Computing with Confidence

Scott Ferson
Applied Biomathematics
100 North Country Road
Setauket, New York 11733 USA
scott@ramas.com, sandp8@gmail.com

Michael Balch
Applied Biomathematics
100 North Country Road
Setauket, New York 11733 USA
michael.balch@arctan-group.com

Kari Sentz
Los Alamos National Laboratory
P.O. Box 1663, MS F609
Los Alamos, New Mexico 87545 USA
ksentz@lanl.gov

Jack Siegrist
Applied Biomathematics
100 North Country Road
Setauket, New York 11733 USA
jack@ramas.com

Abstract

Traditional confidence intervals are useful in engineering because they offer a guarantee of statistical performance through repeated use. However, it is difficult to employ them consistently in analyses and assessments because it is not clear how to propagate them through mathematical calculations. Confidence structures (c-boxes) generalize confidence distributions and provide an interpretation by which confidence intervals at any confidence level can be specified for a parameter of interest. C-boxes can be used in calculations using the standard methods of probability bounds analysis and yield results that also admit the confidence interpretation. Thus analysts using them can now literally compute with confidence. We illustrate the calculation and use of c-boxes for some elementary inference problems and describe R functions to compute them and some Monte Carlo simulations demonstrating the coverage performance of the c-boxes and calculations based on them.

Keywords. confidence intervals, confidence structures, c-boxes, p-boxes, probability bounds analysis, binomial probability, imprecise beta model, *t*-distribution

1 Introduction

When frequentist confidence intervals are constructed across many separate data analyses based on different experiments, the proportion of such intervals that contain the true value of the parameter will match[1] the confidence level, which can be specified in advance to produce any statistical performance that may be desired. Such a guarantee is very attractive to engineers because it allows them to ensure that their conclusions based on confidence intervals will perform according to a specified standard. Bayesian methods in general lack such guarantees that could ensure statistical performance over the long run, and this fact may explain much of the reticence among engineers about adopting the Bayesian framework (Mayo 1996; cf. Vick 2002). On the other hand, Bayesian methodology allows convenient use of its posterior estimates in subsequent calculations, which is usually quite difficult with confidence intervals because it is not clear how knowledge of confidence intervals for parameters can be translated into a confidence interval for an arbitrary function of those parameters using traditional methods.

This paper introduces the notion of confidence structures, or c-boxes. These structures are defined by a traditional confidence interpretation yet admit computations that produce results that also have the confidence interpretation. The next section briefly reviews confidence distributions, which c-boxes generalize. The following sections informally describe c-boxes, give some numerical examples, and compare one of these examples with Walley's imprecise beta model. The paper includes a discussion of the prospects of using c-boxes to compute with confidence, both literally and figuratively, including how to project c-boxes characterizing parameters to estimate the distributions of observable random variates from distributions that depend on those parameters. We provide software functions to compute c-boxes for several important cases and simulate their coverage properties by Monte Carlo methods. Such simulations are useful to determine whether and how conservative the c-boxes are, and thus how useful they are likely to be in practice.

[1] That is, the average frequency of coverage will be at least the specified confidence level.

2 Confidence and Confidence Distributions

The notion of a confidence interval was introduced by Neyman (1937). A confidence interval for parameter θ with coverage α has the property that, among all confidence intervals computed by the same method, at least a proportion α will contain the true value of θ. A confidence interval can serve as an estimate of the parameter that is more sophisticated than any point estimate could be because it encodes not only the available data but also the sampling uncertainty they imply. Valid confidence intervals are more than merely subjective characterizations of uncertainty; they represent rigorous claims and their use establishes a standard of statistical performance that in principle can be checked empirically with Monte Carlo simulations. Credible intervals (sometimes called Bayesian confidence intervals in a usurpation of language) are often considered to be the Bayesian analogs of confidence intervals (Lee 1997), but credible intervals have no general accompanying guarantee like that of the frequentist notion.

Confidence distributions were introduced by Cox[2] (1958), but received little attention in the literature until a recent spike of interest (Efron 1998; Schweder and Hjort 2002; Singh et al. 2005; Xie et al. 2011; Xie and Singh 2012; inter alia). A confidence distribution is a *distributional estimate* for a parameter, in contrast with a point estimate like a sample mean or an interval estimate such as a confidence interval. It has the form of a distribution function on the space of possible parameter values that depends on a statistical sample in a way that encodes confidence intervals at all possible confidence levels. A confidence distribution for a parameter $\theta \in \Theta$ is a function C: $\Theta \rightarrow (0,1)$ such that, for every α in (0,1), $(-\infty, C^{-1}(\alpha)]$ is an exact lower-sided 100α% confidence interval for θ , where the inverse function $C^{-1}(\alpha) = C_n^{-1}(x_1,\ldots,x_n, \alpha)$ is increasing in α. This definition obviously also implies $[C^{-1}(\alpha), C^{-1}(\beta)]$ is a 100(β–α)% confidence for the parameter θ. Although related to many other ideas in statistical inference (Singh et al. 2005; Xie et al. 2011), a confidence distribution can be considered a purely frequentist concept (Schweder and Hjort 2002; Singh et al. 2005).

