Imprecise predictive inference based on low stochastic structure assumptions

F.P.A. Coolen
University of Durham

Abstract

We present the outlines of a relatively new method for predictive inference. The method is closely related to standard non-parametric approaches, and can be regarded as an attempt to draw statistical conclusions while adding only a minimum of structural assumptions to data. The inferences have a frequentist nature, but are also justified from a Bayesian point of view, and could be regarded as a robust approach to some standard problems in statistics. This paper is mainly intended to highlight some recent results and to stimulate discussion on foundations of statistics and decision making, particularly on the apparent conflict between inferential methods based on all information available, including aspects that may justify the use of particular parametric models, and methods based only on available statistical data, for as far as that is possible. We briefly review some possible applications, with special attention to comparison of populations and Bayes' problem. For many details we refer to other recent papers. In the final section we briefly discuss when it may be appropriate to apply this method, and we mention some topics for future research.

Address for correspondence:
Department of Mathematical Sciences
University of Durham
Science Laboratories, South Road
Durham DH1 3LE, England
e-mail: Frank.Coolen@durham.ac.uk

1 Introduction

This paper introduces and reviews a method for predictive statistical inference, that is an attempt to learn about future observations from past observations while adding only few additional structural assumptions. The method is based on Hill's assumption $A_{(n)}$ [15], which gives a direct conditional probability [10] for a future observable random quantity, conditioned on observed values of related random quantities. In fact, this conditional probability can be used as a predictive posterior probability in a general Bayesian framework [3, 12] since Hill [17] showed that there exists a (rather complicated) prior that leads exactly to this predictive posterior. With regard to the Bayesian approach it seems sufficient to remark that $A_{(n)}$ is a De Finetti coherent procedure [8, 16, 17, 18]. Our aim, however, is to present this method unrelated to any other inferential method, as a contribution to foundations of statistics. Some references given in the paper will serve as useful starting points for comparing our method to other inferential methods.

Suppose our interest is in predictions related to a real-valued random quantity X_{n+1}, or several such random quantities X_i, $i \geq n+1$, on the basis of observed values of n such random quantities, ordered as $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$. In this paper we assume that ties do not exist (nor will do in future observations), for ease of presentation. Our results are easily generalized to allow ties [17]. The ordering of the first n observations is an essential assumption underlying Hill's assumption $A_{(n)}$, we will discuss this later. Let us denote the intervals created by the n observations by

$$I_0 = (-\infty, x_{(1)}),\ I_l = (x_{(l)}, x_{(l+1)}),\ \text{for } l = 1, \ldots, n-1, \text{ and } I_n = (x_{(n)}, \infty).$$

If we know that all random quantities are positive, the first interval will have 0 as left boundary, and for some inferences we need finite bounds for observations, in which case the interval I_0 (I_n) will be assumed to have a finite left (right) boundary.

The assumption $A_{(n)}$ is that

$$P(X_i \in I_l) = \frac{1}{n+1}, \quad \text{for } l = 0, \ldots, n,$$

and for all $i \geq n+1$. It should be remarked that $A_{(n)}$ does not assume anything else, and is clearly a post-data assumption which is related to (finite) exchangeability (see De Finetti [8, ch. 11]). Hill [16] gives a detailed presentation and discussion of $A_{(n)}$. The random quantities X_i, $i \geq n+1$, are not assumed to be conditionally independent. A simple example to appreciate the dependence related to $A_{(n)}$ is as follows: Suppose we have a single observation, x_1, providing two intervals, I_0, I_1. The assumption $A_{(1)}$ now states that $P(X_i \in I_0) = P(X_i < x_1) = \frac{1}{2}$ for all $i \geq 2$. Let us consider X_3, and in particular how probability statements about X_3 change when learning X_2. If we remain interested in the event $X_3 < x_1$, the probability $P(X_3 < x_1) = \frac{1}{2}$ will change, assuming $A_{(2)}$, according to whether the observation X_2 will be less than or greater than x_1, $P(X_3 < x_1 | X_2 < x_1) = \frac{2}{3}$ or $P(X_3 < x_1 | X_2 > x_1) = \frac{1}{3}$, respectively. This is related to the probability $P(X_3 < x_1) = \frac{1}{2}$ without conditioning on the as yet unknown X_2 by the theorem of total probability:

$$\begin{aligned} P(X_3 < x_1) &= P(X_3 < x_1 | X_2 < x_1)P(X_2 < x_1) + P(X_3 < x_1 | X_2 > x_1)P(X_2 > x_1) \\ &= \frac{2}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{2} = \frac{1}{2}. \end{aligned}$$

A direct consequence of $A_{(2)}$ is that these probabilities for X_3 keep the same values if the unknown X_2 is replaced by its observed value x_2, so $P(X_3 < x_1 | x_2 < x_1) = \frac{2}{3}$ and $P(X_3 < x_1 | x_2 > x_1) = \frac{1}{3}$. This simple example makes clear the learning process about X_3, based on appropriate assumptions $A_{(n)}$, especially the change between statements based only on x_1 or on x_1 and x_2 jointly.

