Forensic Statistics: Ready for consumption?

Richard Gill
Mathematical Institute, Leiden University

http://www.math.leidenuniv.nl/~gill

In a nutshell (I)

Everyday statistics: The role of a statistician in research and consultation ... Two way interaction, adapting models to findings, adapting questions to findings. Two popular paradigms: frequentist, Bayesian. Pros and cons; modern pragmatic synthesis (not a dichotomy but a spectrum). Different applications require a different place in the spectrum (or even a move in another dimension).

Statistics in the court room is however not everyday statistics. Present consensus in forensic statistics: the statistician should merely report the likelihood ratio (LR). This because combining information and drawing conclusions is the job of the jury/the judges. The statististician must just report what her expertise tells her about the question put her by the judge (statistics: modelling/interpreting/learning from chance). NB difference between statistics in police criminal investigation and in the court room.

Problems with LR:

- who determines the hypotheses?
- which data?
- must the defense specify/accept a hypothesis?
- importance of how the data was obtained: evidence = message + messenger
- composite hypotheses
- posthoc hypotheses
- interpretation, dangers [ignorance=uniform probability? 3 doors problem. Lucia]

In a nutshell (II)

Examples:

1.) DNA matching. Database-search controversy

2.) Forensic glass; modelling of between and within source variatie (Aitken et al.)
We need to develop (empirically calibrated) likelihood ratio
(solve curse of dimension: empirical Bayes?, statistical learning? targeted likelihood)

3.) Lucia de B. shift-roster data

4.) Tamara Wolvers case: combination of various (poor) DNA traces

In each of the examples, even the simplest, I'll show that there are a lot of problems with the LR approach. Big challenges (both from legal and statistical point of view). Two-way interaction is necessary, preferably before we meet in the court-room!

References:

Robertson and Vignaux: don't teach statistics to lawyers!
Seeking truth with statistics:
http://plus.maths.org/latestnews/may-aug04/statslaw/index.html
Meester & Sjerps: Database search controversy and two-stain problem
Sjerps: Statistiek in de rechtszaal. Stator. http://www.kennislink.nl/web/show?id=111865

Everyday statistics

- Intensive two-way interaction between statistician and subject-matter expert (client)

  Cyclic process of re-evaluation of data/ models/questions

or

- Use of standard methodology in standard situation where the user knows what "standard" means (2 ×)

cf. 3 door problem;
Probiotica research;
Prosecutors and defence-attorney's fallacy

Not in the court-room

- Classical (frequentistic) statistics:

    significance tests

    confidence intervals

    p-values ...

  are neither appropriate nor understood

- Bayesiaanse (subjective) statistics is too complex, not appropriate (illegal)

- No place for discussion with subject-matter expert

What are we left with?

- Likelihood ratio (LR): numerical expression of "weight of evidence"

- LR = Prob (evidence | prosecution)

  ÷ Prob (evidence | defense)

- Bayes theorem:

  posterior odds

  = prior odds

  × LR

Bayes, sequential

- posterior odds (given A, B, C) =

  prior odds × LR for A, B, C

- LR for A, B, C

  = LR for A

  × LR for B given A

  × LR for C given A, B

extend to *tree* and then to marginalisation and conditioning in arbitrary trees – Bayes nets

Example 1: DNA match

- Chance of profile “A” is 1 in 5,000
- DNA perpetrator (“*crime stain*”) has profile “A”
- DNA suspect has profile “A”
- Prob(match | perpetrator profile, prosecution) = 1
- Prob(match | perpetrator profile, defence) =

  1 / 5,000

- LR= P(data | H_P) / P(data | H_D)=5,000

DNA match after "database search"

- Suspect found in data-base of 5,000 people, in which he is the only match
- Prob. of a unique match is approx. e^{-1}, "weight of evidence" is about 2.7
- LR of 5,000 was for a "post-hoc" hypothesis

Alternative LR for DNA match

- Compute simultaneous probability of *all* profiles in database *and* "crime-stain" under two hypotheses (perpetrator in / not in database)
- LR = quotient of these two probs

(in our case: a unique match, profile "A")

LR =

1 / size database × frequency profile "A"

= 1

[but if database = whole population?!]

DNA match: 1 or 2.7 or 5,000 !?

- What is “the evidence” ?
- What are the hypotheses?
- Meester and Sjerps: the “a priori” chance that the suspect is the source of the DNA in the crime-stain is very different when he was found from the database, than when he was already a suspect! It’s not the statistician’s job to specify these prior probabilities!

