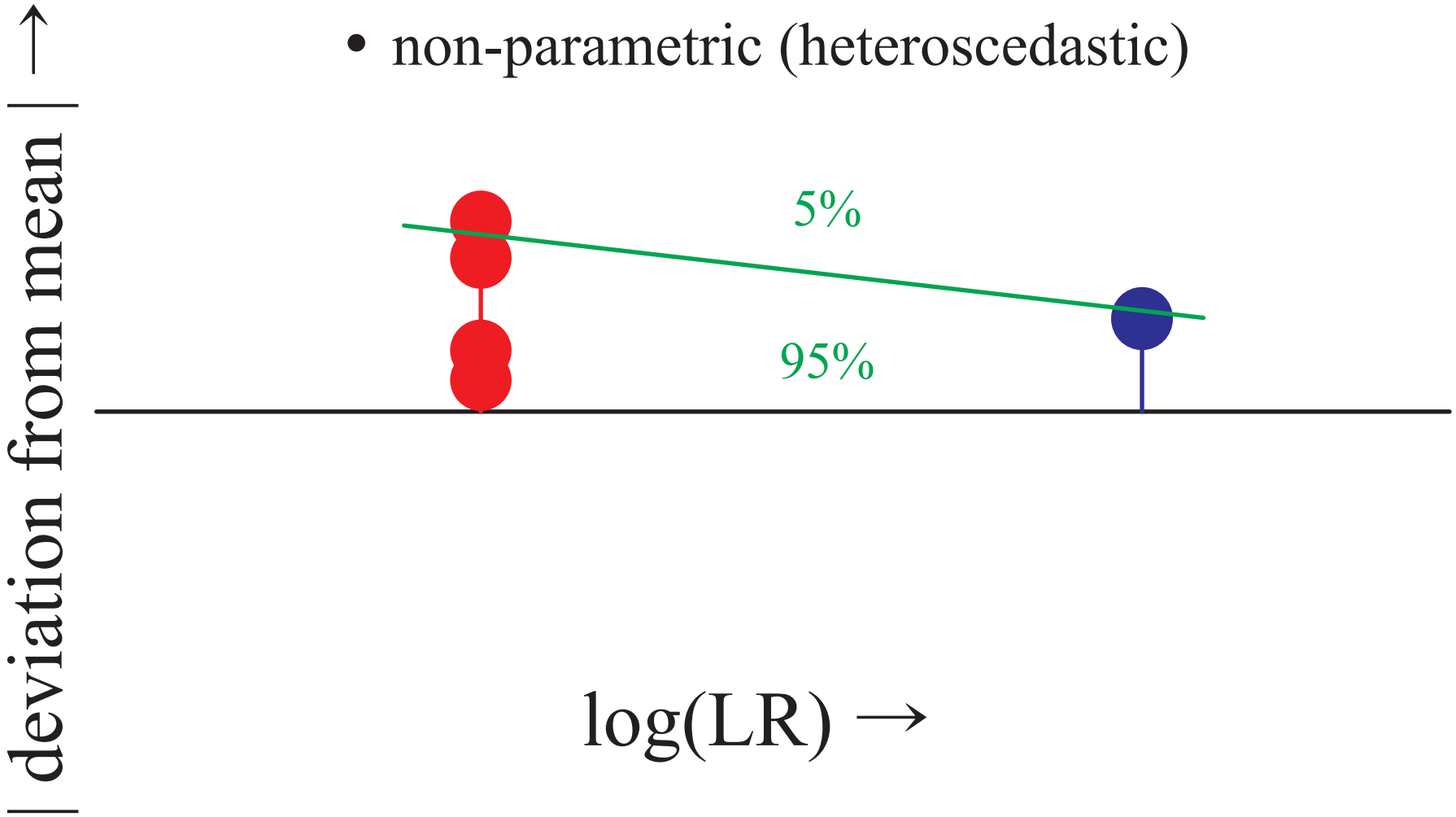
Measuring Reliability

- non-parametric (heteroscedastic)