De Finetti's [8] representation theorem uses a similar setting to justify a Bayesian framework to learn about an underlying parameter, and a probability distribution for that parameter, but he relies on the assumption that indeed there is an infinite sequence of random quantities involved, whereas our interest here (as in many practical situations) is in a finite number of future observations. Even more, the Bayesian approach as justified by De Finetti's [8] important results, explicitly needs a specified prior distribution, and together with the conditional independence of future observations (conditional on an unknown parameter) this adds quite a bit more structure to the data then we want and achieve. Our approach to Bayes' problem in section 3 will make this difference clear, especially when compared to standard Bayesian inference [3]. Our approach seems suitable if there is hardly any knowledge about the random quantities of interest, other than the first n observations, or, which may be more realistic, if one explicitly does not want to use such information. This may occur, for example, if one wants to study the (often hidden) effect of additional structural assumptions underlying statistical models or methods. Inferences based on such restricted knowledge have been called low structure inferences [12, sect. 2.1.2] and black-box inferences [18].

The assumption $A_{(n)}$ is not sufficient to derive precise probability results for many problems of interest. However, it does provide bounds for probabilities and expectations, as presented in this paper, and this is essentially an application of De Finetti's 'fundamental theorem of probability' [8, sect. 3.10]. The bounds that we derive are imprecise probabilities and imprecise previsions (expectations) in the sense of Walley [22]. Adopting a subjective interpretation of probability and prevision, suppose that we are interested in an uncertain quantity A. In a subjective framework [22] that is a generalization of De Finetti's theory [8], your lower prevision $E_l(A)$ for A is the supremum of all 'prices' you want to pay to get the uncertain quantity A, and your upper prevision $E_u(A)$ for A is the infimum of all 'prices' for which you want to sell A (some unit of linear utility is needed for the prices, see Walley [22, sect. 2.2]). If one is not familiar with these concepts, $E_l(A)$ and $E_u(A)$ can be considered as lower and upper bounds for the expected value of A. Imprecise probabilities are simply imprecise previsions for events, so with A an indicator function that is 1 if the event occurs and 0 else. We denote a lower probability for A by $P_l(A)$ and an upper probability for A by $P_u(A)$.

The results in this paper also have another possible interpretation (and justification), as bounds of confidence statements in a nonparametric predictive frequentist setting, see Geisser [12, sect. 2.1.2] for more details about basic results related to this interpretation. Most statistical concepts exploit (finite) exchangeability or stronger assumptions, and agree with the $A_{(n)}$-type assumptions before data are actually observed. Once data are observed, however, an assumed parametric model in effect introduces dependence between the numerical information from the data, and information about the ranks of possible future data related to the current data. This then undermines the validity of predictive statements as purely based on exchangeability. This dependence between numerical information and ranks is explicitly absent when using a nonparametric method, and is excellently shown by our inferences based on $A_{(n)}$ alone. If good reasons for a certain (family of) parametric model(s) are present, indeed one may want to use such for inferences. However, if it is purely done for mathematical convenience or necessity, one should be careful as the model assumed cancels out the weak exchangeability assumption after the data are observed, and what is the justification of the model? Often, in case of few data many models seem justifiable, whereas in case of many data no simple model seems to fit anymore. In the mean time, $A_{(n)}$-based inferences are entirely flexible, valid for few data, although high imprecision may be the fair price of only little information, and valid for many data as its assymptotics are obviously closely related to those of the empirical distribution function. The strength of the assumption $A_{(n)}$ can best be indicated by citing the final paragraph of Hill [16]: 'Let me conclude by observing that $A_{(n)}$ is supported by all of the serious approaches to statistical inference. It is Bayesian, fiducial, and even a confidence/tolerance procedure. It is simple, coherent, and plausible. It can even be argued, I believe, that $A_{(n)}$ constitutes the fundamental solution to the problem of induction'.

In section 2 and 3 we present some results for two fundamental problems in statistics, comparison of two populations and Bayes' problem. For more detailed presentation of these results we refer to recent papers, the goal of this contribution is to briefly present possible inferences and stimulate further discussion of foundations of statistics. Some further recent results and additional aspects of interest are briefly discussed in section 4.

2 Comparing Populations

In this section we consider an elementary problem in statistics: comparison of real-valued random quantities corresponding to two independent populations [4]. An often used approach is to test equality of parameters of assumed parametric models, or, in nonparametric approaches, to use the ranks of the observations (e.g. Wilcoxon's test) and base inferences on limit properties for statistics, with vague justifications for applications to finite (often small) numbers of data.

We compare real-valued random quantities, X_i and Y_j from the first and second population, respectively, by making predictive inferences for X_{n+1} and Y_{m+1} given observations $x_1 < \ldots < x_n$ and $y_1 < \ldots < y_m$, where the assumed orderings are without loss of generality. As before, we assume that there are no ties for ease of presentation. A natural way to model preference for X_{n+1} to Y_{m+1} is by the lower prevision $E_l(X_{n+1}-Y_{m+1}) > 0$, which means that we would want to get $X_{n+1}-Y_{m+1}$ for free, or even perhaps for a positive price. Another way to model such preference is by $E_l(X_{n+1}) > E_u(Y_{m+1})$, which implies $E_l(X_{n+1} - Y_{m+1}) > 0$ but is stronger than that. This second way of modelling preference has the advantage that we can analyse the populations on their own, which especially simplifies matters when dealing with independent populations, and this will be the first method presented in this section. Next to this, we can also use imprecise probabilities for the event $X_{n+1} > Y_{m+1}$, which is presented as the second method in this section.