(posthoc problem)

- The LR for a post-hoc hypothesis is only meaningful in a *total* Bayesian approach
  [cf. lottery winner]
- The "evidence" is not just the *DNA match* but also the reason why the match was found – the message + messenger! [Indeed: *missing* evidence is also evidence!]
- The LR should be determined on the basis of a priori specified hypotheses and for carefully described "evidence"; only then is it interpretable
  [a LR of 5,000 occurs less than once in 5,000 times, if H_D is true]

Example 2: Forensic glass

- Database: measurements of elemental composition of glass fragments (% Si, Na, Al, ...)

  within source and *between source* variation

- Case: 2 samples: fragment(s) broken window pane at scene of crime, fragment(s) in the suspect's clothing

- Combine *similarity* of the 2 samples with their *rarity* in the light of other samples (cf. database)

cf: LCN and incomplete DNA-profile; signatures and handwriting; fingerprints; texts; extasy pills; ...

Forensic glass

- prosecution: 2 fragments same pane
- defence: 2 fragments different panes

- Aitken et al.: *estimate* LR = $p(x,y)/p(x)p(y)$ with advanced applied statistical methodology ...

Forensic glass

This can be simplified slightly so that the numerator of the LR

$$\frac{1}{m}(2\pi)^{-p}\left|\frac{U}{n_c}+\frac{U}{n_r}\right|^{-1/2}\left|C+\frac{U}{n_c+n_r}\right|^{-1/2}|h^2C|^{-1/2}$$
$$\left|(C+\frac{U}{n_c+n_r})^{-1}+(h^2C)^{-1}\right|^{-1/2}$$
$$\exp\left\{-\frac{1}{2}(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{y}}_2)^T\left(\frac{U}{n_c}+\frac{U}{n_r}\right)^{-1}(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{y}}_2)\right\}$$
$$\sum_{i=1}^{m}\exp\left\{-\frac{1}{2}(\bar{\boldsymbol{y}}_{12}-\bar{\boldsymbol{x}}_i)^T\left[\left(C+\frac{U}{n_c+n_r}\right)\right.\right.$$
$$\left.\left.+(h^2C)\right]^{-1}(\bar{\boldsymbol{y}}_{12}-\bar{\boldsymbol{x}}_i)\right\}$$

The first term in the denominator is

$$\int f(\bar{\boldsymbol{y}}_1|\mu)f(\mu)\,\mathrm{d}\mu$$
$$=\frac{1}{m}(2\pi)^{-p/2}\left|C+\frac{U}{n_c}\right|^{-1/2}|h^2C|^{-1/2}\left|\left(C+\frac{U}{n_c}\right)^{-1}+(h^2C)^{-1}\right|^{-1/2}$$
$$\sum_{i=1}^{m}\exp\left\{-\frac{1}{2}(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{x}}_i)^T\left[\left(C+\frac{U}{n_c}\right)+(h^2C)\right]^{-1}(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{x}}_i)\right\}$$

The second term in the denominator is

$$\int f(\bar{\boldsymbol{y}}_2|\mu)f(\mu)\,\mathrm{d}\mu$$
$$=\frac{1}{m}(2\pi)^{-p/2}\left|C+\frac{U}{n_r}\right|^{-1/2}|h^2C|^{-1/2}\left|\left(C+\frac{U}{n_r}\right)^{-1}+(h^2C)^{-1}\right|^{-1/2}$$
$$\sum_{i=1}^{m}\exp\left\{-\frac{1}{2}(\bar{\boldsymbol{y}}_2-\bar{\boldsymbol{x}}_i)^T\left[\left(C+\frac{U}{n_r}\right)+(h^2C)\right]^{-1}(\bar{\boldsymbol{y}}_2-\bar{\boldsymbol{x}}_i)\right\}$$

cf. master-thesis Sonja Scheers

Forensic glass

- Challenging statistics (high dimensional compositional data, many zero’s; parametric? non-parametric?)
- At their best, the models are a rough approx.
- The data-base is not really a random sample...
- In the situation when the evidence counts, we are making a gross extrapolation
- Need: validation, calibration.
  Sufficiency: the likelihood ratio of the likelihood ratio is itself. So the empirical likelihood ratio of the likelihood ratio should be itself!

