Probabilistic Statements Regarding Weight of Evidence

Steve Lund and Hari Iyer
Statistical Engineering Division

Disclaimer

Points of view in this presentation do not necessarily represent the official position or policies of the National Institute of Standards and Technology or the U.S. Department of Commerce.

PROBLEM STATEMENT

- A crime has been committed. A sample (Q) from an unknown source is recovered from the crime scene. A reference sample (R) is obtained from a person of interest.
- A forensic examiner compares the Questioned sample (Q) with the Reference sample (R) and reports his/her findings along with other relevant information.

  - H_1: Q and R come from the same source
  - H_2: Q and R come from different sources

- Triers of fact (**TOFs**) are often insufficiently equipped to assess how much influence the presented information should have on their degrees of belief in H_1 and H_2.
- They look to the expert to provide a ‘fit for purpose’ summary that can help in making a fair assessment in this situation.

TWO APPROACHES

(1) Likelihood Ratio as weight of evidence

(2) Approach based on **Classification Methods**

LIKELIHOOD RATIO

- A persuasive motivation comes from **Bayesian decision theory.**
- An expert summarizes the 'weight of evidence' in the form of a (personal) **Likelihood Ratio (LR)**.
- Values of LR greater than 1 support one hypothesis (say H_1) and values smaller than 1 support the other.
- LR has an **intrinsic** meaning to the individual who computed it.
- Each TOF is **expected** to use this LR to convert his/her prior degree of belief regarding H_1 to obtain his/her posterior degree of belief in H_1 based on the evidence.

Classification Methods

- A **classification score** is developed that summarizes the degree of correspondences and discrepancies between features when comparing Q and R.
- Higher values of the score support one hypothesis (say H_1) and lower values support the other.
- Good classification methods can effectively **discriminate** between **mated** and **non-mated** pairs.
- Performance of competing candidate methods are compared using empirically evaluated **error rates**.
- **Receiver Operating Characteristic** (ROC) curves facilitate such comparisons.

What are the strengths and weaknesses of each method?

Likelihood Ratios As Weight of Evidence

- The **motivation for the LR approach** is provided by Bayes Rule

$$\text{Posterior Odds}_{TOF} = \text{Prior Odds}_{TOF} \times LR_{TOF}$$

- However, the **proposed use of Bayes Rule** is described by

$$\text{Posterior Odds}_{TOF} = \text{Prior Odds}_{TOF} \times LR_{Expert}$$

- Can LR values be calculated by one party on behalf of another? (without an actual elicitation exercise).
- Bayesian decision theory is regarded as well suited for an individual decision maker. But **is it suited for use in criminal proceedings** where **experts are not the decision makers but triers of fact are**?
- Is there any **uncertainty associated with the LR provided by the expert and should this be reported**?

Uncertainty Associated with a Likelihood Ratio

Leading proponents of LR claim there is NO UNCERTAINTY associated with a LR.

Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio by Franco Taroni, Silvia Bozza, Alex Biedermann and Colin Aitken, Law, Probability and Risk, March 2, 2016.

Their claim is valid when a LR is used by the person who computed it to help in their own personl decision making situation.

However, a trier of fact should wonder which LR value is to be believed if different experts arrive at vastly different LR values each of which is correct for the expert who computed it.

That is, how 'transferable' is a LR value computed by one expert?

Factors Influencing Assessment of a Likelihood Ratio

A partial list of factors that influence a LR assessment:

(1) Choice of a r**elevant population**

(2) Choice of **priors on members of relevant population**

(3) Choice of **models**

(4) Choice of **statistical methods**

(5) Treatment of **measurement errors**

(6) **Sampling variability and sampling bias**

(7) Choice of **reference databases**

Assumptions Lattice and Uncertainty Pyramid

A single LR value from one expert is not sufficient.

A good faith assessment should be made of the range of **other plausible LR values,** and shared with TOFs.

This will help TOFs assess the **fitness for purpose of the LR value** offered by an expert.

Assumptions Lattice and **Uncertainty Pyramid** provide a step in this direction.