An important example of a confidence distribution is for the parametric mean of a normal distribution based on random sample data x_i, $i = 1, 2, \ldots, n$. The confidence distribution in this case is

$$C_n(\mu) = F_{Tn-1}((\mu - \bar{x})\sqrt{n}/s)$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and F_{Tn-1} denotes the cumulative distribution function of Student's t-distribution with n–1 degrees of freedom. Confidence intervals for the normal's mean can be constructed directly from this confidence distribution as the inverse image of any subset of the confidence distribution's range that has measure equal to the intended confidence level. In particular,

$$[C_n^{-1}(\alpha), C_n^{-1}(\beta)] = \bar{x} + s\,[F_{Tn-1}^{-1}(\alpha), F_{Tn-1}^{-1}(\beta)] / \sqrt{n}$$

is a 100(β–α)% confidence interval on the mean. For the sake of clarity and convenience for readers, these formulas can be rendered as code for the R statistical computing language (R Development Core Team 2011):

```
pcnorm.mu = function(mu, x)
  pt(sqrt(length(x))*(mu-mean(x))/sd(x),length(x)-1)

cinorm.mu = function(x, c=0.95, alpha=(1-c)/2, beta=1-(1-c)/2)
  mean(x)+qt(c(alpha,beta),df=length(x)-1)*sd(x)/sqrt(length(x))
```

The function pxnorm.mu accepts random normal sample values in the x array and returns the value of the confidence distribution for every value in the mu array. The cinorm.mu function also takes the random samples in the x array, and returns a confidence interval for the mean of the normal distribution that generated those sample values at a confidence level set by the argument c, which defaults to 95%, or by alpha and beta if they are specified.

A Monte Carlo simulation can be implemented using the following R function to check that the confidence distribution indeed allows valid confidence intervals at any level to be constructed from it:

```
covnorm.mu = function(n,mu,sigma,many=1e4,lots=1e3, ... ) {
 ab = alphabeta(...)
 m = seq((mu-5*sigma),(mu+5*sigma),length.out=many)
 cov = 0
 for (i in 1:lots) {
   x = rnorm(n, mu, sigma)
   h = pcnorm.mu(m, x)
   ci = range(m[(ab[1]<=h) & (h<=ab[2])])
   if ((ci[1]<=mu)&(mu<=ci[2])) cov=cov+1 }
 cat(' Intended',diff(ab)*100,'%\n','Observed',100*cov/lots,'%\n')
 cov/lots }
alphabeta = function(c=0.95,a=(1-c)/2,b=1-(1-c)/2) sort(c(a, b))
```

This function can be exercised with a call like covnorm.mu(n, μ, σ), specifying just a positive integer n and the true mean and standard deviation to use in the simulation, which will return a value around 0.95, or a call like covnorm.mu(n, μ, σ, a=α, b=β) may also specify particular α and β levels.

Although a confidence distribution has the form of a probability distribution, it is usually not considered to be a probability distribution. It corresponds to no randomly varying quantity; the parameter it describes is presumed to be fixed and nonrandom. Some also emphasize that the value of the function C is not probability of θ, but

[2] Fraser (2011) argues that confidence distributions can be found in the work of Fisher (1930; 1935) under the name 'fiducial', and even in that of Bayes (1763) namelessly.

rather confidence[3] about θ (Cox 2006; cf. Lindley 1958). A confidence distribution is merely a ciphering device that encodes confidence intervals for each possible confidence level. Nevertheless, it might be reasonable and convenient to adopt a notation that only implicitly denotes the confidence distribution, so that, for instance, in the case of the normal mean, we can write

$$\mu \sim \bar{x} + s\, T_{n-1}/\sqrt{n}$$

where T_{n-1} denotes a random variable from Student's *t*-distribution (Student 1908) with $n-1$ degrees of freedom. This notation avoids the need to name the confidence distribution function. Note that this use of the tilde ~ extends conventional uses in statistics. We suggest that it can still be read as "has the distribution", or perhaps "has uncertainty like", but it obviously does not suggest that the left-hand side is a random variable. The left-hand side after all is a value that is fixed, though unknown. Instead, it says that the inferential uncertainty about the fixed parameter μ is characterized by the transformed *t*-distribution.

Despite their intimate connection with *t*-distributions, confidence distributions are not widely known in statistics, at least not under that name. Efron (1998) characterized bootstrap distributions as (approximate) confidence distributions, and so confidence distributions are widely used in modern statistics, albeit under the guise of bootstrap distributions.

The notion of confidence distributions is not without critics. Early association with fiducial inference has led to some confusion. Some readers seem to have difficulty accepting confidence distributions on their own terms. The arguments of Robert (2012) are paraphrased a bit more bluntly in his blog (http://xianblog.wordpress.com/2012/06/11/confidence-distributions/): "Either the confidence distribution corresponds to a genuine posterior distribution, in which case I think the only possible interpretation is a Bayesian one. Or the confidence distribution does not correspond to a genuine posterior distribution, because no prior can lead to this distribution, in which case there is a probabilistic impossibility in using this distribution." Of course confidence distributions are not trying to be Bayesian posterior distributions, so it should hardly be disquieting if they fail to be. The requisite interpretation of confidence distributions is of course Neyman confidence, which Bayesian posteriors do not generally have.

One potential practical disadvantage of confidence distributions is that they are not unique. Multiple functions may fill the bill, and there seems to be no general way to pick the best confidence distribution from among them. Of course, confidence intervals themselves are not unique either. There are usually lots of reasonable ways to construct a confidence interval for any parameter, even for fixed data and model. Neither form of non-uniqueness seems to impede the purpose of guaranteeing long-term statistical performance.

Another significant limitation on the use of confidence distributions is that not every important inferential problem has a solution. Confidence distributions are often constructed by inverting the upper limits of lower one-sided confidence intervals of all levels, but this is not possible for all important inferential problems. Notably, in particular, *there is no confidence distribution for the binomial probability*.

