Probability bounding

Scott Ferson
Applied Biomathematics

100 North Country Road, Setauket, New York 11733
scott@ramas.com, 631-751-4350, fax -3435

What is probability?

- Laplace v. Pascal
- Classical (combinatorial) probability
  Is this a winning hand in cards?
- Frequentist probability
  Is this fertilizer any good?
- Subjectivist probability
  Is O.J. guilty?   Does God exist?

Risk analysis

- Some (e.g., Cox 1946, Lindley 1982) argue probability is the *only* consistent calculus of uncertainty
- Subjectivists showed probability provides the consistent calculus for propagating rational subjective beliefs
  - If you're rational and you're forced to bet, probability is the only way to maintain your rationality
  - But being rational doesn't imply an agent has to bet on every proposition
- Just because it would be consistent doesn't mean we *should* use beliefs in risk analysis
- Only the frequentist interpretation seems proper in risk analysis

Incertitude

All we know about *A* is that it's between 2 and 4
All we know about *B* is that it's between 3 and 5
What can we say about *A*+*B*?

Modeling this as the convolution of independent uniforms is the traditional probabilistic approach
but it underestimates tail risks

Probability has an inadequate model of ignorance

Two great traditions

Probability theory

What we think is likely true

Interval analysis

What we know to be false

We need an approach with the best features of each of these traditions

Generalization of both

- Probability bounds analysis gives the same answer as interval analysis does when only range information is available
- It also gives the same answers as Monte Carlo analysis does when information is abundant
- Probability bounds analysis is a generalization of both interval analysis and probability theory

Probability bounds analysis

- Distinguishes variability and incertitude
- Makes use of available information
- All standard mathematical operations
- Computationally faster than Monte Carlo
- Guaranteed to bound answer
- Often optimal solutions

- Very intrusive (requires coding into software)
- Methods for black boxes need development

Consistent with probability

- Kolmogorov, Markov, Chebyshev, Fréchet

- Same data structures used in Dempster-Shafer theory and theory of random sets, except we don't use Dempster's Rule

- Similar spirit as robust Bayes analysis, except updating is not the central concern

- Closely allied to imprecise probabilities (Walley 1992), but not expressed in terms of gambles

- Focus is on *convolutions*

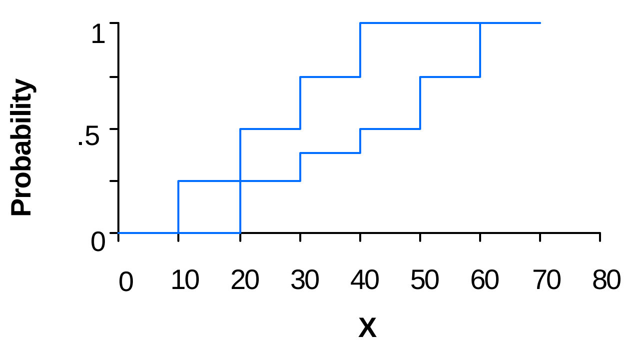
Why bounding?

- Possible even when estimates are impossible
- Results are rigorous, not approximate
- Often easier to compute (no integrals)
- Very simple to combine
- Often optimally tight
- 95% confidence not as good as being sure
- Decisions need not require precision

(after N.C. Rowe)

What is a probability box?

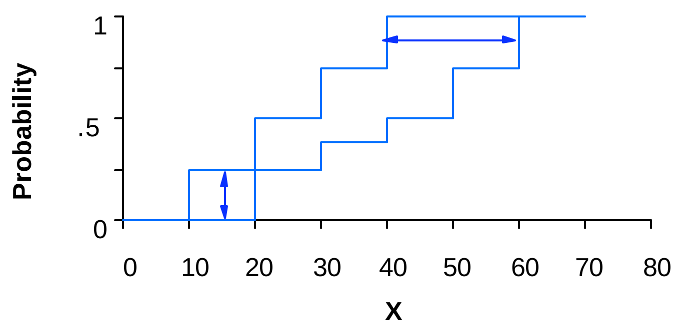
Interval bounds on a cumulative distribution function (CDF)



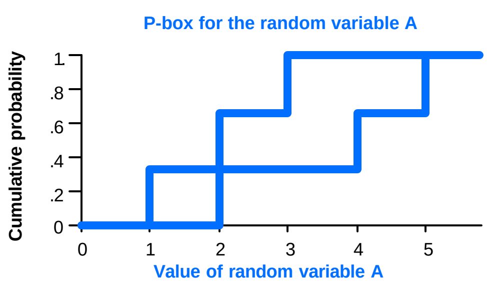
Duality

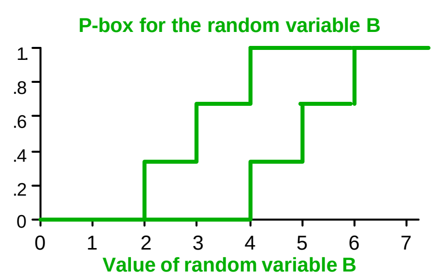
- Bounds on the probability of a value
  likelihood the value will be 15 or less is between 0 and 25%

- Bounds on the value at a probability
  95th percentile is between 40 and 60



Probabilistic arithmetic





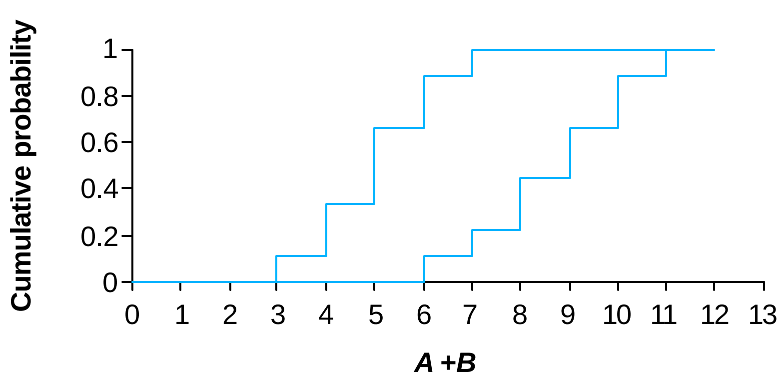
We want to "add" *A* and *B* together,
i.e., compute bounds on the distribution
of the sum *A*+*B*

