Estimating rare-event probabilities without data

Scott Ferson
Institute for Risk and Uncertainty, University of Liverpool
Engineers cannot always get data

• New systems may have no performance history
  – spacecraft of new design or in a new environment
  – biological control strategies using novel genetic constructs that have never existed before

• Two ways to estimate probabilities without data
  – disaggregation into parts whose probabilities are easier to estimate (i.e., breaking it into subproblems)
  – expert elicitation (i.e., guessing)
Is there no uncertainty?

• What is the uncertainty in estimates like
  – “1 in $10^7$”,
  – “about 1 in 1000”, or
  – “never been seen in over 100 years of observation”?

• How should this uncertainty be captured and projected in computations?
Rare events

- Often the driving concern in analyses
- Typically big consequences
- Hardly ever characterized by good data
- Perhaps never seen, or seen only once
Random sample data

• ML says $p$ is zero for never-seen events
  – Nobody believes this is a reasonable estimate

• Bayesian estimator is more reasonable…
Data is “once out of 1000”

Means
0.001    Haldane
0.0015   Jeffreys
0.001986 Zellner
0.001996 Bayes-Laplace
Random sample data

• ML says $p$ is zero for never-seen events
  – Nobody believes this is a reasonable estimate

• The Bayesian estimator is many things
  – ‘Reasonable’ only in that it’s whatever you want

• Modern estimators
  – Imprecise beta (Dirichlet) models
  – Confidence structures
Confidence structure (c-box)

- P-box-shaped estimator of a (fixed) parameter
- Gives confidence interval at *any* confidence level
Estimators

- Point estimates (e.g., sample mean)
- Interval estimates (e.g., confidence intervals)
- Distributional estimates (Bayesian posteriors)
- P-box-shaped estimates (e.g., c-boxes)
Probability of rare event

- Inference about probability from binary data, \(k\) successes out of \(n\) trials

\[ p \sim [\text{beta}(k, n-k+1), \text{beta}(k+1, n-k)] \]

- Identical to Walley’s Imprecise Beta Model with \(s=1\), but needs no prior
Probability $p$ for $k$ of $n$ trials

$p \sim \text{env}(\text{beta}(k, n-k+1), \text{beta}(k+1, n-k))$

Data
$k = 2$
$n = 10$

$(\alpha - \beta)100\%$ confidence interval for $p$

If $1 - \alpha = \beta$, result is identical to classical Clopper–Pearson interval
C-boxes partition the vacuous square
Walley’s IBM

- Assumes beta distributions
  - C-boxes make no shape assumptions
- Needs to select parameter $s$
  - C-boxes have no such parameter
- Works for one problem (binomial $p$)
  - C-boxes are general for many problems
- Not defined by confidence or performance
Plan A
249 out of 1,014 failed

Plan B
39 out of 60 failed

Plan C
17 out of 20 failed

What if we tried all three plans independently?
Conjunction (AND)

Probability that all three plans fail

Has the confidence interpretation
Confidence intervals are clumsy

• Frequentists like confidence intervals but cannot use them in subsequent calculations

• Bayesians can compute with posteriors, but they don’t guarantee statistical performance

• C-boxes take the best from both worlds
Overconfidence

• People, including scientists and engineers, systematically understate their uncertainty
  – 90% confidence intervals ought to enclose the true value 90% of the time on average, but do so only about 30 to 50% of the time
  – Overconfidence “has been found to be almost universal in all measurements of physical quantities”

• Likely to be at least as important in expert elicitation when nothing is being measured
Shlyakhter discounting

- Alex Shlyakhter (1994) suggested this ‘expert’ overconfidence is so pervasive in science we should automatically inflate all uncertainty statements to account for it.