3 Confidence Structures (C-boxes)

Confidence distributions are special cases of more general confidence structures (Balch 2012), which we call 'confidence boxes' or 'c-boxes' because they may often be characterized by two bounding distributions like probability boxes (Ferson et al. 2003). A c-box represents inferential uncertainty about a parameter that characterizes some distribution from which limited or poor or discrete data have been randomly sampled. Like a confidence distribution, a c-box is defined by the property that it can be used to construct Neyman confidence intervals at any confidence level for that parameter. C-boxes generalize confidence distributions because both are estimators of unobservable parameters, but c-boxes can be applied to problems with discrete observations, interval-censored data, and even inference problems in which no assumption about the distribution shape can be made.

Methods for deriving c-boxes are varied (Balch 2012). Generally, wherever a meaningful and valid confidence interval can be defined, a c-box can also be defined. If a confidence interval is based on a pivot, that pivot can be used to directly define a c-box. Any defined confidence distribution can be generalized to a c-box when its data are encoded not as point values but as intervals to account for mensurational uncertainty from the inability to measure individual quantities with perfect precision (Nguyen et al. 2012; Ferson et al. 2007). When a confidence interval is based on a significance function, i.e., a function (of parameters and data) that produces *p*-values in a significance test, the significance function can be used to construct a consonant confidence structure, encoded as a Dempster–Shafer structure which can then be transformed, with some loss of information (Ferson et al. 2003), into a p-box (Balch 2012).

The formula and R function for this c-box of the normal mean can be generalized for the case of interval-censored data using a straightforward but non-trivial algorithm that

[3] Of course, confidence is a probability in a different domain; confidence is the probability realized by frequency that those defined intervals $(-\infty, C^{-1}(\alpha)]$ actually enclose the parameter over some in some future, perhaps hypothetical series of experiments.

extremizes $C_n(\mu)$ over possible configurations of point x-values within their respective interval ranges (Nguyen et al. 2012; Ferson et al. 2007). In case the intervals all overlap any value of μ, the result is vacuous (i.e., the interval [0,1]) for that value. For example, if interval-censored random samples from a normal distribution are {[8,11], [5.5,6.9], [−1.3,0.3], [3.5,7.5], [0.8,1], [2.8,4.2], [1.8,5.2], [2.2,5.2], [3.5,5.7], [5.3,6.1]}, a c-box for the normal mean is shown in Figure 1.

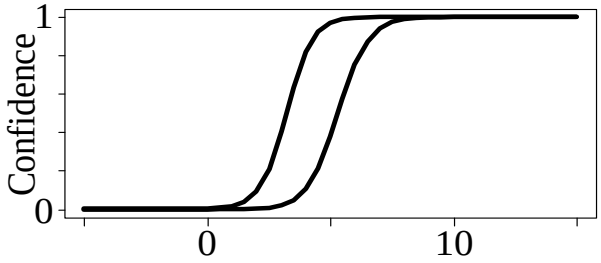

Figure 1: C-box for the normal mean from interval data.

To extract a confidence interval from a c-box, select values of α and β that imply a desired confidence level $100(\beta-\alpha)\%$, and map these values from the confidence axis to the x-axis. The larger value β is mapped through the *right* bound, and the smaller value α is mapped through the *left* bound.

4 Computing with Confidence

Many authors (e.g., Grosof 1986) have suggested using ordinary confidence procedures to obtain interval inputs for use with interval analysis (Moore 1966) for bounding numerical results that depend on sample data. For example, EPA (2002) guidance instructs risk analysts to use the upper bound from the 95% confidence interval for a pollutant's mean concentration rather than the actual sample mean of observed concentration values in order to be protective of the public health in the face of sampling uncertainty arising from sometimes very small sample sizes. Although this may be a reasonable strategy when there is only a single variable for which sampling uncertainty is a major concern, it is not statistically defensible when such uncertainties for several variables must be combined together. Statistical confidence intervals are not rigorous intervals guaranteed to enclose the value they estimate, and therefore confidence intervals do not formally admit interval calculation in the sense of Moore (1966).

Some limited statements are possible using ad hoc application of Bonferroni or Šidák corrections or Boole or Fréchet inequalities (e.g., Ferson 1996). For example, if we combine, say by addition, two 95% confidence intervals using simple interval arithmetic, we might expect the result to be a ~90% confidence interval for the sum because the conjunction of the two probability statements would imply multiplying the two probability levels, at least assuming independence between them. If seven such confidence intervals were combined in some mathematical function, the implied probability level under independence would be less than 70%. Without the independence assumption, the level could fall as low as 65%. To achieve 95% confidence for the result, one would presumably have to use input confidence intervals with confidence level equal to the seventh root of 95%, which is greater than 99%. Because confidence intervals often get substantially wider as the confidence level rises, this approach is rarely workable in practice.

The alternative approach of computing with confidence distributions is also not practical just because (precise) confidence distributions often do not exist for important problems. This limitation may be alleviated by c-boxes because they generalize confidence distributions and more easily provide solutions. Although Cox (2006) counseled that analysts should not try to use confidence distributions in calculations as though they were true probability distributions, Balch (2012) proved that two or more independent c-boxes can be propagated through a function to yield a valid c-box. This is much more efficient than propagating individual confidence intervals because the combinations do not require application of the Bonferroni or Šidák corrections and they deliver results at all confidence levels all at once.