To convolve *A* and *B*, just take the Cartesian product

A + *B* indep.	$A \in [1, 2]$, $p_1 = 1/3$	$A \in [2, 4]$, $p_2 = 1/3$	$A \in [3, 5]$, $p_3 = 1/3$
$B \in [2, 4]$, $q_1 = 1/3$	$A+B \in [3, 6]$, prob = 1/9	$A+B \in [4, 8]$, prob = 1/9	$A+B \in [5, 9]$, prob = 1/9
$B \in [3, 5]$, $q_2 = 1/3$	$A+B \in [4, 7]$, prob = 1/9	$A+B \in [5, 9]$, prob = 1/9	$A+B \in [6, 10]$, prob = 1/9
$B \in [4, 6]$, $q_3 = 1/3$	$A+B \in [5, 8]$, prob = 1/9	$A+B \in [6, 10]$, prob = 1/9	$A+B \in [7, 11]$, prob = 1/9

Sum under independence

P-box for *A* +*B* assuming independence

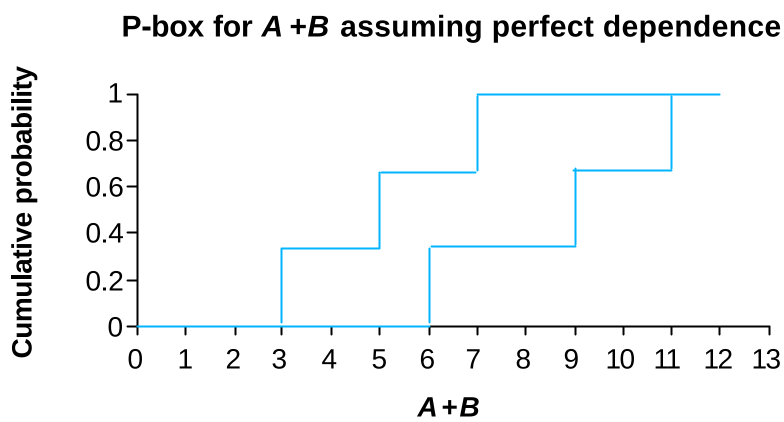


What of other dependencies?

- Independent
- Perfectly positive (maximal correlation)
- Opposite (minimal correlation)
- Positively associated
- Negatively associated
- Particular correlation coefficient
- Nonlinear dependence (copula)
- Unknown dependence

A + *B* *perfect*	$A \in [1, 2]$ $p_1 = 1/3$	$A \in [2, 4]$ $p_2 = 1/3$	$A \in [3, 5]$ $p_3 = 1/3$
$B \in [2, 4]$ $q_1 = 1/3$	$A+B \in [3, 6]$ prob = 1/3	$A+B \in [4, 8]$ prob = 0	$A+B \in [5, 9]$ prob = 0
$B \in [3, 5]$ $q_2 = 1/3$	$A+B \in [4, 7]$ prob = 0	$A+B \in [5, 9]$ prob = 1/3	$A+B \in [6, 10]$ prob = 0
$B \in [4, 6]$ $q_3 = 1/3$	$A+B \in [5, 8]$ prob = 0	$A+B \in [6, 10]$ prob = 0	$A+B \in [7, 11]$ prob = 1/3

Sum under perfect dependence



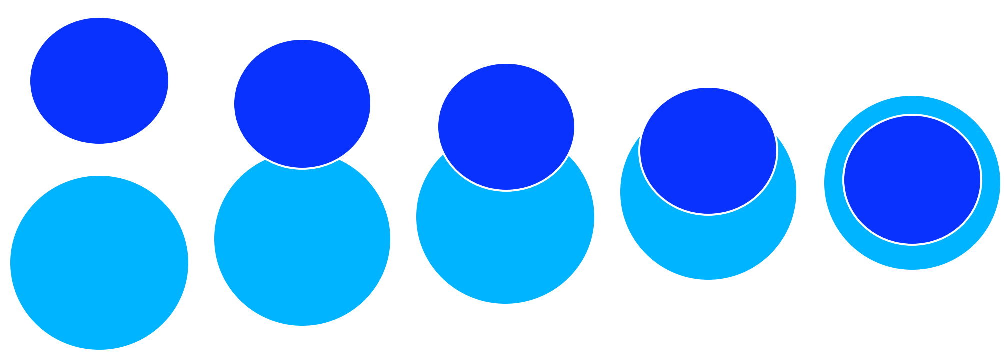
Fréchet inequalities

- Conjunction

$$\max(0, \Pr(F) + \Pr(G) - 1) \le \Pr(F \;\&\; G) \le \min(\Pr(F), \Pr(G))$$

- Disjunction

$$\max(\Pr(F), \Pr(G)) \le \Pr(F \vee G) \le \min(1, \Pr(F) + \Pr(G))$$

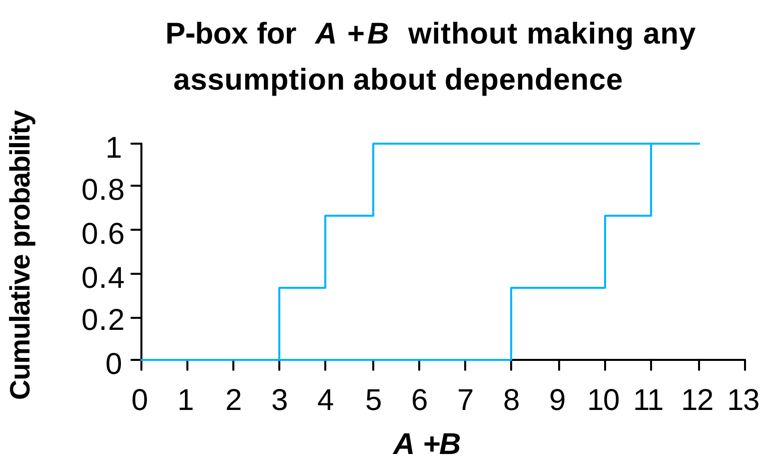


No dependence assumption

- Interval estimates of probability don't reflect that sum must equal one
- Resulting p-box would be too fat
- Linear programming needed for optimal answer using this approach