- 95% confidence intervals should be wider by a factor of $3.8/1.96 \approx 2$
Elicitation penalty

- Expert opinions aren’t like random data
- Their uncertainty should be inflated
- The penalty size should be derived empirically from validation studies of prior elicitations
The numbers in “1 in 300” are not counts but estimates with imprecision implied by sigdigs.

\[
P(\text{“1 in 300”}) = \left[ \frac{0.5}{350}, \frac{1.5}{250} \right]
\]

The *envelope* of all corresponding c-boxes.
**Significant-digit intervals**

- Significant digits imply an interval (± half the magnitude of the last significant decimal place)

<table>
<thead>
<tr>
<th>$x$</th>
<th>$s(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>[0, 0.5]</td>
</tr>
<tr>
<td>1</td>
<td>[0.5, 1.5]</td>
</tr>
<tr>
<td>9</td>
<td>[8.5, 9.5]</td>
</tr>
<tr>
<td>10</td>
<td>[5, 15]</td>
</tr>
<tr>
<td>300</td>
<td>[250, 350]</td>
</tr>
<tr>
<td>8150</td>
<td>[8145, 8155]</td>
</tr>
<tr>
<td>$1 \times 10^7$</td>
<td>[$5 \times 10^6$, 1.5$ \times 10^7$]</td>
</tr>
</tbody>
</table>
Examples for elicitations “$k$ in $n$”

Gray c-boxes $B(k, n)$, and black envelopes of c-boxes $B(s(k), s(n))$
More uncertainty for round numbers

- Doubles (or more than doubles) the uncertainty
- Why *more than* doubles?

\[ s(9) = [8.5, 9.5] \quad \text{unit width} \]
\[ s(10) = [5, 15] \quad \text{width of 10} \]

- Presumes greater uncertainty when round numbers are used to characterize a probability
Linguistic uncertainty

• “Words of estimative probability”
• Sherman Kent, Central Intelligence Agency

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Probability (± Range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>certain</td>
<td>100%</td>
</tr>
<tr>
<td>almost certain</td>
<td>(93 ± ~6)%</td>
</tr>
<tr>
<td>probable</td>
<td>(75 ± ~12)%</td>
</tr>
<tr>
<td>chances about even</td>
<td>(50 ± ~10)%</td>
</tr>
<tr>
<td>probably not</td>
<td>(30 ± ~10)%</td>
</tr>
<tr>
<td>almost certainly not</td>
<td>(7 ± ~5)%</td>
</tr>
<tr>
<td>impossible</td>
<td>0%</td>
</tr>
</tbody>
</table>

Holes; Not designed for humans (jargon); Never widely used
Hedges

• Words or phrases that modify numbers, often to express uncertainty
  about...
  ...approximately...
  ...almost...
  ...at least...
  ...at most...
  ...and change
  ...and some
  around...

• There is knowledge trapped in hedges
Empirical characterisation

• Questionnaires
  – Makes people dizzy

• Von Ahn “games with a purpose”
  – Encode the question as a game whose solution answers the question

• Amazon Mechanical Turk
  – Pay people to answer questions or do small tasks
Amazon Mechanical Turk

- ~400 turkers, whose native language is English
- < $50
- Multifarious phrasings and contextualisations

- People interpret uncertainty narrowly
- Allow us to quantitatively characterise hedges
“About 260”
Future work

• Method will be applied to a real fault tree analysis conducted on a malaria control program involving GMOs

• Combining / pooling expert opinions

• Amazon Mechanical Turk study to derive an empirical uncertainty penalty
Acknowledgments

• Keith Hayes, CSIRO

• Michael Balch, Alexandria Validation

• National Institutes of Health
Confidence distributions

- Not widely used in statistics
- Introduced by Cox in the 1950s
- Closely related to well known ideas
  - Student’s $t$-distribution
  - Bootstrap distributions
- Don’t exist for the binomomial rate
Confidence boxes

- Structures that let you infer confidence intervals for a parameter, at any confidence level
- Can be propagated just like p-boxes
- Allow us to compute with confidence
C-boxes

• Not unique
  – Just as confidence intervals are not unique
  – May create some flexibility

• Depend on stopping rule
  – But not knowing the stopping rule may just mean the c-box is wider (and knowing it tightens it)

• Don’t seem overly conservative in practice
Overconfidence

• Humans are too confident
  – Intervals they give are consistently too narrow
  – Stock projections, project timelines, etc., etc.
  – Scientific measurements understate imprecision too

• Empirical evidence documents understatement of uncertainty in measurements of all kinds of physical constants and chemical values
  – Youden, and Morgan and Henrion document many examples
Overconfidence in measuring $c$

http://www.sigma-engineering.co.uk/light/lightindex.shtml