For example, suppose one were interested in computing a 95% confidence interval on the mean difference between two normal populations with both unknown mean μ and unknown standard deviation σ. Suppose we collect four random samples from each population, say, {2.71, 5.46, 5.45, 5.50}, and {1.88, 1.54, 1.15, 0.46}. One approach to obtaining the desired interval would be to take the interval-difference of the 97.468% confidence intervals on the two population means. The resulting estimate would be $\mu_2 - \mu_1 = [0.37, 6.67]$ with 95% confidence. Alternatively, one could take the stochastic difference of the two c-boxes on the uncertain means which are (shifted and scaled) t-distributions. This yields a much tighter 95% central confidence interval on the difference, [1.10, 5.94], although it is somewhat more difficult to compute because it involves a subtractive convolution rather than merely an interval difference. Still, it can be calculated via Monte Carlo simulation in R using only three lines:

```
rcnorm.mu = function(m, z)
   mean(z)+sd(z)*rt(m, length(z)-1)/sqrt(length(z))
d = sort(rcnorm.mu(m, x) - rcnorm.mu(m, y))
range(d[round(c(0.025*m, (1-0.025)*m))])
```

where `x` and `y` are the vectors of sample values, `m` is the number of Monte Carlo simulations. In fact, this result is the same as the 95% credible interval that would be obtained using Bayesian inference with a Jeffreys prior. The convolution of the confidence distributions yields confidence intervals by a purely frequentist analysis that supports a traditional confidence interpretation in this

and other cases generally. The following R function can be used to implement straightforward Monte Carlo simulations that demonstrate the confidence intervals produced by this approach have the prescribed coverage:

```
covnorm.mudiff=function(n,mu,sigma,many=1e4,lots=1e3,...){
 ab = alphabeta(...)
 truediff = mu[1] - mu[2]
 cov = 0
 for (i in 1:lots) {
  x = rnorm(n[1], mu[1], sigma[1])
  y = rnorm(n[2], mu[2], sigma[2])
  ci=range(sort(rcnorm.mu(many,x)-
              rcnorm.mu(many,y))[round(many*ab)])
  if ((ci[1] <= truediff) & (truediff <= ci[2])) cov = cov + 1 }
 cat(' Intended',diff(ab)*100,'%\n','Observed',100*cov/lots,'%\n')
 cov/lots }
```

This function can be called like covnorm.mudiff(*n*, μ, σ), where *n*, μ and σ are now each *pairs* describing the sample sizes and parameters for the two populations. For instance, covnorm.mudiff(c(10,20),c(5,1),c(2,3)) will return a value around 0.95.

5 C-box for the Binomial Probability

A Bernoulli random variable has only two possible values, perhaps designated {failure, success}, or more conveniently {0, 1}. A binomial random variable is a random variable whose value is a count of Bernoulli successes observed over $n > 0$ independent identical trials, each of which has the same probability p of success, which produces k successes from those n trials (where $0 \leq k \leq n$). A fundamental problem in risk analysis and statistics generally is to characterize what can be inferred about p from observing k successes out of n trials, under the assumption that the trials are independent and the binomial probability p is fixed across the trials.

In fact, the original problem in the famous paper of Bayes (1763) was about the estimation of the binomial probability. The paper begins "*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named" (Bayes 1763, page 376). The same page also says "By *chance* I mean the same as probability." We take this to be asking, given k successes and $n-k$ failures out of n trials where $k \sim \text{binomial}(n, p)$, what is $\Pr(p \in [p_1, p_2])$, for any values p_1 and p_2?

Balch (2012) offers a c-box solution to this problem:

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

where p is the binomial parameter (which is a fixed but unknown value), and the two beta distributions are the left and right edges of the c-box that characterizes the inferential uncertainty about p. Note that we continue to use the ~ symbol even though the right-hand side has the form of a p-box. The ~ can be read as "has uncertainty like". We understand this to entail that the parameter on the left-hand side has inferential uncertainty characterized by a confidence distribution consistent with or inside the c-box, that is, a distribution that is bounded in the cumulative by the two edge distributions of the c-box.

Figure 2 depicts an example using $k = 2$ and $n = 10$ in a graph whose abscissa consists of the possible values of the parameter p and whose ordinate is confidence (probability).

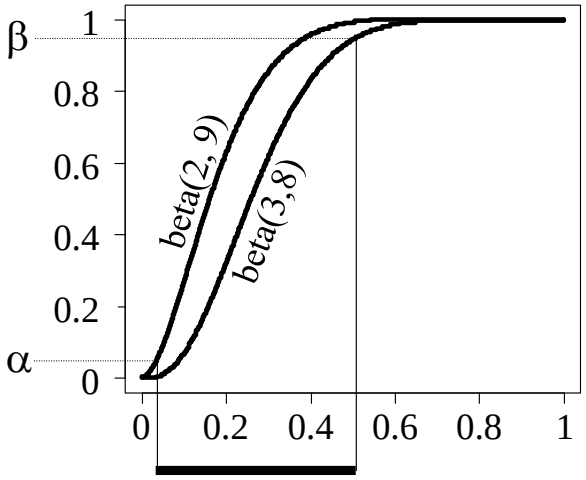

Figure 2: C-box and a 100(β–α)% confidence interval for probability from 2 successes in 10 trials.

The c-box in Figure 2 has a confidence interpretation, which means that one can generate from it true confidence intervals for the binomial probability p at any desired level of confidence. For example, the depicted interval is the symmetric 90% confidence interval [0.037, 0.507]. The confidence intervals obtained in this way are identical to the classical Clopper–Pearson (1934) confidence intervals on the binomial probability. One-sided confidence intervals can be obtained by setting α to zero or β to one. The c-box approach readily provides results for cases involving $k = 0$ and $k = n$, and even the no-data case where $n = 0$, without the overthinking required by a Bayesian analysis constrained to a single precise distribution (Winkler et al. 2002).