For the first method presented in this section we must restrict the values that the random quantities can have to $l_x < X_i < r_x$, $i = 1,\ldots,n+1$, and $l_y < Y_j < r_y$, $j = 1,\ldots,m+1$, with real-valued l_x, r_x, l_y, r_y assumed to be known. Based on observations $x_1 < \ldots < x_n$ and $y_1 < \ldots < y_m$, the appropriate assumptions $A_{(n)}$ and $A_{(m)}$ give predictive probabilities for X_{n+1} and Y_{m+1}, as presented in section 1. Using these predictive probabilities, the lower prevision for X_{n+1} is easily derived by putting the mass $\frac{1}{n+1}$ as far as possible to the left in each interval, leading to

$$E_l(X_{n+1}) = \frac{1}{n+1}\left(l_x + \sum_{i=1}^{n} x_i\right).$$

Analogously, the upper prevision is derived by putting the mass to the extreme right per interval,

$$E_u(X_{n+1}) = \frac{1}{n+1}\left(r_x + \sum_{i=1}^{n} x_i\right).$$

Similarly, we get

$$E_l(Y_{m+1}) = \frac{1}{m+1}\left(l_y + \sum_{j=1}^{m} y_j\right)$$

and

$$E_u(Y_{m+1}) = \frac{1}{m+1}\left(r_y + \sum_{j=1}^{m} y_j\right).$$

Strong preference for X_{n+1} when compared to Y_{m+1} can be modelled by $E_l(X_{n+1}) > E_u(Y_{m+1})$, which leads to a sufficient condition for strong preference for X_{n+1} to Y_{m+1} given by

$$\frac{1}{n+1}\left(l_x + \sum_{i=1}^{n} x_i\right) > \frac{1}{m+1}\left(r_y + \sum_{j=1}^{m} y_j\right).$$

Analogously, a sufficient condition for strong preference for Y_{m+1} to X_{n+1} is given by

$$\frac{1}{m+1}\left(l_y + \sum_{j=1}^{m} y_j\right) > \frac{1}{n+1}\left(r_x + \sum_{i=1}^{n} x_i\right).$$

For all other situations we do not explicitly say which of the two next observations is preferred. If $m = n$ these sufficient conditions are $\sum_{i=1}^{n} x_i - \sum_{j=1}^{n} y_j > r_y - l_x$ and $\sum_{j=1}^{n} y_j - \sum_{i=1}^{n} x_i > r_x - l_y$, respectively, leaving the cases with $l_y - r_x \leq \sum_{i=1}^{n} x_i - \sum_{j=1}^{n} y_j \leq r_y - l_x$ without explicitly stated preference for either group. One may suggest that the situation $E_u(X_{n+1}) > E_u(Y_{m+1}) > E_l(X_{n+1}) > E_l(Y_{m+1})$ models a weaker form of preference for X_{n+1}, which can also be analysed easily.

It is clear that we need the restriction to bounded real-valued random quantities. If $l_x = -\infty$, the lower prevision for X_{n+1} would also be $-\infty$, while for $r_x = \infty$ the upper prevision would be ∞. Without these bounds we would never be able to express strong preference using these imprecise previsions. However, it seems that for practical applications one is usually able to state bounds for the observations, and one can easily study the effect of the choice of bounds on the final inference (see example 1 in this section).

Next we consider comparison between two groups based on imprecise probabilities, where it is not necessary to state bounds for the random quantities. For ease of notation we define $x_0 = -\infty$. Furthermore, we introduce z_i as the number of observed y-values per interval bounded by consecutive x-values, so

$$z_i = \#\{y_j | x_i < y_j < x_{i+1},\ j = 1, \dots, m\},\ i = 0, \dots, n-1,$$

and

$$z_n = \#\{y_j | x_n < y_j < \infty\}.$$

We derive lower and upper probabilities for the event $X_{n+1} > Y_{m+1}$ by looking at extreme positions of these random quantities given the predictive probabilities for the intervals [4]. The lower probability $P_l(X_{n+1} > Y_{m+1})$ is derived by putting the probability mass $\frac{1}{n+1}$ within each interval for X_{n+1} at the infimum of the values per interval ($-\infty$ for the first interval), and the mass $\frac{1}{m+1}$ within each interval for Y_{m+1} at the supremum of the values per interval (∞ for the last interval). It is clear that without additional assumptions nothing further can be said about the actual distribution of the probability mass per interval, and thus no tighter bounds can be achieved. The lower probability for the event $X_{n+1} > Y_{m+1}$ is (taking into account that $P(Y_{m+1} < -\infty) = 0$):

$$P_l(X_{n+1} > Y_{m+1}) = \frac{1}{(n+1)(m+1)} \sum_{j=0}^{n-1} (n-j) z_j.$$

The upper probability $P_u(X_{n+1} > Y_{m+1})$ is derived by putting the probability mass $\frac{1}{n+1}$ within each interval for X_{n+1} at the supremum of the values per interval (∞ for the last interval), and the mass $\frac{1}{m+1}$ within each interval for Y_{m+1} at the infimum of the values per interval ($-\infty$ for the first interval). The upper probability for the event $X_{n+1} > Y_{m+1}$ is (taking into account that $P(Y_{m+1} < \infty) = 1$):

$$P_u(X_{n+1} > Y_{m+1}) = \frac{1}{(n+1)(m+1)} \left\{ n + m + 1 + \sum_{j=0}^{n-1} (n-j) z_j \right\}.$$