- Frank, Nelsen and Schweizer (1987) give a way to compute the optimal answer directly

Best-possible answer



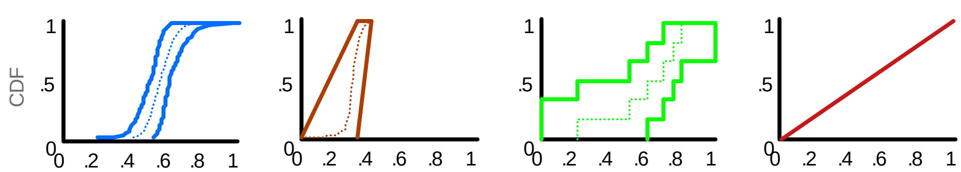
Numerical example

We want bounds on *A*+*B*+*C*+*D* but have only partial information about the variables:

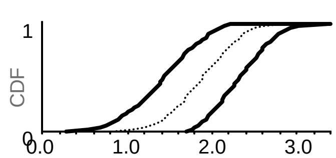
Know the distribution of *A*, but not its parameters.
Know the parameters of *B*, but not its shape.
Have a small data set of samples values of *C*.
D is well described by a precise distribution.

What can we say if we assume independence?
What can we say if we don't make this assumption?

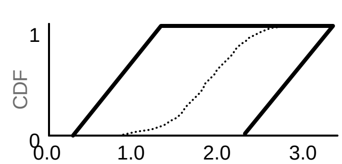
Sum of four p-boxes



A = {lognormal, mean=[.5,.6], variance=[.001,.01]}
B = {min=0, max=.4, mode=.3}
C = {data = (.2, .5, .6, .7, .75, .8)
D = {shape = uniform, min=0, max=1}



Under independence



No assumptions

Summary statistics of risk

Summary	*Independence*	*General*
95th %-ile	[2.1, 2.9]	[1.3, 3.3]
median	[1.4, 2.4]	[0.79, 2.8]
mean	[1.4, 2.3]	[1.4, 2.3]
variance	[0.086, 0.31]	[0, 0.95]

How to use output p-boxes

When uncertainty makes no difference
(because results are so clear), bounding gives confidence in the reliability of the decision

When uncertainty swamps the decision,
(i) use results to identify inputs to study better,
(ii) use other criteria within probability bounds

Seven challenges in risk analysis

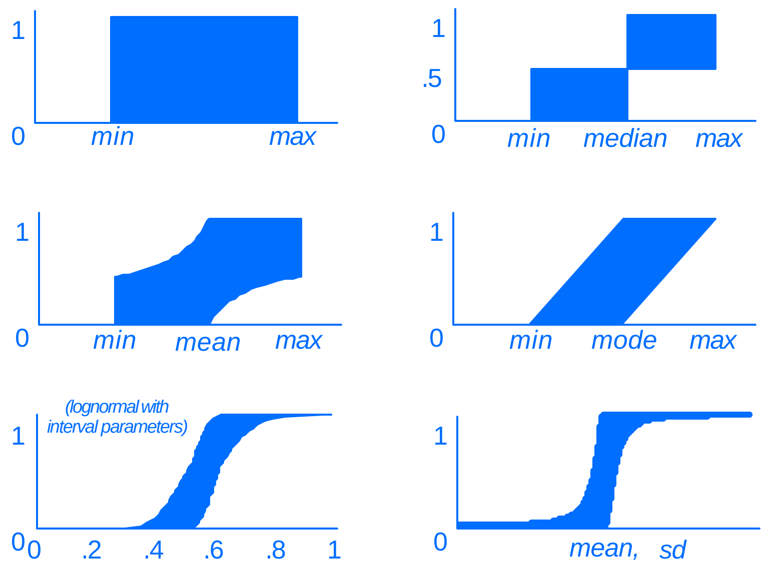
1. Input distributions unknown
2. Large measurement error
3. Censoring
4. Small sample sizes
5. Correlation and dependency ignored
6. Model uncertainty
7. Backcalculation very difficult

For each challenge, we give a poor but commonly used strategy, the current state-of-the-art strategy, and the probability bounding strategy.

1. Input distributions unknown

- Default distributions
- Maximum entropy criterion
- Probability boxes

Probability boxes



Constraints yield p-boxes

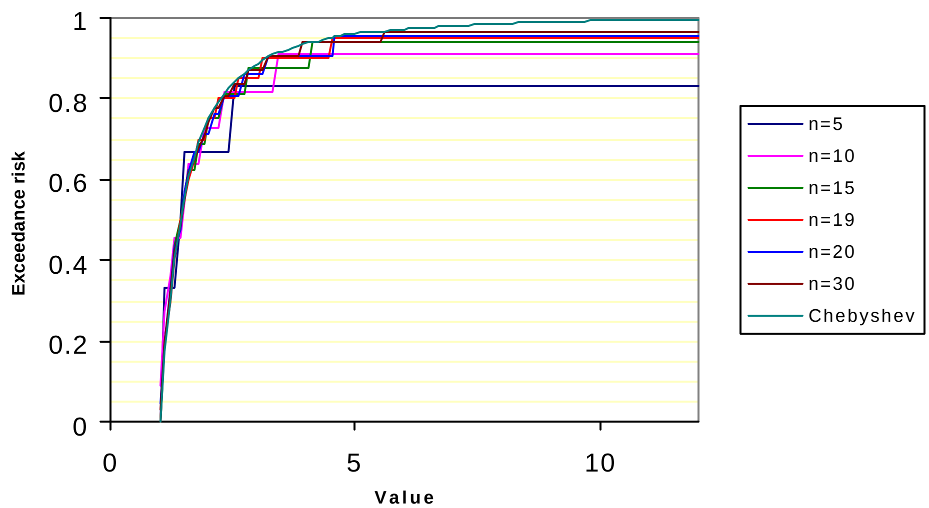
Best-possible bounds are known for these sets of constraints:

{minimum, maximum}
{minimum, maximum, mean}
{minimum, maximum, median}
{minimum, maximum, mode}
{minimum, maximum, quantile}
{minimum, maximum, percentile}
{minimum, maximum, mean = median}
{minimum, maximum, mean = mode}
{minimum, maximum, median = mode}
{minimum, maximum, mean, standard deviation}
{minimum, maximum, mean, variance}

{mean, minimum}
{mean, maximum}
{mean, variance}
{mean, standard deviation}
{mean, coefficient of variation}
{min, mean, standard deviation}
{max, mean, standard deviation}
{shape=symmetric, mean, variance}
{shape=symmetric, mean, standard deviation}
{shape=symmetric, mean, coefficient of variation}
{shape=positive, mean, standard deviation}
{shape=unimodal, min, max, mean, variance}

New (possibly non-optimal) p-boxes can be constructed by intersection

When parameters are estimates



Shown are the best-possible lower bounds on the CDF when the mean and standard deviation are estimated from sample data with varying sample size *n*. When *n* approaches infinity, the bound tends to the classical Chebyshev inequality. (Saw et al. 1986 *Amer. Statistician* 38:130)

Named distributions

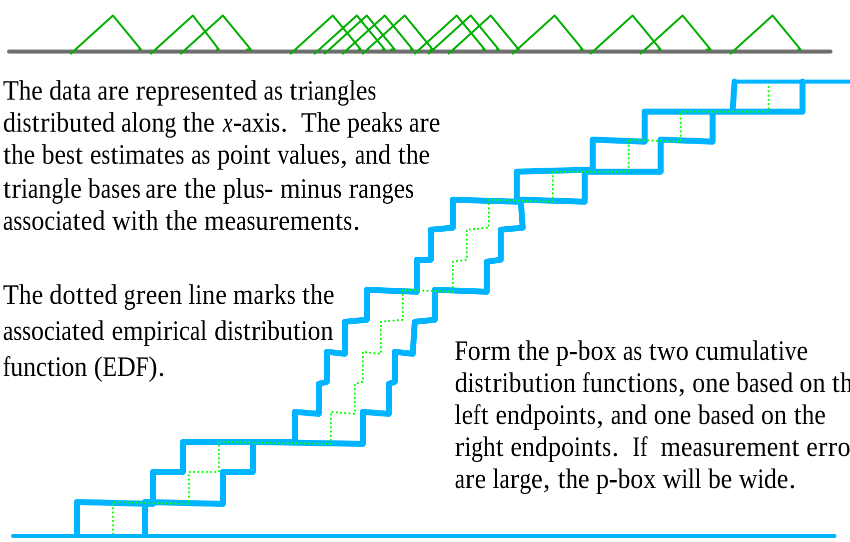
Bernoulli	F	Pascal
beta	gamma	Poisson
binomial	Gaussian	power function
Cauchy	geometric	Rayleigh
chi squared	Gumbel	reciprocal
custom	histogram	rectangular
delta	Laplace	Student's t
discrete uniform	logistic	trapezoidal
Dirac	lognormal	triangular
double exponential	logtriangular	uniform
empirical	loguniform	Wakeby
exponential	normal	Weibull
extreme value	Pareto	X^2

Any parameter for these distributions can be an interval

2. Large measurement error

- Measurement error ignored
- Sampled from in a second-order simulation
- Probability boxes

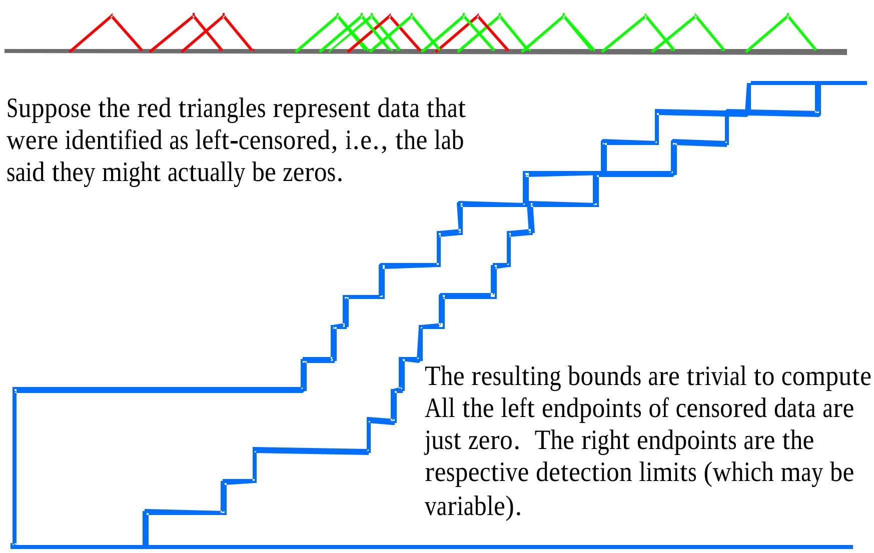
P-box from measurements



3. Censoring

- Substitution methods
- Distributional and "robust" methods
- Probability boxes

P-box under censoring



Censoring

Current approaches

- **Break down when censoring prevalent**
- **Cumbersome with multiple detection limits**
- **Need assumption about distribution shape**
- **Yield approximations only**

P-box approach

- **Works regardless of amount of censoring**
- **Multiple detection limits are no problem**
- **Need not make distribution assumption**
- **Uses all available information**
- **Yields rigorous answers**