Of course the Bayesian and frequentist approaches are trying to do different things. In the c-box approach, p_1 and p_2 are sought to be functions of the data and probabilities are conditional on some hypothetical (but unknown) value of p. In contrast, Bayes explicitly conditions on the data, and asks about the probability of p as a latent variable. These approaches are asking a very different questions: c-boxes ask about coverage for a

fixed value of p, whereas Bayes is asking about the probability of p as a latent random variable.

The c-box and arbitrary confidence intervals for the binomial probability given k successes out of n trials can be computed in R with the functions:

```
pcbinom.p = function(p, k, n)
  list(left=pbeta(p, k, n-k+1), right=pbeta(p, k+1,n-k))

cibinom.p = function(k, n, c=0.95, alpha=(1-c)/2, beta=1-(1-c)/2)
  qbeta(c(alpha,beta), c(k,k+1), c(n-k+1,n-k))
```

Straightforward Monte Carlo simulation can demonstrate the confidence intervals perform statistically.

Note that the c-box also answers Bayes' question about the chance p is in some range, but it gives an interval rather than a single precise probability. The c-box says $\Pr(p \in [p_1, p_2]) \in [\min(0, B_R(p_2) - B_L(p_1)),\ B_L(p_2) - B_R(p_1)]$, where B_L denotes the cumulative beta distribution with parameters k and $n-k+1$, and B_R is the cumulative beta with parameters $k+1$ and $n-k$. The lower bound can be called confidence, and the upper bound plausibility, and together they characterize the chance sought by Bayes.

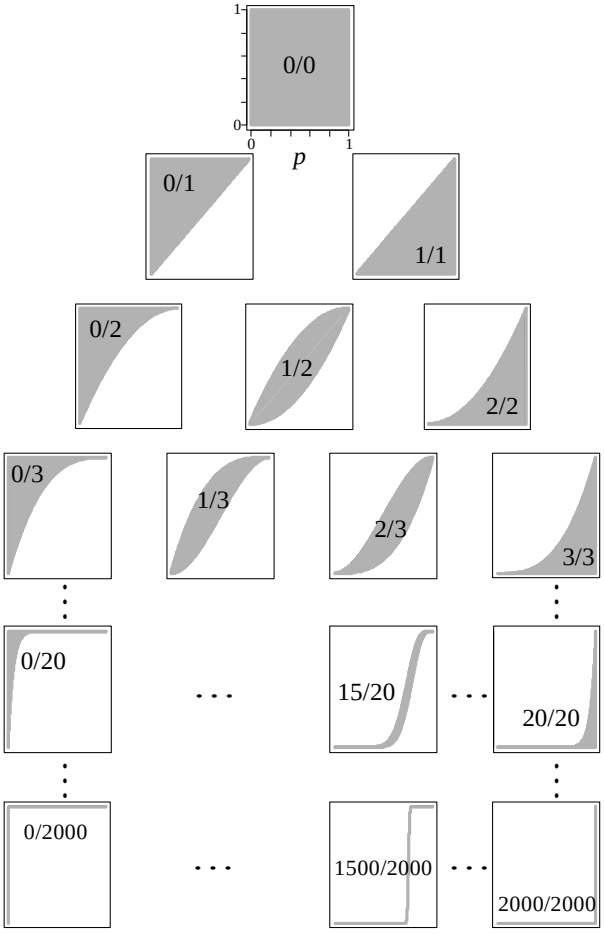

Figure 3: C-boxes for the binomial probability implied by k/n successes out of trials.

Figure 3 shows the first few c-boxes for sample sizes between zero and three. Notice that the c-box for the null case when $n = 0$ corresponds to the entire unit square. Thereafter, the possible c-boxes for any given sample size partition the unit square. As sample size increases, of course the c-box approaches a precisely specified beta distribution which becomes steeper and steeper and centered on the observed frequency k/n.

What determines whether the solution to an inference problem is a precise confidence distribution or a non-degenerate, imprecise c-box? For the normal mean the solution is precise unless the data are themselves imprecise from interval-censoring (as in Figure 1). For binomial probability, however, the solution is imprecise even for well identified data. The reason is what ecologists call "demographic" uncertainty (Akçakaya 1991), which is the variation that arises simply because of the constraint that data must come as integers. The discrete nature of binomial sampling means that evidence cannot reflect patterns as well as continuous data can. Demographic uncertainty is only important for small sample sizes, but it cannot be neglected in such cases.

5.1 Comparison with the Imprecise Beta Model

The c-box solution to the binomial probability estimation problem can be compared to the imprecise beta model (IBM) first suggested by Dempster (1966) but elaborated and championed by Walley (1991; 1996; Walley et al. 1996; Bernard 2005). The IBM employs a class of prior distributions beta(st, $s(1-t)$), $t \in [0,1]$, defined by a single, fixed value $s > 0$ that measures resistance (maybe stubbornness) of the model to new data. After observing k successes in n trials, the posterior is the class beta($st+k$, $s(1-t)+n-k$). Extremizing t from 0 to 1 yields the posterior p-box [beta(k, $s+n-k$), beta($s+k$, $n-k$)] whose expectation is the interval $[k/(s+n), (s+k)/(s+n)]$. As data become available and the model is updated, the left and right beta distributions incrementally converge in accordance with a rate defined by the parameter s. Figure 4 illustrates, for three different values of s, how the vacuous prior (top row) contracts to a posterior with the addition of each binary datum in the sequence {0, 0, 1, 0}. Each graph shows eleven beta distributions evenly distributed across the posterior class.