These imprecise probabilities can be used for predictive comparisons between the two populations. For example, one could say that one prefers X_{n+1} to Y_{m+1} if the lower probability for the event

$X_{n+1} > Y_{m+1}$ exceeds a certain value $1-\alpha$, while Y_{m+1} is preferred to X_{n+1} if the lower probability for the event $Y_{m+1} > X_{n+1}$ exceeds the same $1-\alpha$, where (assuming that $P(X_{n+1} = Y_{m+1}) = 0$) we can use $P_l(Y_{m+1} > X_{n+1}) = 1 - P_u(X_{n+1} > Y_{m+1})$ (this follows from the underlying symmetry). To choose sample sizes when using such a form of preference one may be able to use the fact that the imprecision Δ, defined as the difference between the upper and lower probability for an event, does not depend on the data otherwise than through n and m,

$$\Delta(X_{n+1} > Y_{m+1}) = P_u(X_{n+1} > Y_{m+1}) - P_l(X_{n+1} > Y_{m+1}) = \frac{n+m+1}{(n+1)(m+1)}.$$

If we choose a value of α to express strong preference related to lower probabilities, then X_{n+1} is preferred if $P_l(X_{n+1} > Y_{m+1}) \geq 1-\alpha$ and Y_{m+1} is preferred if $P_l(Y_{m+1} > X_{n+1}) \geq 1-\alpha$. Therefore, an obvious necessary condition for strong preference of either X_{n+1} or Y_{m+1} is

$$P_l(X_{n+1} > Y_{m+1}) + P_l(Y_{m+1} > X_{n+1}) \geq 1-\alpha.$$

So

$$P_l(X_{n+1} > Y_{m+1}) + P_l(Y_{m+1} > X_{n+1}) = 1 - \Delta(X_{n+1} > Y_{m+1}) = \frac{nm}{(n+1)(m+1)} \geq 1-\alpha$$

is a necessary condition (obviously not sufficient). For example, with $m = n$ we would need to take at least

$$n > \frac{\sqrt{1-\alpha}}{1-\sqrt{1-\alpha}}.$$

For $\alpha = 0.1,\ 0.05,\ 0.01$ this implies that we can only have strong preference for either X_{n+1} or Y_{n+1} if n is at least 19, 39, 199, respectively.

In this second method the observed values x_i and y_j have not been used explicitly, only the ordering of these observations plays a role. As such this approach is closely related to standard non-parametric approaches based on ranks.

We end this section with an example to illustrate this approach. A further related example is included in the discussion in section 4.

Example 1:

The following data are presented by Sternberg, Van Kammen and Bunney [21]. In their study, 25 hospitalized schizophrenic patients were treated with antipsychotic medication, and after a period of time were classified as psychotic or nonpsychotic by hospital staff. From each patient samples of cerebrospinal fluid were taken and assayed for the dopamine *b*-hydroxylase activity. The data are given in Table 1 (the units are nmol/(ml)(h)/(mg) of protein). Interest is in the difference between the two groups of patients.

There are 15 observations for patients who were judged nonpsychotic (let us denote these as X-variables), and 10 for patients judged psychotic (Y). There is actually one tie in the X observations, but this does not complicate our analysis (one could imagine them as being very close, but not tied).

X	0.0104	0.0105	0.0112	0.0116	0.0130	0.0145	0.0154	0.0156	0.0170	0.0180
	0.0200	0.0200	0.0210	0.0230	0.0252					
Y	0.0150	0.0204	0.0208	0.0222	0.0226	0.0245	0.0270	0.0275	0.0306	0.0320

Table 1: *Dopamine b-hydroxylase activity data*

We apply the methods of this section to derive predictive inferences for X_{16} and Y_{11}, making the necessary $A_{(n)}$ assumptions. For the first method we need to assume bounds for the possible values of results according to methods X and Y, for the moment say $l_x < X_i < r_x$ and $l_y < Y_j < r_y$. The imprecise previsions for X_{16} and Y_{11} based on these data and bounds are

$$\begin{aligned} E_l(X_{16}) &= \frac{1}{16}(l_x + 0.2464) \\ E_u(X_{16}) &= \frac{1}{16}(r_x + 0.2464) \\ E_l(Y_{11}) &= \frac{1}{11}(l_y + 0.2426) \\ E_u(Y_{11}) &= \frac{1}{11}(r_y + 0.2426)\,. \end{aligned}$$

These would strongly suggest that Y_{11} exceeds X_{16}, so $E_l(Y_{11}) > E_u(X_{16})$, if $11r_x - 16l_y < 1.1712$. We cannot judge on the bounds without further knowledge about the actual situation and the meaning of the figures, but for example if the only acceptable lower bound would be $l_y = 0$, then for all upper bounds $r_x < 0.1064$, Y_{11} would indeed be suggested to exceed X_{16} on the basis of these data, according to the imprecise previsions method.