4. Small sample sizes

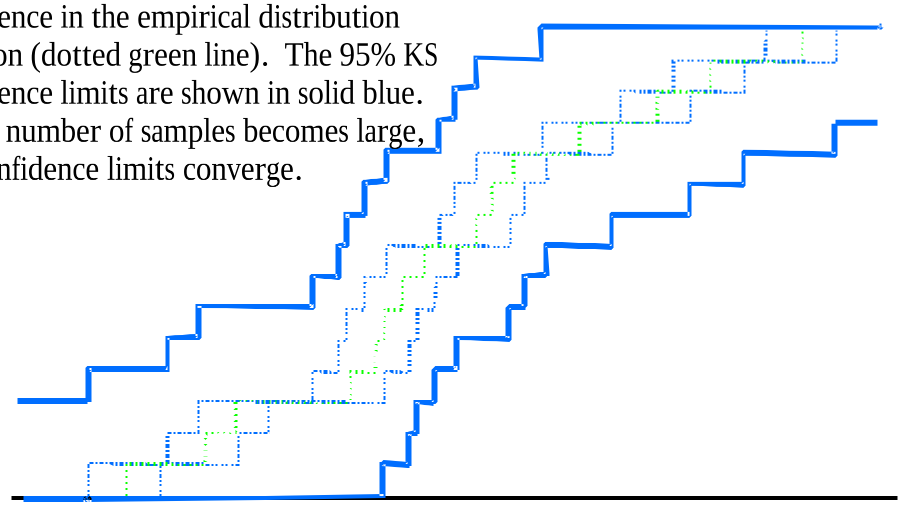
- "Law of small numbers" (Tversky and Kahneman 1971)
- Use confidence intervals in 2-D simulation
- Use confidence intervals to form p-boxes

Extrapolating a subpopulation

- Saw et al. (1986) and similar constraint p-boxes
- Asymptotic theory of extreme values
- Komogorov-Smirnov confidence intervals
  - These are bounds on the distribution as a whole
  - Distribution-free (but does assume iid)
  - EDF(x) ± $D_{max}(\alpha, n)$
  - Compatible with p-boxes including measurement error

P-box with sampling error

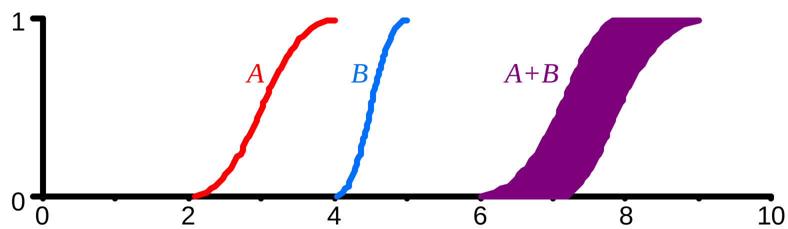
With only 15 data points, we'd expect low confidence in the empirical distribution function (dotted green line). The 95% KS confidence limits are shown in solid blue. As the number of samples becomes large, the confidence limits converge.



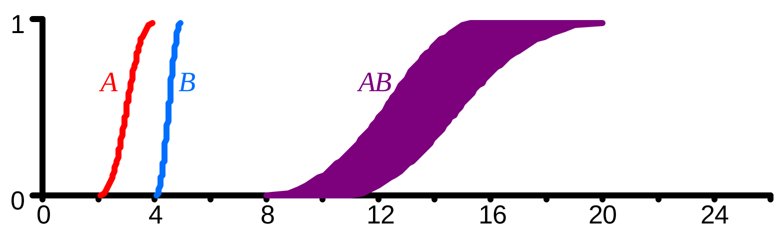
5. Correlations & dependencies

- Assume all variables are mutually independent
- Wiggle correlations between -1 and +1
- Dependency bounds analysis

Dependency bounds analysis

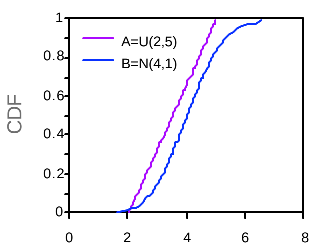


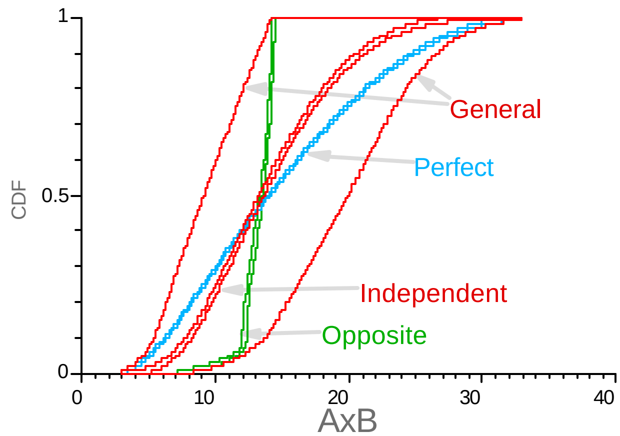
This was a problem of Kolmogorov, only recently solved.
The bounds are rigorous and pointwise best possible.



Wiggling correlations insufficient

If we vary the correlation coefficient between -1 and +1 with currently used correlation simulation techniques, the risk curve would range between the "perfect" and the "opposite" curves below. Dependency bounds analysis shows the actual distribution must be somewhere inside the "general" bounds, and these bounds are known to be best possible.