The IBM is an example of Bayesian sensitivity analysis or robust Bayes analysis (Berger 1985). It may be thought of as many simultaneous Bayesian analyses with many priors ranging between the limiting distributions beta(0,1) and beta(1,0), in which at least one posterior may be improper if k is equal to n or zero. Walley (1991) has demonstrated that robust Bayes analysis is part of a more general theory based on imprecise probabilities of very broad scope and flexibility, for which there is a firm theoretical foundation based on respecting consistency and coherence requirements but which avoids making unwarranted assumptions to obtain

quantitative answers. The most important feature of the IBM is that it does not require the analyst to select some precise probability distribution as prior. The IBM instead intends to specify a reasonable class of priors. The idea is that no single distribution could be reasonable as a model of prior ignorance, but considered as a whole, the class of beta distributions with all possible means specified by IBM is arguably a reasonable model for ignorance.

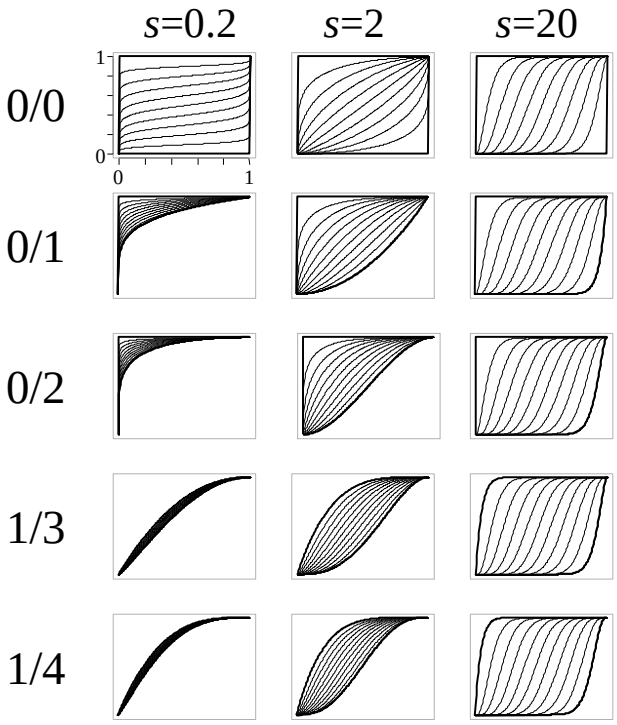

Figure 4: IBMs and their beta distributions for different values of *s* as data accumulates.

In the degenerate initial case, when the sample size is zero before any data are collected, the posterior is the same as the prior, and the IBM yields a vacuous posterior that effectively says the probability could be anywhere in the interval [0,1], which is arguably the only sensible inference when there are no data at all. When the sample size is very large, the posterior is a tight p-box that tends to the observed frequency, as all Bayesian analyses do. In the practical intermediate cases of small sample sizes, the posterior from the IBM is a credal set containing a range of beta distributions whose breadth reflects the uncertainty about the prior that a traditional Bayesian analysis ignores. Importantly, this breadth is not too wide to be useful, but yields answers whose imprecision is roughly what one might expect to see across a community of competent Bayesians (Walley 1991).

A user of the IBM must chose a value for the parameter *s*. This value determines the speed of convergence with which data cause the initially vacuous state of uncertainty to condense into the precise posterior approaching the observed frequency *k*/*n*. High values of *s* cause the IBM to converge slowly. For a given value of *t*, larger values of *s* cause the variance of the distribution beta(st, $s(1-t)$) to be smaller, so when the distribution is considered as an estimate of θ, larger *s* means there is more precision about the parameter. Walley (1996; Walley et al. 1996) recommended using $s = 1$ or $s = 2$, with preference for the larger value.

The c-box approach described in the previous section conforms with an IBM using $s = 1$, although the IBM and c-box have rather different interpretations. Walley (1996) noted the IBM's frequentist coverage characteristics, though he did not mention these coverage characteristics could be propagated through mathematical calculations based on the IBM. The most immediate difference between the IBM and the c-box approach might be that IBM users must select a value for *s*. Users of the c-box approach do not need to choose such a value, as the parameter is not used in the derivation of the approach.

There are also fundamental differences. The prior and posterior structures of the IBM are credal sets, but they are rather delicate credal sets in that they consist only of beta distributions with particular, constant values of *s* (as depicted in Figure 4). A c-box is a much coarser and fuller structure. It effectively includes all the beta distributions that are in the IBM plus infinitely many other distributions that might also be considered reasonable. The choice of the beta family is of course a result of the happenstance of mathematical conjugacy between the beta distribution and binomial sampling. One notable difference and possible conceptual advantage of the c-box approach is that it does not depend on the fiction that the appropriate prior actually or necessarily has some beta shape. Thus, in contrast with the imprecise *beta* model, one might consider the c-box solution to be an imprecise model for the binomial probability, or even *the imprecise model* for the binomial probability. Such presumptuousness in doing so might eventually be forgivable if it turns out that the c-box provides a slightly tidier solution to Bayes' original problem of estimating the binomial probability.

Perhaps more important than any tidiness or even the ability to propagate the confidence interpretation through mathematical functions is the fact that the solution strategy for the inference about binomial probability can now be contextualized as an instance of a general approach based on confidence that can be applied in many other inference problems. In contrast with the IBM, which seems to be a *sui generis* solution for one parameter of one particular sampling model, the c-box solution clearly generalizes to other problems. Balch (2012) discusses these prospects.