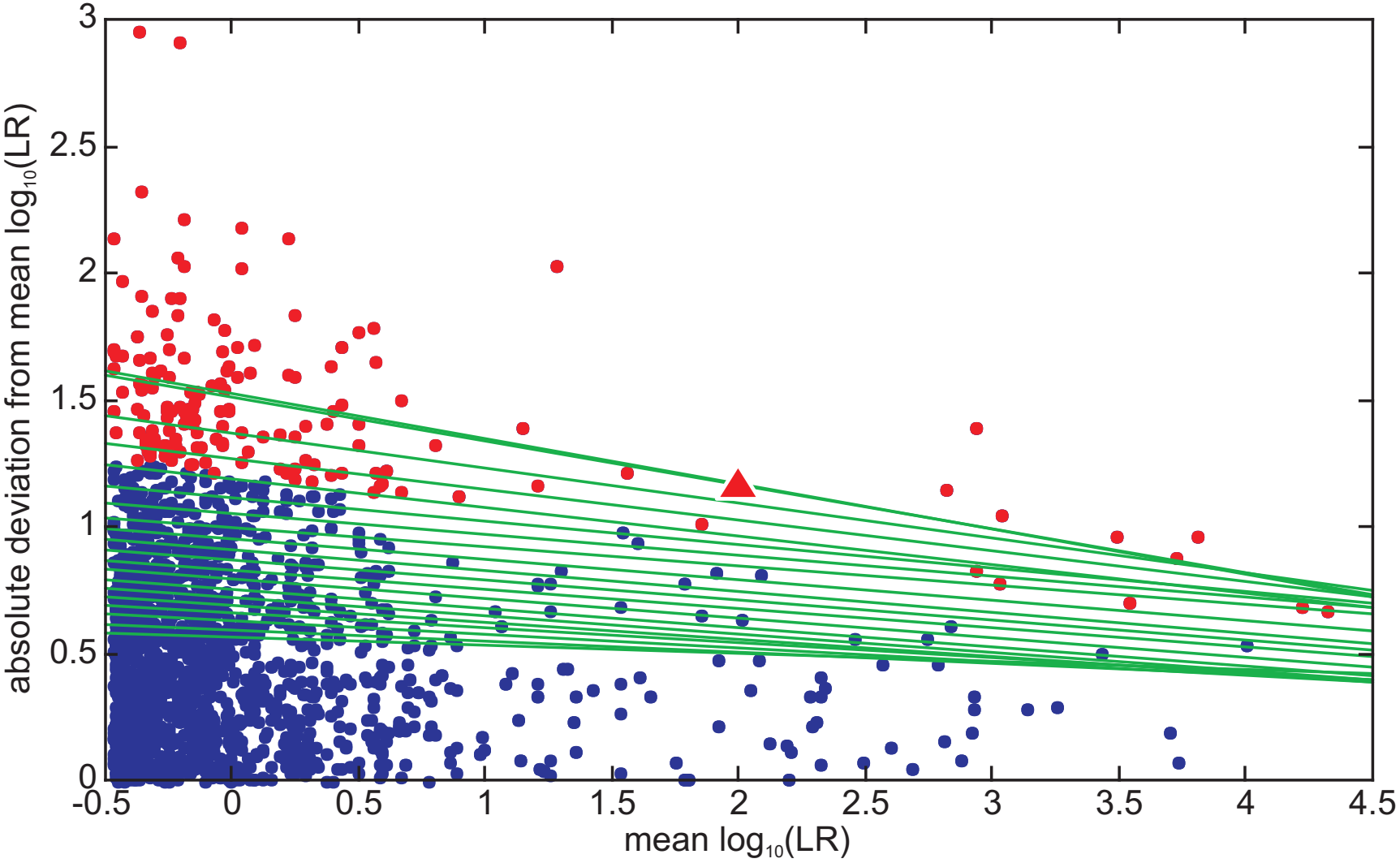
Measuring Reliability

- non-parametric (heteroscedastic)



