Nonparametric predictive comparison of proportions*

F.P.A. Coolen, P. Coolen-Schrijner
Department of Mathematical Sciences, Durham University
Durham, DH1 3LE, England

Abstract

We use the lower and upper predictive probabilities from Coolen [5] to compare future numbers of successes in Bernoulli trials for different groups. We consider both pairwise and multiple comparisons. These inferences are in terms of lower and upper probabilities that the number of successes in m future trials from one group exceeds the number of successes in m future trials from another group, or such numbers from all other groups. We analyse these lower and upper probabilities via application to two data sets from the literature, and discuss the imprecision in relation to m.

Key Words: Bernoulli trials; Lower and upper probabilities; Multiple comparisons; Nonparametric predictive inference; Pairwise comparisons.

1 Introduction

Coolen [5] presented lower and upper probabilities, also called 'imprecise probabilities' [21] or 'interval probability' [22, 23], for prediction of Bernoulli random quantities, which have strong internal consistency properties within theory of interval probability [1, 6, 22, 23]. These lower and upper probabilities followed from an assumed underlying model similar to Bayes' original representation [2, 5], yet without a prior distribution, with future outcomes of random quantities related to observations by Hill's assumption $A_{(n)}$ [17, 18]. Recently, similar inferences for real-valued random quantities based on $A_{(n)}$, generally called 'nonparametric predictive inference', have been presented for a variety of statistical and operational research problems, see e.g. [7, 8, 10, 11], but Coolen's [5] lower and upper probabilities for Bernoulli random quantities have not yet been applied to further statistical inferential problems. In this paper, we present such inferences for predictive comparison of different groups of proportions based on available data consisting of numbers of successes and failures per group. Similar inferences for real-valued random quantities were presented by Coolen and van der Laan [8], and by Coolen and Yan [9] for lifetime data including right-censored observations.

*This paper was presented at the Workshop 'Interval Probability - Methodological Foundation, Theory, and Application', at the Ludwig-Maximilians University Munich, July 2004, in honour of the 75th birthday of Professor Kurt Weichselberger, to whom we dedicate this paper.

In Section 2 of this paper, we briefly review the main results from Coolen [5], and present the upper probabilities used in this paper. In Section 3 we present the lower and upper probabilities for pairwise and multiple comparisons, and in Section 4 these are illustrated and discussed via examples. We end this paper with some concluding remarks in Section 5.

2 Nonparametric predictive inference for Bernoulli quantities

In this section, we summarize results from Coolen [5] on nonparametric predictive inference for Bernoulli random quantities. We refer to [5] for justifications, which are based on representing Bernoulli data as outcomes of an experiment similar to that used by Bayes [2], with Hill's assumption $A_{(n)}$ [17, 18] used to derive direct predictive probabilities [13] for future observations using available data. The lower and upper probabilities presented in [5] fit in the framework of nonparametric predictive inference (NPI) [1], hence we also call them 'NPI-based' lower and upper probabilities. Due to the use of $A_{(n)}$ in deriving these lower and upper probabilities, they fit in a frequentist framework of statistics but can also be interpreted from Bayesian perspective [18, 19]. As they are conditional lower and upper probabilities which are introduced without reference to probabilities for the unconditional events, they can be interpreted in a way similar to Dempster's 'direct probabilities' [13]. For further discussion of such inferences see Augustin and Coolen [1].

Suppose that we have a sequence of $n+m$ exchangeable Bernoulli trials, each with 'success' and 'failure' as possible outcomes, and data consisting of s successes in n trials. Let Y_1^n denote the random number of successes in trials 1 to n, then a sufficient representation of the data for our inferences is $Y_1^n = s$, due to the assumed exchangeability of all trials. Let Y_{n+1}^{n+m} denote the random number of successes in trials $n+1$ to $n+m$. Let $R_t = \{r_1, \ldots, r_t\}$, with $1 \leq t \leq m+1$ and $0 \leq r_1 < r_2 < \ldots < r_t \leq m$, and, for ease of notation, let us define $\binom{s+r_0}{s} = 0$. Then the NPI-based upper probability [5] for the event $Y_{n+1}^{n+m} \in R_t$, given data $Y_1^n = s$, for $s \in \{0, \ldots, n\}$, is

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

The corresponding lower probability [5] is derived via

$$\underline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s) = 1 - \overline{P}(Y_{n+1}^{n+m} \in R_t^c | Y_1^n = s),$$

where $R_t^c = \{0, 1, \ldots, m\} \backslash R_t$. This relation between these upper and lower probabilities is justified in Coolen [5], and agrees with the fact that these are F-probabilities in Weichselberger's theory of interval probability [1, 22, 23], and hence are strongly internally consistent. The proof of the F-probability property is similar to the proof for this property in case of multinomial data, as given by Coolen and Augustin [6].