6 Predictive Distributions and P-boxes

If the first estimation problem given a sample of observable values $X_i \sim F(\theta)$ is to characterize the

sampling or inferential uncertainty associated with a putatively fixed but unknown parameter θ governing the stochastic process that created those observable values, the second estimation problem, which is discussed in this section, is to characterize what can be inferred about a future observable value X_{n+1} that might be collected. In addition to the sampling uncertainty associated with the inference step that arises from not having measured every possible sample value, this characterization also has a component of pure aleatory uncertainty associated with the underlying stochastic process *F*.

The characterization is a predictive distribution, or more generally a predictive p-box. This output is analogous to a Bayesian posterior predictive distribution and related to prediction intervals common in frequentist analyses. Note that the output is a proper p-box because it is a collection of probability distributions constrained by a pair of bounding distributions. But this p-box is special in that it also inherits the confidence interpretation.

The predictive distribution or p-box can be understood to be, and evaluated as, the composition $F(C(\theta))$ of the distribution function *F* and the c-box *C* estimating the parameter θ. For example, the Bernoulli distribution can be composed with the c-box for the binomial probability to create the predictive p-box for the next randomly sampled Bernoulli deviate. For this case, the composition can be done analytically: Given a Bernoulli process generating zeros and ones where the probability of one is *p* which has a constant but unknown value, and *n* random observations of which *k* values are ones and $n-k$ values are zeros, the predictive p-box, i.e., the p-box estimate of the distribution for the next binary observation, is $[B(k/(n+1)), B((k+1)/(n+1))]$, where B denotes a Bernoulli distribution. Likewise, the predictive p-box for the next binomial deviate, that is, the number of ones in *N* Bernoulli trials, is [BB(*k*, $n-k$ +1, *N*), BB(*k* +1, $n-k$, *N*)] where BB denotes a beta-binomial distribution.

Straightforward Monte Carlo simulations can demonstrate that the interval $[BB_1^{-1}(\alpha), BB_2^{-1}(\beta)]$ will contain the next binomial deviate with coverage probability $\beta-\alpha$, where BB_1^{-1} and BB_2^{-1} are the quantile functions of the beta-binomial distributions BB(*k*, $n-k$ +1, *N*) and BB(*k* +1, $n-k$, *N*) respectively.

When the c-box is described numerically rather than analytically, probability bounds analysis provides for numerical composition. For one-parameter distribution families, this involves discretizing the parameter's c-box $C = [C_1(\theta), C_2(\theta)]$ into to $m+1$ equal-confidence intervals $[C_1^{-1}(i/(m+1)), C_2^{-1}((i+1)/(m+1))]$, $i = 0, 1, ..., m$, where the superscripts denote appropriate inverse or quasi-inverse functions. Each of these intervals in turn define a p-box. Each of these p-boxes is the distribution function *F* with that interval for the parameter θ. All of the p-boxes are then aggregated using stochastic mixture which reverses the dissolution into many intervals. Equal weights are used for the mixture so long as the original discretization of the c-box was into intervals with equal partitions of confidence. (For details about this operation, see sections 2.3 and 3.2.1.6 of Ferson et al. 2003.)

7 Summary and Conclusions

This paper gives a brief introduction to a new class of estimators for a broad variety of inference problems called confidence boxes (c-boxes) that both embody a traditional confidence interpretation yet also support propagation of inferential uncertainty through mathematical operations. C-boxes can be thought of as the confluence of classical notions of confidence (Neyman 1937) embodied in confidence distributions (Cox 1958) with more recent ideas about imprecise probabilities (Walley 1991) expressed as probability boxes (Ferson et al. 2003). The paper omits the derivations of the c-box solutions described by Balch (2012), but emphasizes that their statistical performance can be checked via Monte Carlo simulations and provides R functions for this purpose.

C-boxes capture much of the flexibility of Bayesian posteriors. However, by consistently supporting a Neyman confidence interpretation, c-boxes also establish a clear connection to the underlying empirical reality, a connection which both Walley (1991) and Mayo (1998) have called for. This means that engineering and statistical calculations can be constructed using c-boxes that ensure a particular standard of performance. This approach should be useful for many applications in medical statistics, engineering in novel environments, market research, survey sampling, etc., whenever statistical performance is desired but sample data are in short supply.

In the inference for the binomial probability, the c-box is very similar to the imprecise beta model (IBM, Walley 1996). However, the c-box arises in a purely frequentist framework, and it does not refer to or depend on any priors. Its results include more than beta distributions. Unlike the IBM, the c-box approach for the binomial probability has clear connections to other inference problems such as those involving normal sampling models, and the pathway for extending these solutions to other problems is much more straightforward.

Because confidence boxes can be used in subsequent calculations involving compositions and convolutions using standard methods of probability bounds analysis, and the resulting structures also have the same Neyman confidence interpretation, analysts using c-boxes will be able, both figuratively and literally, to compute with confidence. For instance, a c-box for a parameter can be composed with the distribution function of a sample model to create a p-box that characterizes the distribution

of the next sample. The result is a new type of p-box that also has the confidence interpretation. Convolutions of c-boxes yielding sums, differences or other mathematical results likewise preserve the confidence interpretation.

Point estimators ignore uncertainties altogether. Interval estimators such as confidence intervals can be unwieldy for several reasons. Even detail-rich distributional estimators like confidence distributions or Bayesian posteriors may give an incomplete characterization under demographic uncertainty when continuous parameters must be estimated from discrete data. C-boxes are more general than distributional, interval or point estimators. C-boxes can express inferential uncertainty arising from demographic uncertainty, as well as both sampling uncertainty from small sample sizes and mensurational uncertainty arising from the inability to measure quantities with infinite precision. The new estimators have the form of p-boxes, so that they may rightly be described as *p-box estimators* of parameters. C-boxes provide inferential tools to complement and support the theory of p-boxes and probability bounds analysis.