Straightforward application of the second method leads to imprecise probabilities

$$\begin{aligned} P_l(X_{16} > Y_{11}) &= 0.1136 \\ P_u(X_{16} > Y_{11}) &= 0.2614, \end{aligned}$$

so $P_l(Y_{11} > X_{16}) = 1 - 0.2614 = 0.7386$ and $P_u(Y_{11} > X_{16}) = 1 - 0.1136 = 0.8864$. The exact implications of these values should not be suggested by the statistician, but should be discussed with the topic experts. However, especially the combination of both methods in this example gives strong support to the claim that the next Y measurement will give a higher result than the next X measurement.

3 Bayes' Problem

In this section we briefly consider another fundamental problem in statistics [5]: You want to assess the probability that an event will occur, and you have information concerning past occurrences in similar situations. Pearson [19, 20] called this 'the fundamental problem of practical statistics', and the problem has a long history ([22, note 1 to sect. 5.3, p. 521]; for a detailed overview of early work on this problem, up to and including Pearson, we refer to Dale [9]). One of the earliest contributions to this problem is the paper by Bayes [1]. Coolen [5] presents a new solution to the problem derived by combining Hill's assumption $A_{(n)}$ [15] with a less restrictive version of Bayes' postulate, and this solution is briefly reviewed in this section.

Bayes' solution to the fundamental problem is based on the following postulate. Suppose an experiment consists of repeatedly throwing balls on a square table, such that the landing place of a ball is uniformly distributed over the table. Attention is restricted to the position of the ball projected on a single side of the table, say the x-axis. The position of the first ball on this x-axis plays an important role, for all other balls interest is in whether or not they land to the left of the first ball. Refer to the event that a ball lands to the left of the first ball as a success. It is easily seen that the uniformity assumption can be replaced by any other distribution, but this should still be equal for all balls, including the first ball that determines the probability of a success for the other balls.

Given the number of successes for a certain number of balls, the problem of interest is to say something about the probability of a success for a future ball, or more generally to say something about the number of successes for several future balls. When thinking about Bayes' postulate as a suitable process determining successes or failures in trials, it seems unrealistic that a probability of success in a trial is determined by a first similar trial (the position of the first ball). Our approach replaces Bayes' first ball by the assumption that there is some point S on the x-axis such that a ball to the left of S is called a success, without any further assumptions about S, and there is no explicit interest in S itself. Remark that the use of the first ball in Bayes' postulate is directly related to the role of a parameter and prior distribution in Bayesian statistical inference, where learning from observed trials is dealt with by changing beliefs, as represented by probability distributions for the parameter (prior changes to posterior), so in effect for the location of the first ball. With such a posterior distribution for the location of the first ball, all sorts of inferences for future trials are possible via conditioning on the position of the first ball. However, we step aside from this approach by suggesting that the position of this (hypothetical) first ball is not of interest apart from linking data to future observations, and our main result is that we base such a link on the assumption $A_{(n)}$ without needing this first ball. This also has the effect that Bayes' implicit assumption that all balls land on the table by the same process is replaced by the weaker assumption $A_{(n)}$, according to which these processes might be quite different in nature, but which represents that we currently have a lack of further detailed knowledge about processes per ball.

Assume an imaginary experiment, playing a role similar to Bayes' postulate, with balls thrown on a table, and only consider the projected position on the x-axis. Consider the size of the table and the place of origin to be unknown, and no knowledge is assumed about the distributions of the places where balls land on the table. For simplicity, assume that the x-axis represents the set of real numbers, $I\!R$. Different throws of balls are referred to as trials, in the same sense as De Finetti [8, sect. 1.5]: 'Trials of a phenomenon; one may allude to some exterior analogy, but one does not mean to assume anything which would imply either equal probability, or independence, or anything else of probabilistic relevance'.

Let the position on the x-axis of the ith ball be given by the real-valued random quantity X_i. Let $S \in I\!R$ and the ith trial is said to be a success if $X_i < S$ and a failure if $X_i > S$. Introduce the random quantities S_i as the indicator functions of the events $X_i < S$, so $S_i = 1$ iff $X_i < S$, and $S_i = 0$ iff $X_i > S$. Further, let $Y_i^j = \sum_{l=i}^{j} S_l$. a random quantity that counts successes. It should be emphasized that it is explicitly not assumed that $P(S_i = 1) = \theta$ for all i. If this would be the case, then obviously one would be able to learn about θ from the number of observations smaller than S. We refer to Hill [16] for further comparison with De Finetti's representation theorem, it is especially the absence here of information on the weighting distribution (prior or posterior) in the representation theorem that does not allow us to assume conditional independence for our finite (usually small) numbers of observations.

Suppose that you are interested in the number of successes in m future trials, given the number of successes in n past trials, so given $Y_1^n = s$ you are interested in Y_{n+1}^{n+m}. Realized random quantities (although we do not get the explicit numerical values) $x_1, \ldots, x_n$ can be related to future random quantities $X_{n+1}, \ldots, X_{n+m}$ via Hill's assumption $A_{(n)}$, as presented in section 1. In fact, when interested in m future observations (remember that these are not conditionally independent), a similar assumption needs to be made for each future observation consecutively, so one needs to assume $A_{(n)}, \ldots, A_{(n+m-1)}$. Hill [15] shows that the assumption $A_{(n)}$ implies $A_{(k)}$ for all $k \leq n$, so assuming $A_{(n+m-1)}$ is sufficient, but we explicitly mention all assumptions involved for a clear presentation.