In this paper, we use these lower and upper probabilities for events $Y_{n+1}^{n+m} \geq y$ and $Y_{n+1}^{n+m} < y$. Due to the above relation between lower and upper probabilities, we can develop the entire theory

in this paper in terms of upper probabilities. From the upper probability for $Y_{n+1}^{n+m} \in R_t$ given above, we derive, for $y \in \{0, 1, \ldots, m\}$ and $0 < s < n$,

$$\overline{P}(Y_{n+1}^{n+m} \geq y | Y_1^n = s) = \binom{n+m}{n}^{-1} \left[\binom{s+y}{s} \binom{n-s+m-y}{n-s} + \sum_{l=y+1}^{m} \binom{s+l-1}{s-1} \binom{n-s+m-l}{n-s} \right],$$

and for $y \in \{1, \ldots, m+1\}$ and $0 < s < n$,

$$\overline{P}(Y_{n+1}^{n+m} < y | Y_1^n = s) = \binom{n+m}{n}^{-1} \left[\binom{n-s+m}{n-s} + \sum_{l=1}^{y-1} \binom{s+l-1}{s-1} \binom{n-s+m-l}{n-s} \right].$$

For $m = 1$, the two non-trivial values of these upper probabilities are $\overline{P}(Y_{n+1}^{n+1} \geq 1 | Y_1^n = s) = \frac{s+1}{n+1}$ and $\overline{P}(Y_{n+1}^{n+1} < 1 | Y_1^n = s) = \frac{n-s+1}{n+1}$.

If the observed data are all successes, so $s = n$, or all failures, so $s = 0$, then these upper probabilities are, for all $y \in \{0, 1, \ldots, m\}$,

$$\begin{aligned} \overline{P}(Y_{n+1}^{n+m} \geq y | Y_1^n = n) &= 1, \\ \overline{P}(Y_{n+1}^{n+m} \geq y | Y_1^n = 0) &= \frac{\binom{n+m-y}{n}}{\binom{n+m}{n}}, \end{aligned}$$

and for all $y \in \{1, \ldots, m+1\}$,

$$\begin{aligned} \overline{P}(Y_{n+1}^{n+m} < y | Y_1^n = n) &= \frac{\binom{n+y-1}{n}}{\binom{n+m}{n}}, \\ \overline{P}(Y_{n+1}^{n+m} < y | Y_1^n = 0) &= 1. \end{aligned}$$

We will use these lower and upper probabilities for comparison of proportions data from $k \geq 2$ different groups. Throughout, notation related to group i, $i \in \{1, \ldots, k\}$, will have an additional index i. In particular, the random number of successes in m_i future trials for group i, given s_i successes in the first n_i trials, is denoted by $Y_{i,n_i+1}^{n_i+m_i} | Y_{i,1}^{n_i} = s_i$. Throughout this paper, we will restrict attention to the same number of future trials for each group under consideration, so $m_i = m > 0$ for all i, and we discuss the influence of the choice of m in the examples in Section 4.

3 Comparison of proportions

We first consider nonparametric predictive pairwise comparison of groups of proportions data, followed by the generalization to multiple comparisons. Throughout, we assume that these groups are fully independent, in the sense that any information about one group does not influence the lower and upper probabilities for random quantities from another group. In our NPI-based method, we consider the random number of successes in m future trials per group. For pairwise comparison of groups 1 and 2, based on data $Y_{1,1}^{n_1} = s_1$ and $Y_{2,1}^{n_2} = s_2$, we use the following lower and upper

probabilities:

$$\overline{P}(Y_{1,n_1+1}^{n_1+m} > Y_{2,n_2+1}^{n_2+m} | Y_{1,1}^{n_1} = s_1, Y_{2,1}^{n_2} = s_2) =$$
$$\sum_{y=0}^{m} \overline{P}(Y_{2,n_2+1}^{n_2+m} < y | Y_{2,1}^{n_2} = s_2) \times \left[\overline{P}(Y_{1,n_1+1}^{n_1+m} \geq y | Y_{1,1}^{n_1} = s_1) - \overline{P}(Y_{1,n_1+1}^{n_1+m} \geq y+1 | Y_{1,1}^{n_1} = s_1) \right],$$

