Embracing Bayesian inference in metrology

Tony O’Hagan

Outline

- Motivation
  - To Bayes or not to Bayes?
  - Who is it for?
- Horses for courses
  - Internal
  - Technical
  - Scientific
  - Statutory/legal

Disclaimer

- I’m a statistician
  - Not a metrologist
- I’ve been listening to metrologists
  - Who have helped me gain some understanding of the field
  - But any blunders in this talk are my own fault

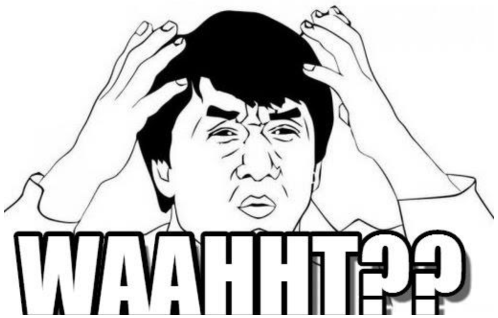
Motivation

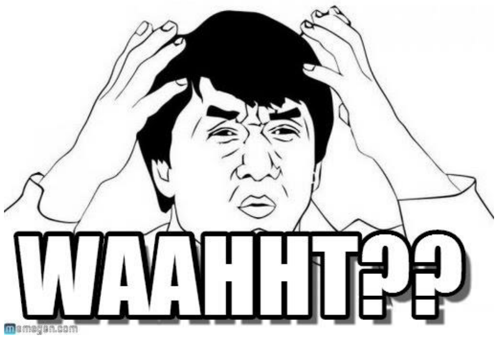
To B or not to B?

- That is the question!
  - Long-standing debate in Metrology
  - Many metrologists resistant to B(ayes)
    - Preferring traditional frequentist methods
  - GUM explicitly adopting Bayesian framework
  - But is Bayes the right way?
- Of course!
  - I’m a committed Bayesian
    - So I would say that!
  - Here are some slides I presented at MATHMET 2014
    - My usual uncompromising kind of talk!

What the VIM3 says

- Measurement is a process
  - When applied in a particular instance it produces
  - A measurement result
  - For the measurand of interest
- The measurement result (2.9)
  - Set of values attributed to the measurand
  - Could be a pdf (Note 1)
  - Could be a measured quantity value + measurement uncertainty (Note 2)
- Measurement uncertainty (2.26)
  - Non-negative value characterising dispersion
  - Could be standard deviation/uncertainty or … (Note 2)





Stop messing about!

- The measurement result has to be a **distribution** (pdf)
- Anything less is a cop-out and inadequate
  - Estimate + std uncertainty is only a **partial** measurement result
    - Doesn't tell the recipient how likely it is for the measurand to exceed some critical value
    - Doesn't allow the recipient to compute a credible interval
    - Can only be propagated by crude approximations
    - Can be highly misleading when uncertainty is asymmetric (skewed)
- Probability is the proper quantification of uncertainty
  - And for a variable we require a probability distribution
  - The probability distribution of the measurand
    - Based on the available information and the measurement process

Once we have the distribution

- We can derive summaries
  - Like an estimate,
    - The mean of the distribution or the median
  - The standard deviation
    - Or some other summary of uncertainty such as the IQR
  - Credible intervals
  - Probabilities below thresholds
- They are just for *communication*
  - The result is the distribution
  - But summaries help to communicate relevant features for the recipient

What is a probability?

- Curiously, the VIM does not define the term!
- In the GUM there are two kinds of probability (3.3.6)
  - Type A standard uncertainty is obtained from a **probability density function** derived from an **observed frequency distribution**
  - Type B standard uncertainty is obtained from an assumed probability density function based on the degree of belief that an event will occur [often called subjective **probability**]
- In GUM Supplement 1 there is only one (Introduction)
  - The PDF for a quantity expresses the state of knowledge about the quantity, i.e. it quantifies the degree of belief about the values that can be assigned to the quantity based on the available information.

What kind do we need?

- The measurement result is a probability distribution for the measurand
- The probabilities have to be degrees of belief
  - Frequency probability only applies to things that are repeatable
    - At least in principle, we have to be able to observe many instances and count how often particular values arise
  - The measurement result can only be based on frequency probability if the measurand is repeatable
    - But the measurand has a fixed, unique value – it's not random
    - We may be able to make repeated *measurements* of it, but that's a completely different thing
- Every measurement result must be expressed using subjective probability

Typically uncompromising!