In effect, the assumptions $A_{(n)}, \ldots, A_{(n+m-1)}$ imply that all orderings of the $X_i, i = 1, \ldots, n+m$, are equally likely, not only before any observations are made, but also with some observed x_i values fixed all possible orderings with regard to the other X_i's varying remain equally likely. On the basis of these assumptions only, the best we can do is deriving upper and lower bounds for $P(Y_{n+1}^{n+m} \in R_t \mid Y_1^n = s)$, where $R_t = \{r_1, \ldots, r_t\}$ with $1 \leq t \leq m+1$ and $0 \leq r_1 < r_2 < \ldots < r_t \leq m$. These bounds, which are upper and lower probabilities [22] and denoted by P_u and P_l, respectively, are determined by counting. The upper probability is derived by counting the number of orderings for which $Y_{n+1}^{n+m} \in R_t$ is possible after observing $Y_1^n = s$, the lower probability by counting the number of orderings for which $Y_{n+1}^{n+m} \in R_t$ is necessary after observing $Y_1^n = s$ [5]. Counting is more complicated here than for precise probabilities, as upper probabilities are not additive but sub-additive [22, sect. 1.6].

Defining $\binom{s+r_0}{s} = 0$, a general form for the upper bound derived in this way is [5]

$$P_u(Y_{n+1}^{n+m} \in R_t \mid Y_1^n = s) = \frac{1}{\binom{n+m}{n}} \sum_{j=1}^{t} \left[\binom{s+r_j}{s} - \binom{s+r_{j-1}}{s} \right] \binom{n-s+m-r_j}{n-s}.$$

In this sum the jth term is the number of orderings according to which s successes in the first n

observations can be followed by r_j successes but (if $j \geq 2$) not by r_{j-1} in the next m observations. A simple result (and consequence of coherence) is

$$P_l(Y_{n+1}^{n+m} \in R_t \mid Y_1^n = s) = 1 - P_u(Y_{n+1}^{n+m} \in R_t^c \mid Y_1^n = s),$$

so it is sufficient to determine only the upper bounds and give the related results. Some special cases are

$$P_u(Y_{n+1}^{n+m} = r_1 \mid Y_1^n = s) \;=\; \frac{1}{\binom{n+m}{n}} \binom{s+r_1}{s} \binom{n-s+m-r_1}{n-s} = \frac{\binom{m}{r_1}\binom{n}{s}}{\binom{n+m}{s+r_1}},$$

$$P_u(Y_{n+1}^{n+m} \in R_t \mid Y_1^n = 0) \;=\; \frac{1}{\binom{n+m}{n}} \binom{n+m-r_1}{n},$$

which equals 1 if $r_1 = 0$, and

$$P_u(Y_{n+1}^{n+m} \in R_t \mid Y_1^n = n) \;=\; \frac{1}{\binom{n+m}{n}} \binom{n+r_t}{n},$$

which equals 1 if $r_t = m$.

If attention is restricted to only a single future observation, we have

$$P_l(Y_{n+1}^{n+1} = 1 \mid Y_1^n = s) \;=\; \frac{s}{n+1}$$

and

$$P_u(Y_{n+1}^{n+1} = 1 \mid Y_1^n = s) \;=\; \frac{s+1}{n+1}.$$

As an indication of the numerical values of our imprecise probabilities we give the lower and upper probabilities for the case $n = m = 2$ in table 2.

$\times\frac{1}{6}$	$s=0$		$s=1$		$s=2$	
R_t	P_l	P_u	P_l	P_u	P_l	P_u
0	3	6	1	3	0	1
1	0	3	1	4	0	3
2	0	1	1	3	3	6
0,1	5	6	3	5	0	3
0,2	3	6	2	5	3	6
1,2	0	3	3	5	5	6

Table 2: *Bayes' problem, results for $n = 2, m = 2$*

Finally, a few results [5] related to these upper probabilities are worth to be mentioned. For all $1 \leq m_1 \leq m - 1$ we have

$$P_u(Y_{n+1}^{n+m} = r \mid Y_1^n = s) \;=\;$$
$$\sum_{j=0}^{\min(m_1,r)} P_u(Y_{n+1}^{n+m_1} = j \mid Y_1^n = s) \times P_u(Y_{n+m_1+1}^{n+m} = r - j \mid Y_1^{n+m_1} = s + j),$$

which is a convolution property, that also relates to the theorem of total probability. The same result holds for lower probabilities. If we consider the ratio of upper probabilities for Y_{n+1}^{n+m} equals $\gamma_1 m$ versus $\gamma_2 m$ (both assumed to be integers), for $\gamma_1, \gamma_2 \in [0,1]$, given the same values n and s, then the limiting value of this ratio for $m \to \infty$ equals the likelihood ratio for γ_1 versus γ_2, if the data were interpreted as s successes in n independent observations, all with identical binomial distributions with success parameter γ_1 or γ_2, respectively. Remark that, in taking the limit for $m \to \infty$, we need to assume $A_{(n)}, A_{(n+1)}, \ldots$, for the relation to infinite exchangeability we refer to Hill [16]. Finally, all predictive probabilities for this setting, related to Bayesian methods with attempts to choose non-informative priors, fall in between our upper and lower probabilities, including Laplace's rule of succession $P(Y_{n+1}^{n+1} = 1 \mid Y_1^n = s) = \frac{s+1}{n+2}$ [9].