$$\underline{P}(Y_{1,n_1+1}^{n_1+m} > Y_{2,n_2+1}^{n_2+m} | Y_{1,1}^{n_1} = s_1, Y_{2,1}^{n_2} = s_2) =$$
$$\sum_{y=0}^{m} \underline{P}(Y_{2,n_2+1}^{n_2+m} < y | Y_{2,1}^{n_2} = s_2) \times \left[\underline{P}(Y_{1,n_1+1}^{n_1+m} \geq y | Y_{1,1}^{n_1} = s_1) - \underline{P}(Y_{1,n_1+1}^{n_1+m} \geq y+1 | Y_{1,1}^{n_1} = s_1) \right],$$

$$\overline{P}(Y_{1,n_1+1}^{n_1+m} \geq Y_{2,n_2+1}^{n_2+m} | Y_{1,1}^{n_1} = s_1, Y_{2,1}^{n_2} = s_2) =$$
$$\sum_{y=0}^{m} \overline{P}(Y_{2,n_2+1}^{n_2+m} \leq y | Y_{2,1}^{n_2} = s_2) \times \left[\overline{P}(Y_{1,n_1+1}^{n_1+m} \geq y | Y_{1,1}^{n_1} = s_1) - \overline{P}(Y_{1,n_1+1}^{n_1+m} \geq y+1 | Y_{1,1}^{n_1} = s_1) \right],$$

$$\underline{P}(Y_{1,n_1+1}^{n_1+m} \geq Y_{2,n_2+1}^{n_2+m} | Y_{1,1}^{n_1} = s_1, Y_{2,1}^{n_2} = s_2) =$$
$$\sum_{y=0}^{m} \underline{P}(Y_{2,n_2+1}^{n_2+m} \leq y | Y_{2,1}^{n_2} = s_2) \times \left[\underline{P}(Y_{1,n_1+1}^{n_1+m} \geq y | Y_{1,1}^{n_1} = s_1) - \underline{P}(Y_{1,n_1+1}^{n_1+m} \geq y+1 | Y_{1,1}^{n_1} = s_1) \right],$$

where the NPI-based lower and upper probabilities per group are as presented in Section 2. The justification of these lower and upper probabilities is given in the Appendix.

In the examples in Section 4, we denote the first upper probability above also by $\overline{P}(1 > 2)$, and similar for the other lower and upper probabilities. For $m = 1$, we have $\underline{P}(1 > 2) = \frac{s_1}{n_1+1} \times \frac{n_2-s_2}{n_2+1}$ and $\overline{P}(1 > 2) = \frac{s_1+1}{n_1+1} \times \frac{n_2-s_2+1}{n_2+1}$, creating an interval which contains the product of the corresponding observed proportions, $\frac{s_1}{n_1} \times \frac{n_2-s_2}{n_2}$. For pairwise comparison, it would have been sufficient to only give the first two results above, as the last two can be derived from these via $\underline{P}(A) = 1 - \overline{P}(A^c)$. However, we present all these four lower and upper probabilities for clarity, and to be able to generalize these easily to multiple comparisons for more than two groups, in which case the four corresponding lower and upper probabilities are a minimum requirement for our inferences.

For multiple comparisons of $k \geq 2$ groups, we consider the event that the number of successes in m future trials in group i is greater than (or equal to) the maximum of the number of successes in m future trials for each of the $k-1$ other groups. Such a simultaneous comparison of one group with all other groups cannot be inferred directly from pairwise comparisons [3, 8], and is often advocated for problems where one is explicitly interested in distinguishing a 'best' group, where in our NPI-based setting 'best' is formulated as the highest number of successes in m future trials per group. As before, we assume the groups to be fully independent. For this situation, the generalized version of the first upper probability given above is, for $i \in \{1, \ldots, k\}$, with $j \neq i$ used to denote $j \in \{1, \ldots, k\} \backslash i$ and $(\underline{n}, \underline{s})$ used to denote all data for the k groups,

$$\overline{P}(Y_{i,n_i+1}^{n_i+m} > \max_{j \neq i} Y_{j,n_j+1}^{n_j+m} | (\underline{n}, \underline{s})) =$$
$$\sum_{y=0}^{m} \left\{ \left[\prod_{j \neq i} \overline{P}(Y_{j,n_j+1}^{n_j+m} < y | Y_{j,1}^{n_j} = s_j) \right] \times \left[\overline{P}(Y_{i,n_i+1}^{n_i+m} \geq y | Y_{i,1}^{n_i} = s_i) - \overline{P}(Y_{i,n_i+1}^{n_i+m} \geq y+1 | Y_{i,1}^{n_i} = s_i) \right] \right\}.$$