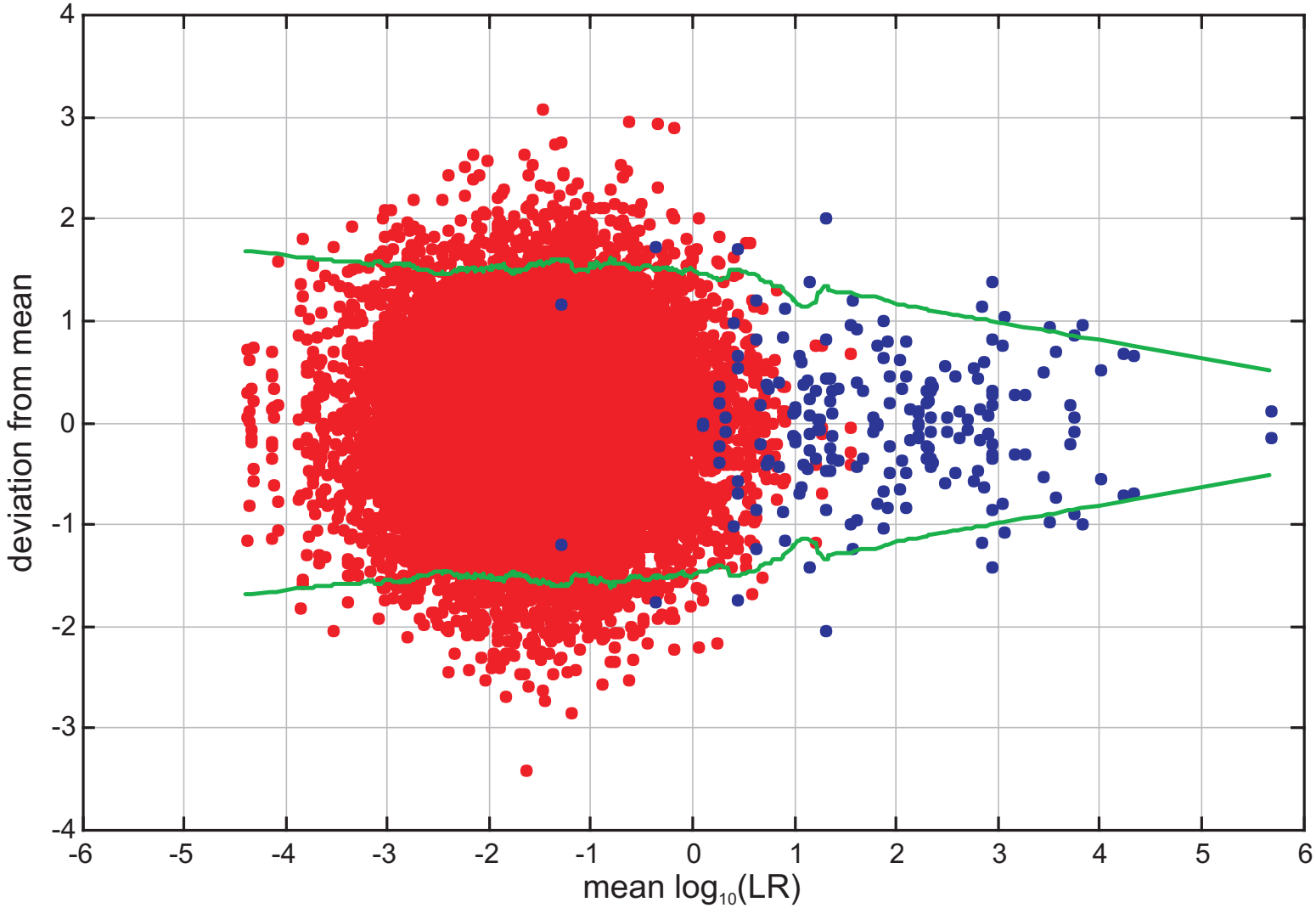
Measuring Reliability

- non-parametric (heteroscedastic)
- local linear regression



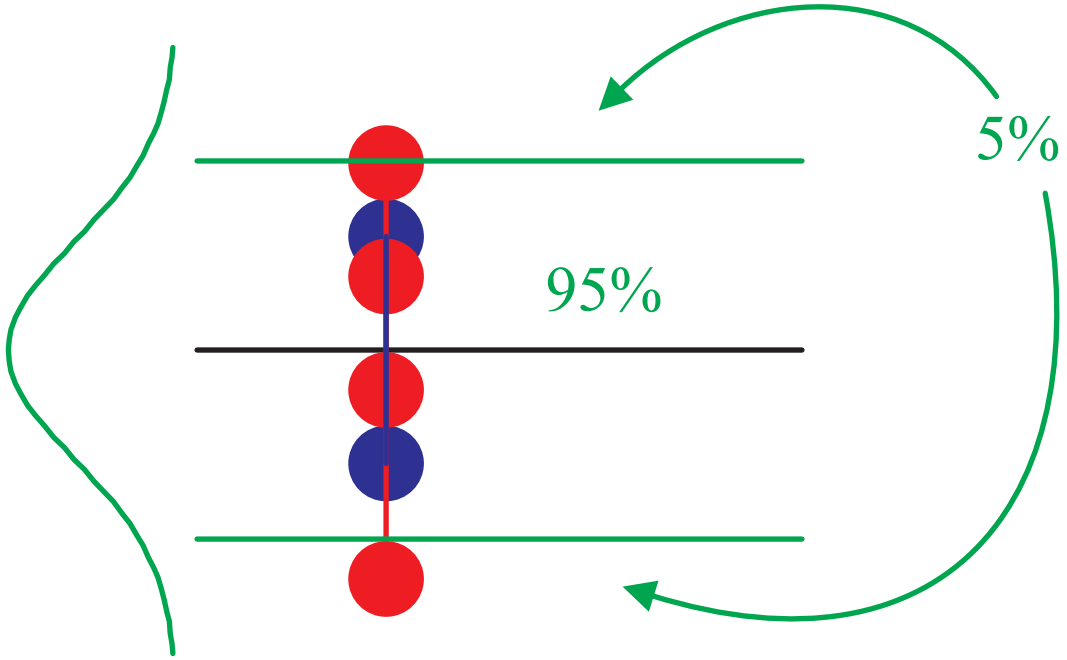
Measuring Reliability

- non-parametric (heteroscedastic)
- local linear regression



Measuring Reliability

← deviation from mean →



- parametric (homoscedastic)
- pooled variance
- *t* distibution
- assume uniform priors

Example of Testing the Validity and Reliability of a Forensic-Comparison System

System and Data (Morrison, Thirivaran, Epps, 2010)

- **Automatic system:**
  - 16 MFCCs (20 ms window, 10 ms overlap) + deltas
  - cumulative density mapping
  - 512 mixture GMM-UBM
  - logistic-regression calibration

- **Databases:**
  - **Background:** 800 recordings from NIST SRE 2004
  - **Calibration:** 2 recordings of each of 32 speakers from NIST SRE 2008 8conv
  - **Test:** 4 recordings of each of 100 speakers from NIST SRE 2008 8conv