6. Model uncertainty

- "My model is correct"
- QA, stochastic mixtures and Bayesian averaging
- Stochastic envelopes

Battery of checks

- Generic checks
  - Dimensional and unit concordance
  - Feasibility of correlation structure
  - Consistence of independence assumptions
  - Single instantiations of repeated variables
- Checks against domain knowledge

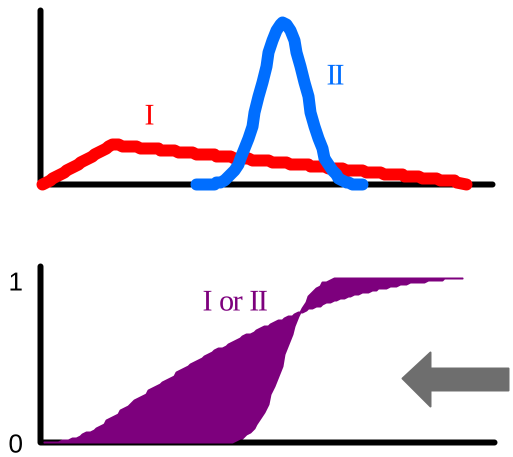
  For instance, in ecological risk analysis...
  - Population sizes nonnegative
  - Trophic relations influence bioaccumulation
  - Food web structure constrained

Doubt about mathematical form

- Stochastic mixture is the traditional way to represent doubt about model form
- Can incorporate judgements about likelihoods of different models
- Easy to use in Monte Carlo simulation

- Averages together incompatible theories
- Can underestimate true tail risks

Stochastic envelope



Models I and II make different predictions about some PDF

P-box capturing the model uncertainty

It can even handle non-stationarity!

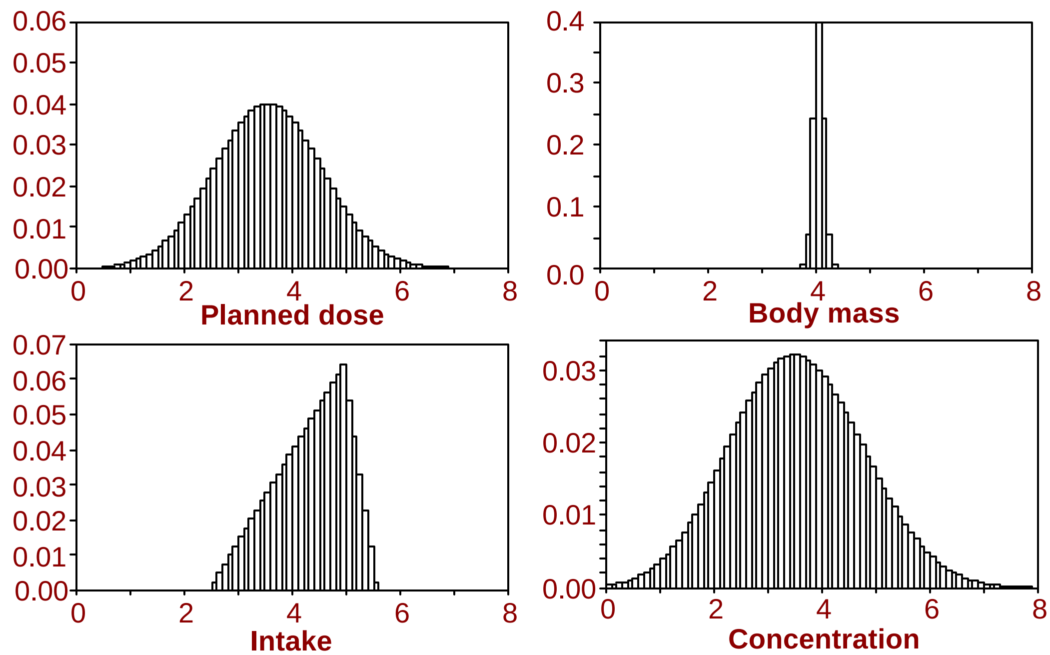
7. Backcalculation

- Revert to deterministic use of point estimates
- Trial-and-error simulation strategies
- Deconvolution of p-boxes