The justification of this upper probability is again based on the result in the Appendix, together with the assumed full independence between the groups, which is also used here for the joint upper probability

$$\overline{P}\left(\bigcap_{j\neq i}\{Y_{j,n_j+1}^{n_j+m} < y | Y_{j,1}^{n_j} = s_j\}\right) = \prod_{j\neq i}\overline{P}(Y_{j,n_j+1}^{n_j+m} < y | Y_{j,1}^{n_j} = s_j),$$

which is easily shown to hold for our NPI-based upper probabilities, as the upper probabilities on the right-hand side of this equality are all actually maxima that are attained for particular configurations in the underlying model [5]. The other three lower and upper probabilities given above for pairwise comparison are similarly generalized. In the examples below, we also denote this upper probability by $\overline{P}(i > \max_{j\neq i} j)$, and similarly for the other lower and upper probabilities.

4 Examples

Example 1: Spiegelhalter, *et al.* [20] present an analysis of several data sets on mortality in heart operations on children. We use one of those data sets to illustrate our methods, without comparing it to other sources of information or discussing the quality of the data. This data set (Table 1) consists of the number n_i, for $i = 1, \ldots, 12$, of heart operations on children under 1 year old at 12 medical centres, during the period 1991 until March 1995, and the corresponding number s_i of mortalities. Table 1 also gives s_i/n_i, and the order of these proportions. An aspect of interest in the original study was whether the proportion of mortalities at Centre 1 exceeds those at the other centres.

Centre	(n_i, s_i)	s_i/n_i	order	Centre	(n_i, s_i)	s_i/n_i	order
1	(181,43)	0.2376	1	7	(253,27)	0.1067	10
2	(200,27)	0.1350	6	8	(369,57)	0.1545	5
3	(157,26)	0.1656	4	9	(214,28)	0.1308	7
4	(142,15)	0.1056	11	10	(184,31)	0.1685	2
5	(217,36)	0.1659	3	11	(740,67)	0.0905	12
6	(417,49)	0.1175	9	12	(268,32)	0.1194	8

Table 1. Heart operations mortality data.

Table 2 presents some pairwise comparisons for this data set, which illustrate several features of our inferences. We present the lower and upper probabilities for different values of m, and also the corresponding imprecision $\Delta(A) = \overline{P}(A) - \underline{P}(A)$, which provides insight into the link between these lower and upper probabilities and the amount of information available [21]. However, one should be aware that imprecision generally tends to be greater if the lower and upper probabilities are not both very close to either 0 or 1. It may be an interesting topic for future research to study different information measures related to imprecision for such inferences.

$m:$	1	3	5	10	50	250
$\overline{P}(1 > 11)$	0.220	0.457	0.578	0.730	0.964	1.000
$\underline{P}(1 > 11)$	0.215	0.447	0.566	0.716	0.957	0.999
$\Delta(1 > 11)$	0.005	0.010	0.012	0.014	0.007	0.001
$\overline{P}(1 \geq 11)$	0.931	0.882	0.875	0.894	0.981	1.000
$\underline{P}(1 \geq 11)$	0.930	0.878	0.870	0.887	0.976	1.000
$\Delta(1 \geq 11)$	0.001	0.004	0.005	0.007	0.005	0.000
$\overline{P}(3 > 5)$	0.143	0.277	0.335	0.394	0.478	0.527
$\underline{P}(3 > 5)$	0.137	0.264	0.318	0.369	0.426	0.441
$\Delta(3 > 5)$	0.006	0.013	0.017	0.025	0.052	0.086
$\overline{P}(3 \geq 5)$	0.863	0.736	0.682	0.631	0.573	0.558
$\underline{P}(3 \geq 5)$	0.858	0.724	0.666	0.606	0.522	0.473
$\Delta(3 \geq 5)$	0.005	0.012	0.016	0.025	0.051	0.085
$\overline{P}(3 > 4)$	0.153	0.327	0.421	0.536	0.763	0.902
$\underline{P}(3 > 4)$	0.146	0.311	0.397	0.502	0.707	0.846
$\Delta(3 > 4)$	0.007	0.016	0.024	0.034	0.056	0.056
$\overline{P}(3 \geq 4)$	0.913	0.828	0.795	0.777	0.835	0.915
$\underline{P}(3 \geq 4)$	0.907	0.813	0.775	0.749	0.788	0.864
$\Delta(3 \geq 4)$	0.006	0.015	0.020	0.028	0.047	0.051

Table 2. Some pairwise comparisons between centres.