4 Discussion

Many often applied inferential methods assume that random quantities per population are conditionally independent and identically distributed (*ciid*). This assumption is hard to justify in practice, in fact hardly ever attention is paid to it. It may well be that there are relevant covariates that are different for members of the same population, and therefore affect the random quantities of interest, but which values we do not know. De Finetti [8, ch. 11] makes it perfectly clear that the weaker assumption of (finite) exchangeability is the natural assumption to start many statistical analyses. Exchangeability can be assumed even if the populations are non-homogeneous, but we simply do not know any other relevant characteristics of the individuals than the random quantity of interest. Since many populations in problems of applied statistics are non-homogeneous, the ciid assumption may often be too strong. Hill [15, 16, 17] discusses relations between $A_{(n)}$ and exchangeability, and essentially one should be happy to use inferences based on $A_{(n)}$ whenever an exchangeability assumption prior to observing data is not strongly suggested to be inappropriate by the data, or by information related to the data. For example, if there seems to be a clear effect from the time order in which the data became available, then inferences as presented and discussed in this paper are likely to be considered inappropriate, which for example occurs clearly in time series data. In such situations, it seems obvious that one has to choose for more detailed modelling, where subjective elements seem unavoidable, although these are often hidden beneath a layer of mathematically convenient formulae. If one would like to use a highly flexible and powerful subjective framework for inferences, one may want to consider applying Bayes linear methods [14] which have the advantage that all structural judgements on which models are based have to be added explicitly.

Inferences based on $A_{(n)}$ can be used next to other inferences, based on stronger assumptions, simply as a robustness study and to analyse the effect of the assumptions underlying more complex mathematical models. One may want to use $A_{(n)}$-based inferences on their own, as for example suggested in the sections above, especially when there is very little (prior) knowledge about the random quantities of interest, or when one explicitly does not want to use any knowledge next to observed values of such random quantities. A nice discussion on exchangeability, strongly related to this aspect, is provided by Gelman, *et al.*, [13, sect. 5.2], and also example 2 can be useful to understand when our inferences may be regarded to be useful or not [4].

Example 2:

We use data on birthweights for 12 male and 12 female babies as presented by Dobson [11, p. 14], see table 3.

Male (X)	2625	2628	2795	2847	2925	2968
	2975	3163	3176	3292	3421	3473
Female (Y)	2412	2539	2729	2754	2817	2875
	2935	2991	3126	3210	3231	3317

Table 3: *Ordered birthweights (g)*

Next to this information, the original data also provided estimated gestational ages for each baby, and there seemed to be a trend of increasing birthweight with gestational age, so this can be treated as an important covariate. For our method, let us just consider the birthweights without the additional information of the estimated gestational ages. Since we know that this is a significant covariate that is not equal for all babies, an assumption of identical distributions for the weights of all boys will be hard to justify. But if we do not know the actual values of this covariate per baby, we can assume exchangeability of the weights of 13 male babies, and exchangeability of the weights of 13 female babies, before 12 weights of each actually become available. Under these assumptions, let us see what our methods, as presented in section 2, tell us about the weights X_{13} of the next boy and Y_{13} of the next girl to be weighted.

For the imprecise previsions method we assume again that there are bounds known for the possible values, using the first method of section 2, we get

$$\begin{aligned} E_l(X_{13}) &= \frac{1}{13}(l_x + 36288) \\ E_u(X_{13}) &= \frac{1}{13}(r_x + 36288) \\ E_l(Y_{13}) &= \frac{1}{13}(l_y + 34936) \\ E_u(Y_{13}) &= \frac{1}{13}(r_y + 34936)\,. \end{aligned}$$

These imprecise previsions would strongly suggest that X_{13} is greater than Y_{13}, so $E_l(X_{13}) > E_u(Y_{13})$, if $r_y - l_x < 1352$. The difference between the maximum and minimum observed weights is 1061, but probably in this case the experts would not want to assess bounds that are less than 1352 g. apart.

Suppose that bounds $l_x = l_y = 800$ and $r_x = r_y = 5000$ would be acceptable, then these data do not strongly indicate a heigher weight for the next boy than for the next girl.

The second method of section 2 gives imprecise probabilities

$$\begin{aligned} P_l(X_{13} > Y_{13}) &= \frac{86}{169} = 0.509 \\ P_u(X_{13} > Y_{13}) &= \frac{111}{169} = 0.657. \end{aligned}$$

These numbers do not indicate that the data provide very strong evidence for $X_{13} > Y_{13}$. However, if we were offered to buy a bet that pays 1 if $X_{13} > Y_{13}$ and 0 else, we would be willing to buy it for a price even slightly greater than 0.5, whereas we would only want to buy a similar bet on $Y_{13} > X_{13}$ for prices up to $1 - 0.657 = 0.343$. Combining the imprecise previsions and imprecise probabilities we could conclude that there is some evidence in favour of $X_{13} > Y_{13}$, but the evidence is not very strong.

