Data quality, combining multiple data sources, and distribution fitting

Selected points and illustrations!

Dr Marion Wooldridge
26th July 2002

Outline of talk

- Data quality
  - what do we need?
- Combining multiple data sources
  - when can we do it?
  - an example explained
- Fitting distributions
  - how do we do it?
  - an example explained
What do we mean by data?

- Experimental results
  - numerical
- Field data
  - numerical
- Information on pathways/processes
  - words and numbers
- Expert opinion
  - words to convert to numbers?!

Why do we need data?

- Model construction
  - risk pathways
    - what actually happens?
      - do batches get mixed?
      - is the product heated? etc….
- Model inputs
  - estimate probabilities
  - estimate uncertainties
  - estimate variability
Which data should we use?

- The best available!

However…….

- A risk assessment is pragmatic
- A risk assessment is 'iterative'
- There are always data deficiencies
  - These are often 'crucial'

So what does this tell us?…..

What do we mean by pragmatic?

- Purpose of risk assessment…..

- To give useful info to risk manager
  - To aid decisions, usually in short term

- This leads to time constraints

- This generally means using currently available data for decisions ‘now’
  - However incomplete or uncertain that data
What do we mean by ‘iterative’?

- Often several stages, with different purposes
  - ‘Is there really a problem?’
    - rapid preliminary assessment
  - ‘OK, so give me the answer’
    - refine assessment
  - ‘Did we get it right?’
    - revisit assessment if/when new data

The minimum data quality required…..

- will be different at each stage
- so - why is that?…..
‘Is there really a problem?’

- Risk manager needs rapid answer
  - ‘preliminary’ risk assessment
  - Identify data rapidly available
    - may be incomplete, anecdotal, old, from a different country, about a different strain, from a different species, etc. etc……

- Still allows decision to be made
  - a problem…..
    - is highly likely - do more now!
    - may develop - keep watching brief
    - is highly unlikely - but never zero risk!

Stage 1: Conclusion...

- Sometimes, poor quality data may be useful data
  - is it ‘fit-for-purpose’
  - Don’t just throw it out!
OK, so give me the answer

- **refine risk assessment**
  - set up model
    - data on risk pathways
    - model input data
      - utilise ‘best’ data found, identify ‘real’ gaps & uncertainties
      - elicit expert opinions
  - data should be ‘best’ currently available
    - with time allowed to search all reasonable sources
      - still often incomplete and uncertain
      - checked by peer review

So the risk is…….

- The best currently available data gives….  
  - *the best currently available estimate*
- Which is….  
  - *the best the risk manager can ever have ‘now’!*
- Even if…  
  - *it is still based on data gaps and uncertainties*
    - *which it will be!*
    - the ‘best’ data may still be poor data
    - allows targeted future data collection - may be lengthy
Stage 2: Conclusion….

- If a choice
  - Need to identify the best data available
    - What makes it the best data?
  - What do we do if we can't decide?
    - Multiple data sources for model!
- What do we do with crucial data gaps?
  - Often need expert opinion
    - How do we turn that into a number?

‘Did we get it right?’

- revisit assessment if/when new data available
  - targetted data collection
    - may be years away….
    - but allows quality to be specified
      - not just ‘best’ but ‘good’
    - will minimise uncertainty
    - should describe variability
What makes data ‘good’?

- Having said all that - some data is ‘good’!
- Two aspects
  - Intrinsic features of the data
    - universally essential for high quality
  - Applicability to current situation
    - data selection affects quality of the assessment and output
- General principles......

High quality data should be....

- Complete
  - for assessment in hand
  - required level of detail
- Relevant
  - e.g. right bug, country, management system, date etc. etc.
  - nothing irrelevant
- Succinct and transparent
- Fully referenced
- Presented in a logical sequence
High quality data should have….

- Full information on data source or provenance
  - Full reference if paper/similar
  - Name if *pers. comm.* or unpublished
  - Copy of web-page or other ephemeral source
  - Date of data collection/elicitation
  - Affiliation and/or funding source of provider

High quality data should have….

- A description of the level of uncertainty and the amount of variability
  - Uncertainty = lack of knowledge
  - Variability = real life situation

- Units
  - Where appropriate

- Raw numerical data
  - Where appropriate
High quality study data should have....

- Detailed information on study design: e.g.
  - Experimental or field based
  - Details of sample and sampling frame including:
    - livestock species (inc. scientific name)/product definition/population sub-group definition
    - source (country, region, producer, retailer, etc.)
    - sample selection method (esp for infection: clinical cases or random selection)
    - Population size
- Season of collection
- Portion description or size
- Method of sample collection

High quality microbiological data should have....

- Information on microbiological methods including:
  - Pathogen species, subspecies, strain
    - may include antibiotic resistance pattern
  - tests used, including any variation from published methods
  - sensitivity and specificity of tests
  - units used
  - precision of measurement
High quality numerical data should have....

- Results as raw data
  - Including:
    - Number tested
    - Results given for all samples tested
    - For pathogens, number of micro-organisms
      - Not just positive/negative.

A note on comparability....

- High quality data sets can easily be checked for comparability
  - As they contain sufficient detail
    - Are they describing or measuring the same thing?
    - Are the levels of uncertainty similar?

- Poor quality data sets are difficult to compare
  - Lack detail
    - May never know if describing or measuring same thing!
    - Or may be exactly same data!!
    - Difficult/unwise to combine!
and on homogeneity…..