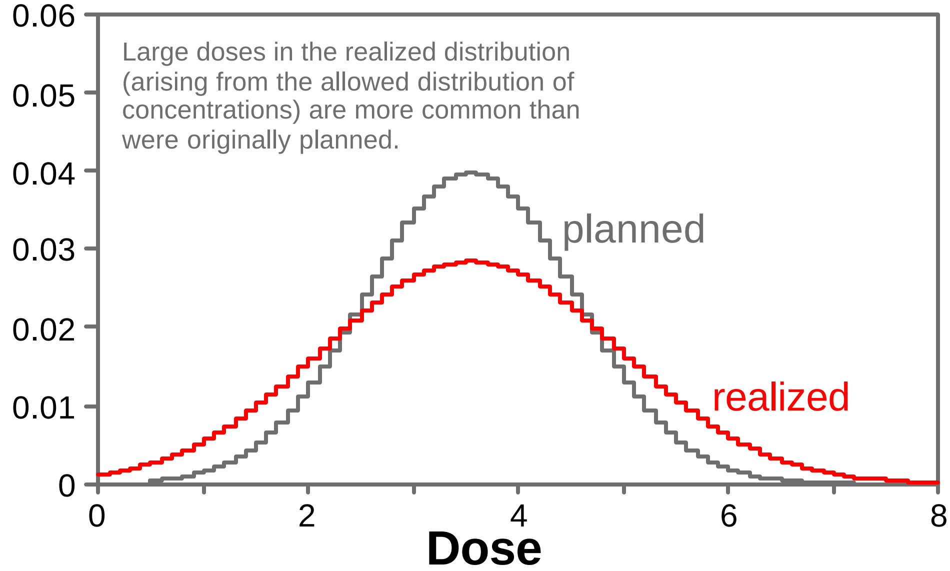
Inverting the defining equation

$$\text{dose} = \frac{\text{conc} \times \text{intake}}{\text{body mass}}$$

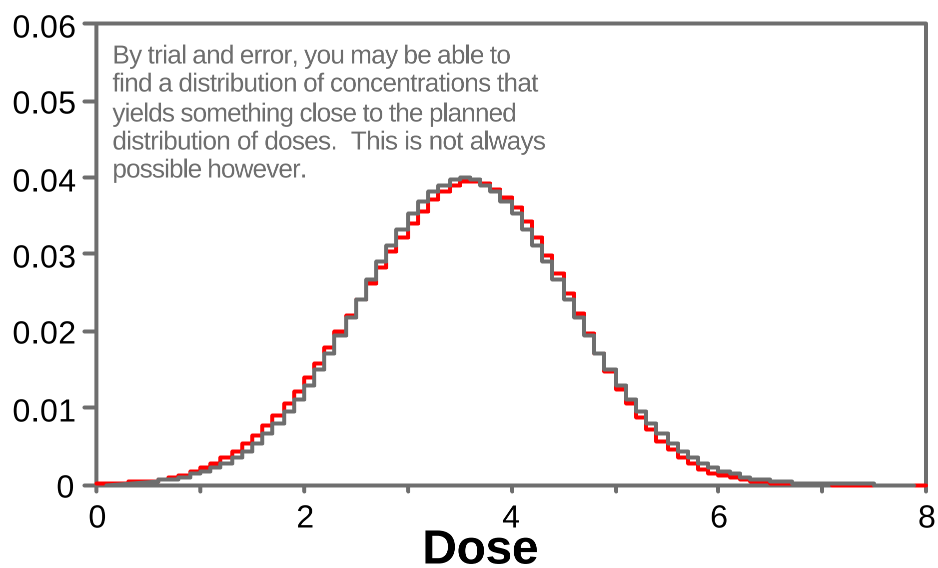
$$\text{conc} = \frac{\text{dose} \times \text{body mass}}{\text{intake}}$$