Acknowledgements

We thank Matthias Troffaes of Durham University and Jason O'Rawe, Dan Rozell and Masatoshi Sugeno of Applied Biomathematics for helpful discussions. Support was provided by the National Library of Medicine, a component of the National Institutes of Health within the United States Department of Health and Human Services, through a Small Business Innovation Research grant (RC3LM010794) to Applied Biomathematics funded under the American Recovery and Reinvestment Act. The views and opinions expressed herein are solely those of the authors and not those of the National Library of Medicine or the National Institutes of Health.

References

[1] H.R. Akçakaya (1991). A method for simulating demographic stochasticity. *Ecological Modelling* 54: 133–136.

[2] M S. Balch (2012). Mathematical foundations for a theory of confidence structures. *International Journal of Approximate Reasoning* 53: 1003–1019.

[3] T. Bayes [and R. Price] (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53: 370–418. http://rstl.royalsocietypublishing.org/content/53/370.full.pdf. Reprinted (1958). *Biometrika* 45: 296–315.

[4] J.O. Berger (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag.

[5] J.-M. Bernard (2005). An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning* 39: 123–150.

[6] C. Clopper and E.S. Pearson (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26: 404–413.

[7] D.R. Cox (1958). Some problems with statistical inference. *The Annals of Mathematical Statistics* 29: 357–372.

[8] D.R. Cox (2006). *Principles of Statistical Inference*. Cambridge University Press.

[9] A.P. Dempster (1966). New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics* 37: 355–374. http://www.stat.purdue.edu/~chuanhai/projects/DS/docs/66Annals.pdf

[10] A.P. Dempster (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics* 38: 325–339.

[11] B. Efron (1998). R.A. Fisher in the 21st century *Statistical Science* 13: 95–122.

[12] EPA [U.S. Environmental Protection Agency] (2002). Calculating upper confidence limits for exposure point concentrations at hazardous waste sites. OSWER 9285.6-10, Office of Emergency and Remedial Response, Washington, DC. http://www.epa.gov/oswer/riskassessment/pdf/ucl.pdf

[13] Ferson, S. (1996). Reliable calculation in probabilistic logic: accounting for small sample size and model uncertainty. *Intelligent Systems: A Semiotic Perspective*, NIST, Gaithersburg, MD. Pp. 115–121.

[14] S. Ferson, V. Kreinovich, L. Ginzburg, K. Sentz and D.S. Myers (2003). *Constructing Probability Boxes and Dempster–Shafer Structures*. SAND2002-4015, Sandia National Laboratories, Albuquerque, New Mexico. http://www.ramas.com/unabridged.zip

[15] S. Ferson, V. Kreinovich, J. Hajagos, W.L. Oberkampf and L. Ginzburg (2007). *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. SAND2007-0939, Sandia National Laboratories, Albuquerque, New Mexico. http://www.ramas.com/intstats.pdf

[16] R.A. Fisher (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society* 26: 528–535.

[17] R.A. Fisher (1935). The fiducial argument in statistical inference. *Annals of Eugenics B*: 391–398.

[18] B.N. Grosof (1986). An inequality paradigm for probabilistic knowledge: the logic of conditional probability intervals. *Uncertainty in Artificial Intelligence*, L.N. Kanal and J.F. Lemmer (eds.), Elsevier Science.

[19] P.M. Lee (1997). *Bayesian Statistics: An Introduction*. Arnold.

[20] D.V. Lindley (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society, Series B* 20: 102–107.

[21] D. Mayo (1996). *Error and the Growth of Experimental Knowledge*. Chicago University Press.

[22] R.E. Moore (1966). *Interval Analysis*. Prentice-Hall.

[23] J. Neyman (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society* A237: 333–380.

[24] H.T. Nguyen, V. Kreinovich, B. Wu and G. Xiang (2012). *Computing Statistics under Interval and Fuzzy Uncertainty*. Springer Verlag.

[25] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/

[26] C.P. Robert (2012). Comments on "Confidence distribution, the frequentist distribution estimator of a parameter—a review" by Min-ge Xie and Kesar Singh. *International Statistical Review* [in press] http://arxiv.org /pdf/1206.1708.pdf. See http://xianblog.wordpress.com/ 2012/06/11/confidence-distributions/

[27] T. Schweder and N.L. Hjort (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* 29: 309–332.

[28] K. Singh, M. Xie and W.E. Strawderman (2005). Combining information from independent sources through confidence distributions. *The Annals of Statistics* 33: 159–183.

[29] Student [W.S. Gosset] (1908). The probable error of a mean. *Biometrika* 6: 1–25. http://www.york.ac.uk/ depts/maths/histstat/student.pdf

[30] S.G. Vick (2002). *Degrees of Belief: Subjective Probability and Engineering Judgment*. ASCE Press, Reston, Virginia.

[31] P. Walley (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall,.

[32] P. Walley (1996). Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B* 58: 3–57.

[33] P. Walley, L. Gurrin and P. Barton (1996). Analysis of clinical data using imprecise prior probabilities. *The Statistician* 45: 457–485.

[34] R.L. Winkler, J.E. Smith and D.G. Fryback (2002). The role of informative priors in zero-numerator problems: being conservative versus being candid. *The American Statistician* 56: 1–4. See also Comments by Browne and Eddings and Reply. *The American Statistician* 56: 252–253.

[35] M. Xie and K. Singh (2012). Confidence distribution, the frequentist distribution estimator of a parameter—a review. *International Statistical Review* [in press].

[36] M. Xie, K. Singh and W.E. Strawderman (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association* 106(493): 320–333.