- Homogeneity of input data aids comparability of model output
- Homogeneity achieved by, e.g.
  - Standardised methods
    - For sampling, testing etc.
  - Standard units
  - Standard definitions
    - For pathogen, host, product, portion size, etc.

Multiple data sources - why?

- No data specific to the problem - but several studies with some relevance
  - E.g. one old, one a different country etc
- Several studies directly relevant to the problem - all information needs inclusion
  - E.g. Regional prevalence studies available; need national
- Expert opinion - but a number of experts
  - And may need to combine with other data
Multiple data sources….how?

- model all separately
  - ‘what-if’ scenarios
  - several outputs
  - for data inputs and ‘model uncertainty’

- fit single distribution
  - point values

- weight alternatives
  - equally or differentially?
  - expert opinion on weights?
  - sample by weight distribution - single output
  - for data inputs and ‘model uncertainty’

Weighting example

- Overall aim:
  - To estimate probability that a random bird from GB broiler poultry flock will be campylobacter positive at point of slaughter

- Current need:
  - To estimate probability that a random flock is positive

- Data available:
  - Flock positive prevalence data from 4 separate studies

Probability flock is positive, $P_{fp}$, per data set

- Raw data used
- $P_{fp}$ per data set
  - 'standard' method
  - Beta distribution
    - $r =$ positive flocks
    - $s =$ flocks sampled
  - $P_{fp} = \text{Beta}(r+1,s-r+1)$
  - for P1, P2, P4
  - describes uncertainty per data set

Data sets:
- 2 x poultry companies
- 1 x colleagues (published) epidemiological study
- 1 x another published study

Other data available:
- Market share by total bird numbers for studies 1,2,3

Assumption made:
- study 4 was 'rest of market'
- approximation

Weighting method:
- uses all available info
- better than 'equal weighting'
So - weights assigned….

- based on market share per study
  - $P_1 + P_2 = 35\%; \ w_1 + w_2 = 0.35$
  - combined here as confidential data
  - $P_3 = 50\%; \ w_3 = 0.50$
  - $P_4 = 15\%; \ w_4 = 0.15$

- probability random flock positive, $P_{fp}$

$$P_{fp} = (P_{1_{fp}} \ w_1) + (P_{2_{fp}} \ w_2) + (P_{3_{fp}} \ w_3) + (P_{4_{fp}} \ w_4)$$

So - when to combine?

- Depends on info needed
  - in example, for random GB flock
  - must combine
  - but - loses info

- Depends on info available
  - and assumptions on relevance/comparability
  - e.g: two studies 10 years old and 5 years old:
    - if prevalence studies - leave separate? Trend info? ‘What-if’ scenarios better?
    - if effects of heat on organism - will this have altered? Check method, detection efficacy, strain etc - but age of study per se not relevant

- Assessors judgement
Fitting distributions

- Could fill a book (has done!)
- Discrete or continuous?
  - Limited values 'v' any value (within the range); e.g.
    - number of chickens in flock - discrete (1,2,3... etc)
    - chicken bodyweight - continuous (within range min-max)
- Parametric or non-parametric?
  - Theoretically described 'v' directly from data; e.g.
    - incubation period - theoretically lognormal - parametric
    - percentage free-range flocks by region - direct from data
- Truncated or not?
  - For unbounded distributions - e.g. incubation period > lifespan?

Where do we start?

- Discrete or continuous?
  - Inspect the data
    - whole numbers = limited values = discrete
- Parametric or non-parametric?
  - Use parametric with care! - usually if:
    - theoretical basis known and appropriate
    - has proved to be accurate even if theory undemonstrated
    - parameters define distribution
  - non-parametric generally more useful/appropriate
- For biological process data
  - for selected distribution - biological plausibility (at least!)
  - gross error check.....
Distribution fitting example...

- **Aim:**
  - To estimate, for a random GB broiler flock at slaughter, its probable age

- **Data available:**
  - Age at slaughter, in weeks, for a large number of GB broiler flocks

  - Funded by, and work undertaken in, The Risk Research Department, The Veterinary Laboratories Agency, UK.

Sequence of steps: 1. Decisions....

- **Discrete or continuous?**
  - Inspect data...age given in whole weeks - but:
    - no reason why slaughter should be at specific weekly intervals
    - time is a *continuous* variable

- **Parametric or non-parametric?**
  - No theoretical basis
  - non-parametric more appropriate

- **Other considerations**
  - Slaughter for meat
    - shouldn’t be too young (too small) or too old (industry economics)
    - suggests bounded distribution
2. What does the distribution look like?

- Scatter plot drawn…..
- Suggests triangular ‘appearance’

![Scatter plot with triangular appearance](image)

3. Is this appropriate?

- Triangular distribution is..
  - continuous
  - non-parametric
  - bounded
- Sounds good so far!
  - 'checked' in BestFit (Palisade)
    - only AFTER logic considered
    - gave Triang as best fit
Comparison of Input Distribution and Triangular distribution

4. Check: Rank 1 using Kolmogorov-Smirnov test in BestFit

Conclusion.....

- Use Triang in model!
  - And it was....

Note: This was also an example of combining (point value) data from multiple sources by fitting a single distribution
Summary….

- **Acceptable data quality**
  - requires judgement

- **Multiple data sets - to combine or not?**
  - requires judgement

- **Distribution - the most appropriate?**
  - requires judgement

- **Risk assessment is art plus science**
  - assessor frequently uses judgement
  - and makes assumptions

- **Only safeguard: transparency!**