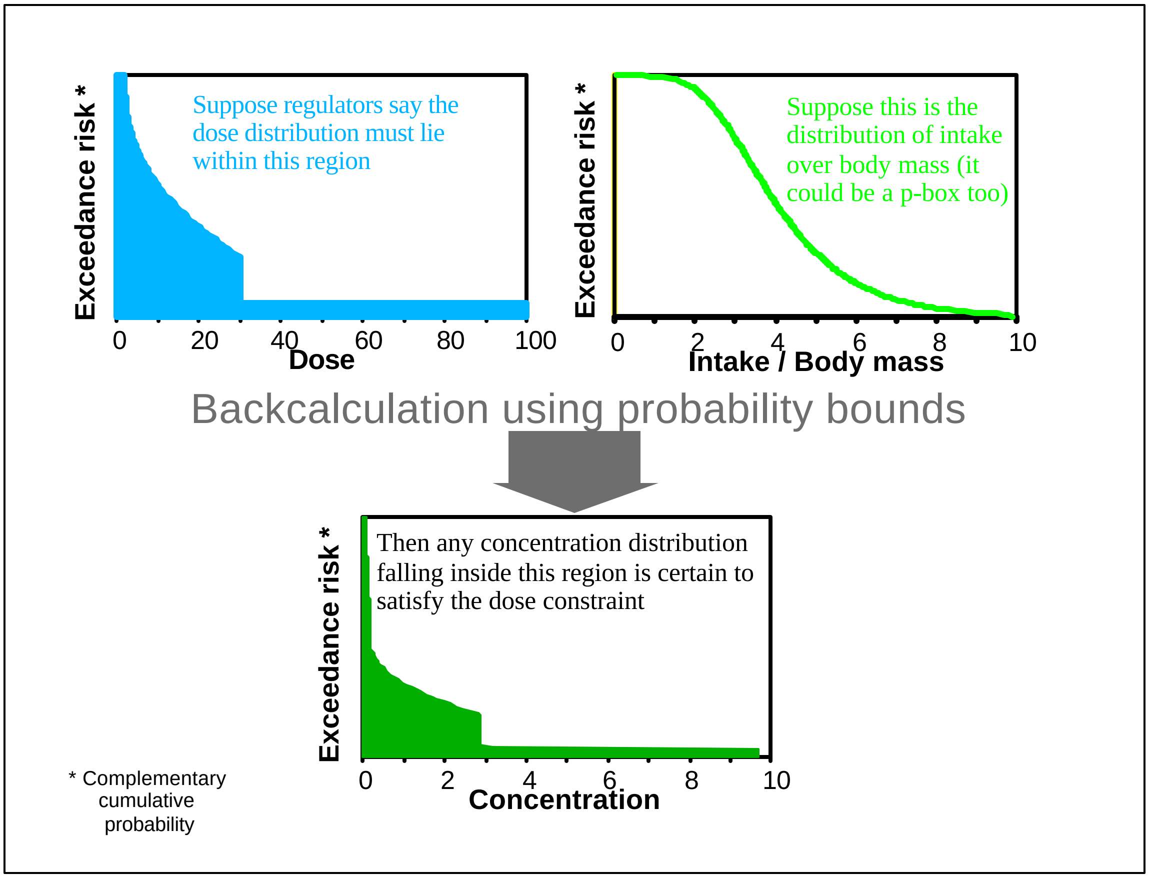


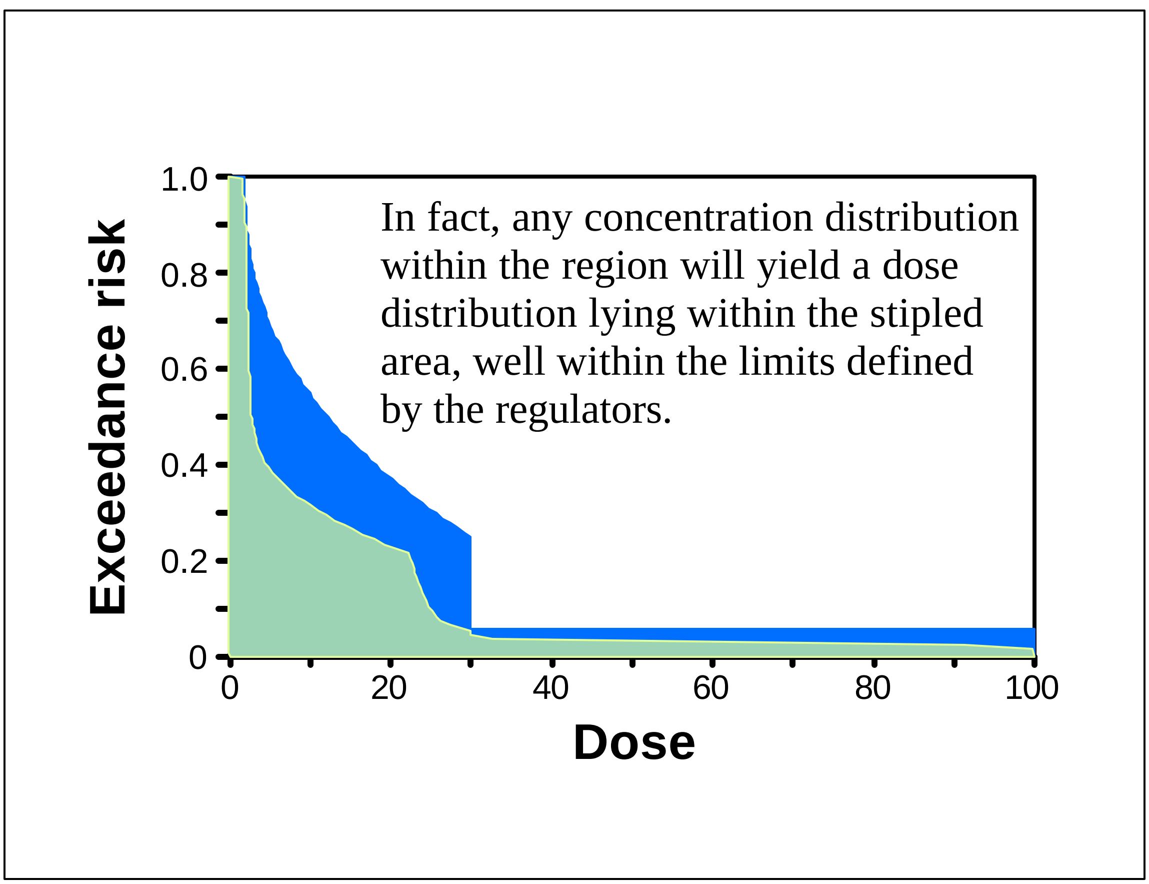
Naive Monte Carlo



Trial and error Monte Carlo




Backcalculation using probability bounds
Exceedance risk *
Suppose regulators say the dose distribution must lie within this region
0 20 40 60 80 100
Dose
Exceedance risk *
Suppose this is the distribution of intake over body mass (it could be a p-box too)
0 2 4 6 8 10
Intake / Body mass
Exceedance risk *
Then any concentration distribution falling inside this region is certain to satisfy the dose constraint
0 2 4 6 8 10
Concentration
* Complementary cumulative probability


Exceedance risk
1.0
0.8
0.6
0.4
0.2
0
In fact, any concentration distribution within the region will yield a dose distribution lying within the stipled area, well within the limits defined by the regulators.
0 20 40 60 80 100
Dose

Backcalculation

- Aside from a few special cases, Monte Carlo methods (including LHS) cannot generally be used to get the target distribution
- Trial-and-error can work but may be impractical
- To get the right answer directly, you need deconvolution
- But known algorithms have terrible numerical problems
- When given arbitrary inputs such as might be defined by regulatory constraints, they usually crash
- P-boxes are a far more natural way to express regulatory constraints
- Because their interval nature relaxes the numerical problems, solutions are also easier to obtain

Advantages of p-boxes

- Marries interval analysis and probability
- Models both interactions and ignorance
- Respects both variability and incertitude
- Handles uncertainty about
  - plus-minus ranges, censoring, sampling error,
  - distribution shapes,
  - correlations and dependencies,
  - model form
  - nonstationarity
- Backcalculation is straightforward
- Simple to use and describe

Disadvantages of intervals

- Same as a (formal) worst case analysis
- Often criticized as hyperconservative
- Cannot take account of distributions
- Cannot take account of correlations and dependencies
- Doesn't express likelihood of extremes

Disadvantages of probability

- Requires a lot of information, or else subjective judgement
- Confounds variability with incertitude
- Cannot handle shape or model uncertainty
- Backcalculation requires trial and error

Disadvantages of 2nd order MC

- Can be daunting to parameterize
- Displays can be ugly and hard to explain
- Some technical problems
  (e.g., when uniform's max<min)
- Expensive calculation (squared effort)
- Cannot handle shape or model uncertainty
- Does not handle incertitude correctly
- Cumbersome in a backcalculation

Disadvantages of p-boxes

- A p-box can't show what's most likely within the box...no shades of gray or second-order information
- Optimal answers may be hard to get when there are repeated variables or when dependency information is subtle
- Propagation through black boxes needs development
- Contradicts traditional attitudes about the universality of pure probability

Present work on ASCI contract

- Representation of information

  How do we get p-boxes? Where do they come from?

- Aggregation methods

  How do we combine estimates from multiple sources?

- Propagation through black boxes

  Can we apply the method to arbitrary engineering problems?

For further information

- Ferson, S. and L.R. Ginzburg. 1996. Different methods are needed to propagate ignorance and variability. *Reliability Engineering and Systems Safety* 54:133-144.
- Ferson, S. 1996. What Monte Carlo methods cannot do. *Human and Ecological Risk Assessment* 2:990-1007.
- Ferson, S. and T.F. Long. 1997. Deconvolution can reduce uncertainty in risk analyses. *Risk Assessment: Measurement and Logic*, M. Newman and C. Strojan (eds.), Ann Arbor Press.
- Ferson, S. 2001. Probability bounds analysis solves the problem of incomplete specification in probabilistic risk and safety assessements. *Risk-Based Decision Making in Water Resources IX*, Y.Y. Haimes, D.A. Moser and E.Z. Stakhiv (eds.), American Society of Civil Engineers, Reston, Virginia, page 173-188.
- Ferson, S. 2001. Checking for errors in calculations and software: dimensional balance and conformance of units. *Accountability in Research: Policies and Quality Assurance* 8:261-279.
- Ferson, S., L.R. Ginzburg and H.R. Akçakaya. Whereof one cannot speak: when input distributions are unknown. *Risk Analysis* [to appear].