- Uncertainty about the measurand must be expressed using subjective probability
  - And hence must be formulated within the Bayesian paradigm
- The proper expression of uncertainty is a (subjective) probability distribution
  - Necessarily involves subjective judgement (prior)
- To state an estimate and a single uncertainty measure is inadequate
  - Useful as summaries but incomplete
  - Only acceptable when presented *in addition* to the complete pdf
- But …
  - This talk is going to be *atypical*

Who is it for?

- For whom are we making measurements?
- I believe there are four distinct cases to consider
  - Internal
    - Measurement for metrologists
  - Technical
    - Measurement for commercial users
  - Scientific
    - Measurement for scientists
  - Statutory/legal
    - Measurement for the courts

Why does it matter?

- Bayesian inference is a framework for coherent personal judgements
  - Prior distribution represents personal knowledge a priori
    - Likelihood is also a personal judgement
  - Posterior distribution represents personal knowledge a posteriori
    - Synthesising prior information and data
- It is the proper basis for decision-making
  - Based on decision-maker's personal knowledge and utilities
- Questions arise over reporting your personal judgements to somebody else
  - Why should they be interested?

Horses for courses

Internal – for metrologists

- Inter-lab comparisons are done by metrologists for metrologists
  - Decisions are made regarding the quality and consistency of work by individual labs
- This should be done using full Bayesian analysis and decision theory
  - Frequentist inference is unsuitable for all the usual reasons
  - Prior distributions should be carefully formulated
    - Using historic performance
    - With agreement of the community

Technical – for commercial users

- The bulk of routine, individual measurements are made for the use of others
  - Often commercial organisations
  - Full Bayesian analysis is appropriate here, too
    - Because the user wants the best available judgment and will accept the metrologist's analysis
  - Prior distributions should be based on the metrologist's own experience and judgement
    - All relevant knowledge should be incorporated
- A full posterior distribution is the measurement
  - With suitable summaries
  - Metrology needs to adapt to this
  - Will require new forms of certification

Calibration – internal or technical

- Calibration tasks may be internal
  - Developing one-off calibration curves for an instrument or material that will be used many times for measurement
- Or technical
  - Calibration for another end user
- In both cases a full Bayesian analysis is correct
  - Reporting the posterior distribution plus summaries
  - When using it for measurement, it's necessary to incorporate uncertainty in the calibration curve
    - As described by the posterior distribution
    - When reporting the (posterior distribution of) an individual measurement
  - Methods may need to be developed for this

Scientific – for scientists

- Another class of problems involve measurement of physical constants
  - As a contribution to science
- Bayesian analysis is still appropriate
  - Metrologist's prior distribution should be justified
    - from available knowledge
  - But has no special weight
- Should be reported in such a way that other scientists may substitute alternative priors
  - Or analysis given for a range of plausible priors

Statutory/legal – for the courts

- Finally, some measurements may be made for statutory purposes, or for use in court as forensic evidence
  - Metrologist is typically not allowed to use his/her own prior knowledge
    - The data must “speak for themselves”
    - Even though this may mean some knowledge is not used in evidence
- This is not an inference problem
  - Neither Bayesian nor frequentist inference should be used
  - What is needed is the likelihood function, or likelihood ratio
  - But in practice, Bayesian methods will still be needed
    - Integrating over other unknowns, such as σ^2
    - Bayes factor

Proper priors

- It is essential to think about prior distributions
  - In principle, only proper priors are permitted
- Conventional, so-called uninformative priors are dangerous
  - E.g. maximum entropy priors
- They can be employed as approximations to weak, but proper, prior distributions
  - But only when there is no doubt that they yield adequate approximations to the posterior distribution
  - It is dangerous to just assume that this will be the case!
  - It is important to verify robustness

Summary

In my opinion

- Bayesian analysis should always be used
  - Frequentist methods are never appropriate
  - Statutory/legal measurements may appear to be an exception
    - Because they do not require inference as such
    - But in practice we still have to adopt a Bayesian approach
- Prior information needs care and thought
  - Strictly, priors should always be proper
    - E.g. elicited expert judgement
  - So-called uninformative priors may be used as approximations but they are dangerous
- Prior distributions should be appropriate to the context
  - Different criteria for the four different contexts