Assumptions Lattice

A1: G_0 is a continuous CDF

A2: G_0 is a continuous, unimodal CDF

A3: G_0 is a continuous, symmetric CDF

A4: G_0 is a continuous, symmetric, unimodal CDF

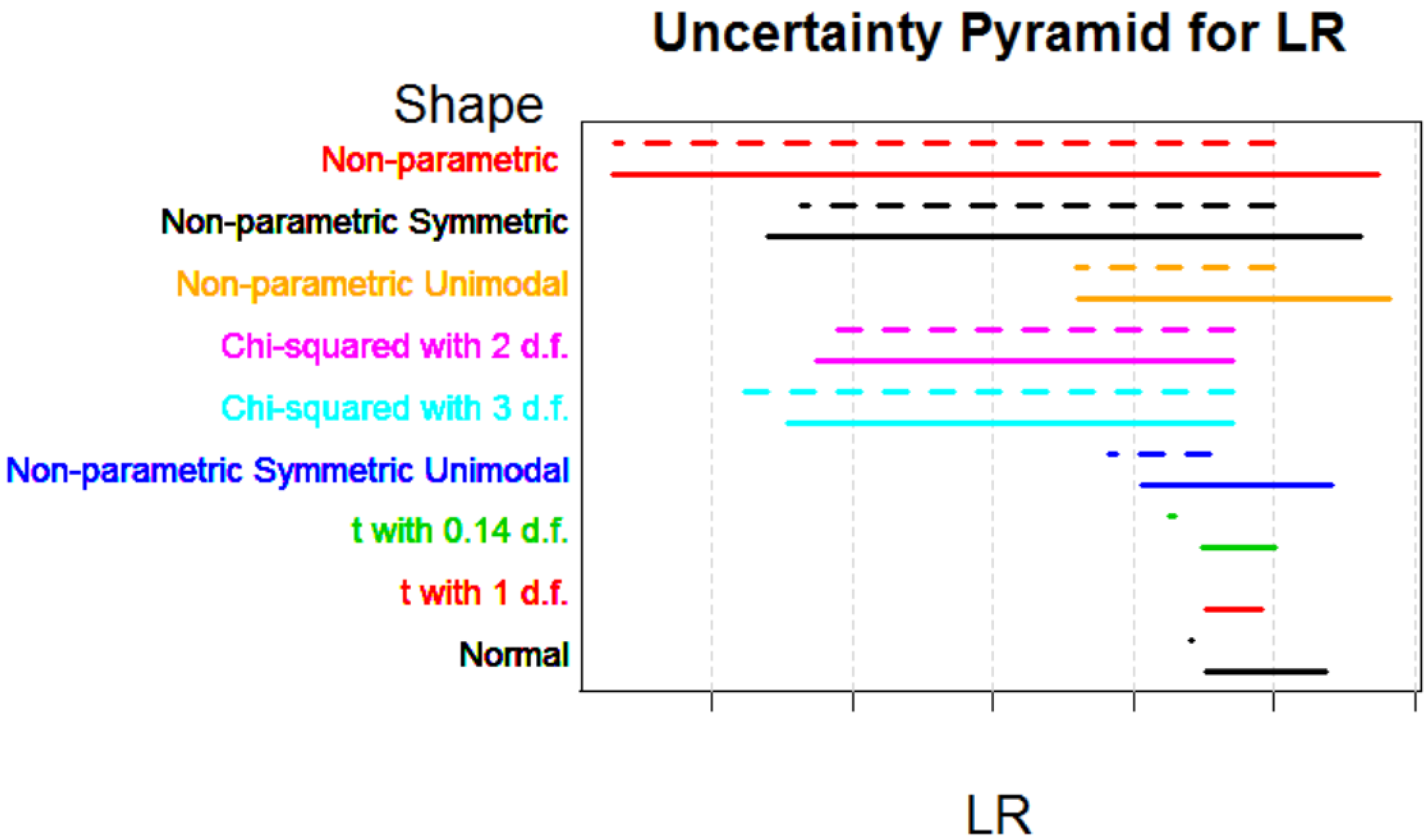
A5: G_0 is the CDF of a scaled t distribution with k degrees of freedom, $1 \leq k \leq \infty$.

A6: G_0 is the CDF of a scaled t distribution with 1 degree of freedom.

A7: G_0 is the CDF of a scaled t distribution with 0.5 degree of freedom.

A8: G_0 is the CDF of a Normal distribution with known standard deviation σ

Uncertainty Pyramid



Classification Methods

- **Reference databases with known ground truth are used** to facilitate an empirical comparison of competing methods.
- Reference databases must be ***sufficiently rich*** to facilitate calculation of **situation relevant error rates**. So **adequacy of reference databases** needs to be addressed.
- Statements regarding **error rates** must include a characterization of specific circumstances for which those error rates apply.
- Some ongoing efforts:

  South Dakota State University, DFSC, CSAFE, SAMSI, NIST

Summary

Likelihood Ratios

Probabilistic assessments (LR) generally involve subjective elements and have unknown degrees of "transferability" **unless** they are accompanied by a sufficiently comprehensive assessment of uncertainty (for instance ***assumptions lattice*** and ***uncertainty pyramid***) to assist triers of fact determine the fitness for purpose of the LR value presented.

Classification Scores

- Performance of classifiers can be empirically evaluated and the results made publicly available. The empirical findings are ***demonstrable*** and ***falsifiable***.
- With sufficiently rich databases, situation relevant empirical error rates can be computed.
- Suitable empirically determined ROC curves can assist triers of fact assess the probative value of a classification score.

INFORMATION TECHNOLOGY LABORATORY

THANK YOU

National Institute of Standards and Technology
U.S. Department of Commerce