Results

- 40 s of speech per offender recording in the test set

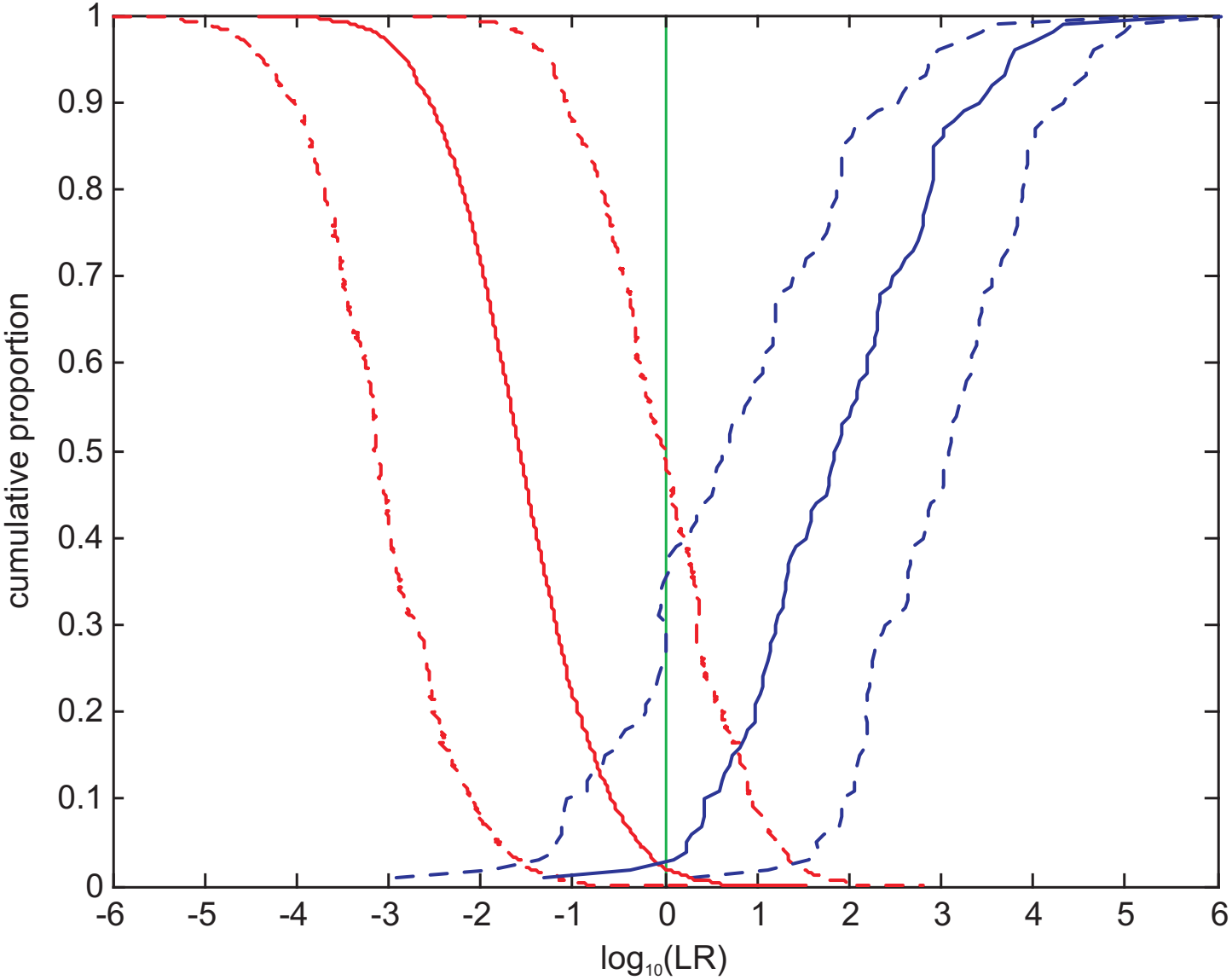
  $C_{llr} = 0.150$ 95% CI (parametric) = ±1.63 $\log_{10}$(LR)

- 20 s of speech per offender recording in the test set

  $C_{llr} = 0.150$ 95% CI (parametric) = ±1.69 $\log_{10}$(LR)