Forensic glass

- Sufficiency: the likelihood ratio of the likelihood ratio is itself!
- Proposal: "estimate" the likelihood ratio anyway you like
- It's a function of the *2 samples* (<u>crime scene</u>, <u>suspect</u>)
- Use the data-base to *sample LR's* under both hypotheses (prosecution, defense: H_P , H_D)
- Estimate the ratio of the densities of the two sampled LR's (which should be monotone)
- Test the hypothesis of monotony

Forensic glass

- Estimation, testing is based on greatest convex minorant of the QQ plot of sample under H_P against the combined sample $H_P + H_D$
- Proposal: “estimate” the likelihood ratio anyway you like
- It’s a function of the *2 samples* (crime scene, suspect)
- Use the data-base to *sample LR’s* under both hypotheses
- Estimate the ratio of the densities of the two sampled LR’s (which should be monotone)
- Test the hypothesis of monotony using non-parametric generalised likelihood ratio test

Example 3: Lucia

Shifts	Incident	No inc.	Total
Lucia	**9**	**133**	142
No L.	**0**	**887**	887
Total	9	1020	1029

Original data

- Fisher exact test

  p = 15 per billion

- Binomial test (days w. incident & L.)

  p = 50 per million

Corrected data

- Fisher exact test

  p = 0.2 pro mille

- Binomial test (days w. incident & L.)

  p = 4 %

Shifts	Incident	No inc.	Total
Lucia	**7**	**135**	142
No L.	**4**	**883**	887
Total	11	1018	1029

- Heterogeneity model, JKZ+RKZ, p = 5%

Lucia: problems

- The data: "selection bias", definition "shift w. incident" – *blinding*?
- [Bayes vs. frequentistic]
- LR: specification hypotheses prosecution, defence? Post-hoc!
- The notion of "chance" is not unequivocal; "ignorance" does not guarantee "*pure* chance"
- Information from other periods in same ward?

Lucia: epidemiological, causal thinking

- Clusters of incidents between long incident-less periods seems to be the *norm*
- Shifts follow a regular pattern

  so if one incident “hits” your shifts it is likely there’ll be more (In Lucia case, 7=2+2+3 incidents belonged to 3 children)

- Serious empirical research into the “normal situation” has *never, ever,* been done!
- World-wide epidemic of *collapsed cases*

Example 4

- Tamara Wolvers: three separate kinds of DNA evidence
- Three separate forensic reports, in each case “the DNA profile does not *exclude* the suspect”
- Neither prosecution nor judge could combine the three match chances (can it be done?? ...)
- The suspect went free
- No “control” measurements (what is normal?)

Conclusion

- Statistics in court is *still far from* everyday statistics; it is challenging and important for lawyers and statisticians
- For the time being: use in detection rather than proof?

Appendix:

Bayes nets, the solution of everything ?

- Bulldozer-ram-robbery
- Sweeney case

Bayes net/graphical model: quantitative combination of (sometimes contradictory) evidence of varying character

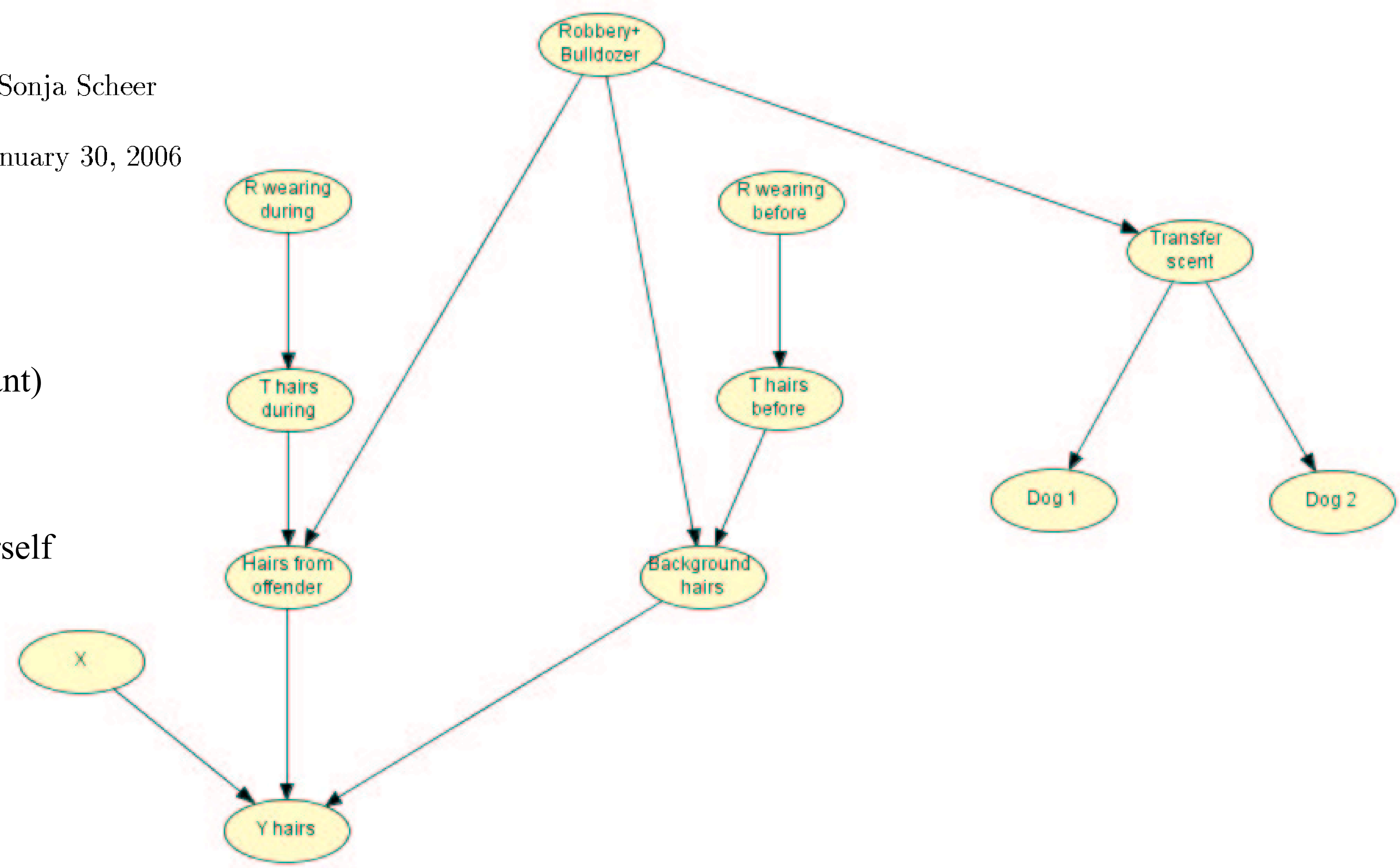
Compute likelihood ratio for complex composite evidence, taking account of dependence and independences (Taroni, Aitken, Dawid, ...)