There is an important remark to be made about this example, which may help to understand the low structure assumption used in this paper. Based on this assumption, we put a probability mass $\frac{1}{13}$ between consecutive observations when we actually have the values. For the predictive distribution for X_{13}, the weight of the next boy, our inferences imply that the probability for the event $2625 < X_{13} < 2628$ is assumed to be $\frac{1}{13}$. It is quite likely that one objects to this inference, thinking that one's actual betting behaviour would not be reflected by this number. This is precisely the situation where one feels unhappy with the bet after seeing the data, and obviously this is related to the presence of some knowledge of birthweights. The essential argument of our approach is that inferences are based on the data only, and it indicates how we can learn from the data. As mentioned before, this feature is excellently discussed by Hill [15, 16, 17], see also the papers by Coolen, *et al.*, [4, 5, 6, 7] for further discussions. If one objects to this example, then delete the nature of the numbers or think of some situation where one would really not have more information than the data only (for example weights of green (X) and yellow (Y) creatures on the planet Mars).

The inferential method presented and discussed in this paper needs to be developed further before it can actually be applied to interesting practical problems, which would be the only way to allow fair comparison to other approaches. Some further recent work is reported in Coolen and Newby [6] and Coolen and Schrijner [7]. Next to that, research is continuing on related methods for ranking and selection of populations, for multinomial inferences (with numbers of categories either known or unknown) and for censored data. Berliner and Hill [2] considered right-censored data in the same context, but as they do not allow imprecision their approach relies on additional assumptions. A main challenge is research into related methods for multidimensional random quantities, Hill [17] briefly outlines one possible approach, but further research is needed. Generally, comparison of our results so far to other methods is useful, as well via analytical methods as by means of simulations studies, or ideally in relation to simple practical applications.

Acknowledgements

I would like to thank a referee for helpful comments on a first version of this paper. Presenting this material during seminars at Eindhoven University of Technology and the University of Twente (November 1997) led to critical comments from and interesting discussions with the audiences, which have been useful for the revision of this paper. I am grateful to everyone who has reacted to the challenge inherent to this fundamental work, and I will welcome any further discussion wholeheartedly. My visit to The Netherlands in November 1997 was funded by a joint subsidy from The British Council and NWO within the framework of their 'UK-Dutch Joint Scientific Research Programme', for which I am most grateful.

References

[1] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53**, 370-418 *and* **54**, 296-325. Reproduced in: Press, S.J. (1989) *Bayesian Statistics*, Wiley, New York, 185-217.

[2] Berliner, L.M. and Hill, B.M. (1988). Bayesian nonparametric survival analysis. *Journal of the American Statistical Association*, **83**, 772-784.

[3] Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* Wiley, Chichester.

[4] Coolen, F.P.A. (1996). Comparing two populations based on low stochastic structure assumptions. *Statistics & Probability Letters*, **29**, 297-305.

[5] Coolen, F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, **36**, 349-357.

[6] Coolen, F.P.A. and Newby, M.J. (1997). Guidelines for corrective replacement based on low stochastic structure assumptions. *Quality and Reliability Engineering International*, **13**, 177-182.

[7] Coolen, F.P.A. and Schrijner, P. (1997). Which queue? A low structure analysis via bounds on expected waiting times. Submitted.

[8] De Finetti, B. (1974). *Theory of Probability* (2 volumes). Wiley, London.

[9] Dale, A.I. (1991). *History of Inverse Probability: From Thomas Bayes to Karl Pearson.* Springer-Verlag, New York.

[10] Dempster, A.P. (1963). On direct probabilities. *Journal of the Royal Statistical Society B*, **25**, 100-110.

[11] Dobson, A.J. (1983). *An Introduction to Statistical Modelling.* Chapman and Hall, London.

[12] Geisser, S. (1993). *Predictive Inference: an Introduction.* Chapman and Hall, London.

[13] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis.* Chapman and Hall, London.

[14] Goldstein, M. (1994). Revising exhangeable beliefs: subjectivist foundations for the inductive argument. In: *Aspects of Uncertainty*, eds. P.R. Freeman and A.F.M. Smith, Wiley, Chichester, 201-222.

[15] Hill, B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, **63**, 677-691.

[16] Hill, B.M. (1988). De Finetti's Theorem, Induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference. *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.). University Press, Oxford, 211-241 (with discussion).

[17] Hill, B.M. (1993). Parametric models for A_n: splitting processes and mixtures. *Journal of the Royal Statistical Society B*, **55**, 423-433.

[18] Lane, D.A. and Sudderth, W.D. (1984). Coherent predictive inference. *Sankhyā: The Indian Journal of Statistics, Series A*, **46**, 166-185.

[19] Pearson, K. (1920). The fundamental problem of practical statistics. *Biometrika* **13**, 1-16.

[20] Pearson, K. (1921). Note on the 'fundamental problem of practical statistics. *Biometrika* **13**, 300-301.

[21] Sternberg, D.E., Van Kammen, D.P. and Bunney, W.E. (1982). Schizophrenia: dopamine *b*-hydroxylase activity and treatment response. *Science*, **216**, 1423-1425.

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

ontvangen: 12-08-1997
geaccepteerd: 07-04-1998