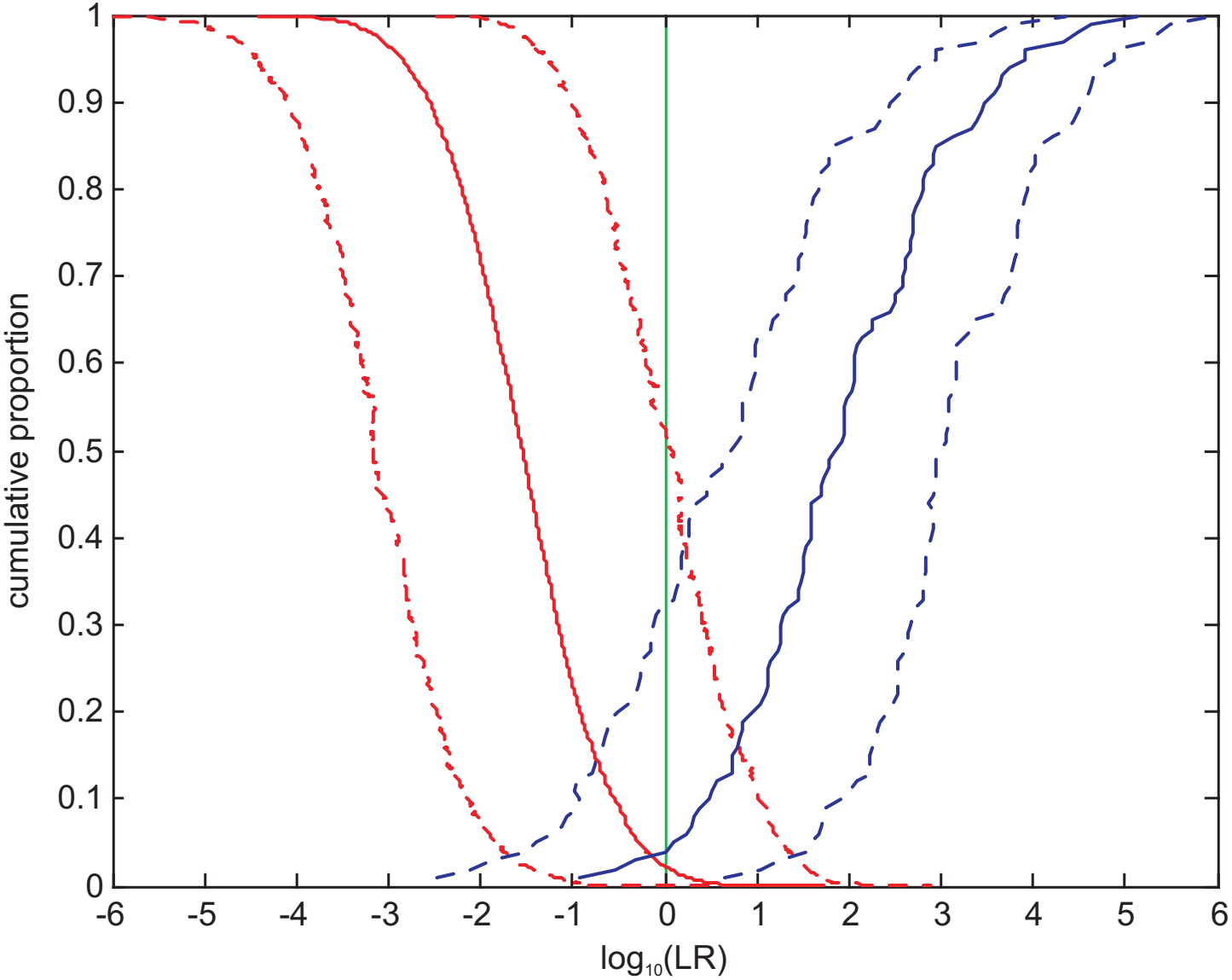
Results

- 40 s of speech per offender recording in the test set



Results

- 20 s of speech per offender recording in the test set



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, say, 100 ($\log_{10}(LR)$ of +2), then the the non-parametric 95% CI estimate would be ±1.17 $\log_{10}(LR)$, and 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 differences between the voice samples if the questioned-voice sample had been produced by the accused than if it had been produced by someone other than the accused.

What this means is that whatever you believed before this evidence was presented, you should now be 100 times more likely than before to believe that the voice on the questioned-voice recording is that of the accused.

Based on my calculations, I am 95% certain that the acoustic differences are at least 7 times more likely and not more than 1450 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.

Latest Thoughts on Measuring the Reliability of a Forensic Comparison System

Measuring Reliability

- In a trial the offender sample is fixed, and precision should be measured given this fixed sample
- Imagine that we have four recordings (A, B, C, D) of each speaker in our test database, and that these are matched to the conditions of the suspect recording from the trial
- Use each recording to build four suspect models for each test speaker
- Calculate likelihood ratios using each suspect model and the fixed offender sample
- Use these likelihood ratios to calculate the precision of the system given the fixed offender sample

Measuring Reliability

- Suspect models from test database compared to fixed offender data from trial

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

Conclusion

Conclusion

- At admissibility hearing (*Daubert*), must supply judge with all relevant information about system performance (validity & reliability a.k.a. accuracy & precision)
- Not to present information about the precision of the system would be to mislead the trier of fact
- Must take account of the speaker level as well as the recording level (akin to activity and source levels)
- Intrinsic variability of voice data (cf. DNA profiles)
- Limited data for suspect models
  – underestimating within-speaker variability
- Limited offender data

Thank You

http://geoff-morrison.net

http://forensic-voice-comparison.net

http://forensic.unsw.edu.au