Bulldozer-ram-robbery

The use of Bayesian networks for combining forensic evidence in a Dutch criminal case

Sonja Scheer

January 30, 2006



<u>Hierarchy of propositions</u>:
source (the stain is from the defendant)
activity (contact, transfer)
crime (guilt, innocence)

The forensic statistician restricts herself to source and activity

Conclusion: ... taught us much, but unsatisfactory

Kevin Sweeney case

The probability that Kevin Sweeney murdered his wife ... is very small indeed

Richard Gill, Aart de Vos

University Leiden, Free University Amsterdam
Draft discussion paper

March 25, 2008

It was a warm summer night in 1995. Kevin Sweeney left his wife Suzanne Davies at their new home in Steensel (near Eindhoven) at 02:00 a.m. Between 02:47 and 03:00, two policemen and the housekeeper walked all around the house not noticing anything, in response to a burglar alarm at the alarm centre. At about 03:45 a fire was reported – clients still on sitting on the terrace of the café across the road saw flames in the upstairs bedroom window. Firemen arrived at 03:55. Suzanne Davies was pronounced dead at 04:37 by carbon monoxide poisoning. Many facts were unclear, but the main riddle is the time span if Kevin set the fire alight before 2.00. House room fires start rapidly. In 6 attempts by TNO (using petrol and a naked flame) the fire spread within 5 minutes. But also fires started by a discarded cigarette start very rapidly.

T	P(T|I)	P(T|G)	likelihood ratio P(T|G)/ P(T|¬ G)	If prior Odds **10** Post odds	P(G|T)	P(G|T)× P(T|I)
2:00						
2:15	3.0E-09	0.9	5.4	54	0.982	2.9E-09
2:30	5.9E-08	0.09	0.54	5.4	0.844	5.0E-08
2:45	1.2E-05	0.009	0.054	0.54	0.351	4.2E-06
3:00	4.8E-04	0.0009	0.0054	0.054	0.051	2.4E-05
3:15	4.8E-02	0.00009	0.00054	0.0054	0.005	2.6E-04
3:30	9.5E-01	0.000009	0.000054	0.00054	0.001	5.1E-04
					P(G|I)	**0.080%**

See also A. Derksen (2008), *Het OM in de Fout*

Kevin Sweeney case

Het 'vergeten' tijdspad.

De anatomische ontleding van een bewijscorpus voor moord door brandstichting; met het 'scheermes' van Ockham.

F.W.J. Vos, 17 mei 2008

Distinguish between definite primary *observation* and secondary *interpretations* thereof;
also the observations which *ought to have been there* ...
showed that our *Bayes net* was based on completely wrong ideas (forensic fire-expert F. Vos).

F. Vos: <u>all observation</u> compatible with a completely "normal" accident

Needed: *expert* combination of fire-forensic, chemical, pathological, toxicological evidence

Conclusion: ... *if you need statistics*... ?