For $m = 1$, the next observation from group 1 only exceeds the next observation from group 11 if these are 1 and 0, respectively. The product of the corresponding observed frequency proportions, $\frac{s_1}{n_1} = 0.2376$ and $\frac{n_{11}-s_{11}}{n_{11}} = 0.9095$, is equal to 0.2161, which is indeed between the corresponding lower and upper probabilities. Centre 1 has a far greater observed proportion of deaths than Centre 11, and there are quite many observations for each centre. Hence, it is very likely, under the assumed exchangeability assumptions which are implicit in our NPI-based inferences [5], that indeed a larger future proportion of deaths will occur for Centre 1 than for Centre 11. However, due to inherent randomness, this does not show strongly for small values of m, for which it is also quite likely to get the same proportion of deaths at both centres, illustrated by the large differences between the lower and upper probabilities for the events '$1 > 11'$ and '$1 \geq 11'$ for small m. There is a tendency for the imprecision Δ to increase in m, but for larger m it decreases again, which is due to the fact that the lower and upper probabilities both get close to 1 or 0. When comparing the lower and upper probabilities for '$1 > 11'$ and '$11 > 1'$ (these latter can be deduced from Table 2 via $\underline{P}(A) = 1 - \overline{P}(A^c)$), it is clear that indeed Centre 1 is pretty strongly expected to have a greater proportion of future deaths than Centre 11. Note, however, that the lower and upper probabilities differ substantially for different m. We belief that a table like Table 2 provides a clear picture of the predictive inference on the pairwise comparison of these centres, while choosing just a few of these lower and upper probabilities may be confusing. Of course, if one has an explicit interest in a particular value of m, e.g. if a decision needs to be made on where to send the next 10 patients for such surgery, than such decisions can be supported directly by use of the relevant lower and upper probabilities.

The observed proportions of Centres 3 and 5 are very close, which is reflected by the fact that, for $m = 250$, the corresponding lower and upper probabilities in Table 2 form intervals containing 0.5, and the fact that the lower and upper probabilities for events '$3 > 5'$ and '$5 > 3'$ are very close. In this case, the imprecision is increasing in m, which is partly due to the fact that the lower and upper probabilities move towards 0.5 for larger m, but also due to the attractive feature of our method that imprecision tends to increase with increasing m, as is e.g. illustrated by the increasing imprecision if $m = 250$ when compared to $m = 50$.

Centres 3 and 4 have the smallest numbers of observations, hence one would expect quite large imprecision in their pairwise comparison. Table 2 shows that this is indeed the case, but the move of the corresponding lower and upper probabilities towards 1 causes imprecision Δ to become smaller than the values for '$3 > 5'$ for large m.

	$m = 10$		$m = 50$	
i	$[\underline{P}, \overline{P}](i > \max_{j \neq i} j)$	$[\underline{P}, \overline{P}](i \geq \max_{j \neq i} j)$	$[\underline{P}, \overline{P}](i > \max_{j \neq i} j)$	$[\underline{P}, \overline{P}](i \geq \max_{j \neq i} j)$
1	[0.177, 0.197]	[0.369, 0.397]	[0.426, 0.482]	[0.526, 0.583]
2	[0.033, 0.039]	[0.112, 0.128]	[0.022, 0.032]	[0.041, 0.057]
3	[0.061, 0.072]	[0.173, 0.196]	[0.073, 0.098]	[0.114, 0.148]
4	[0.017, 0.022]	[0.067, 0.082]	[0.007, 0.011]	[0.014, 0.022]
5	[0.060, 0.070]	[0.173, 0.193]	[0.069, 0.089]	[0.110, 0.139]
6	[0.021, 0.024]	[0.082, 0.092]	[0.008, 0.011]	[0.017, 0.022]
7	[0.016, 0.020]	[0.067, 0.078]	[0.005, 0.008]	[0.011, 0.016]
8	[0.048, 0.054]	[0.148, 0.163]	[0.042, 0.054]	[0.073, 0.091]
9	[0.030, 0.036]	[0.104, 0.120]	[0.018, 0.026]	[0.034, 0.048]
10	[0.064, 0.074]	[0.179, 0.201]	[0.077, 0.101]	[0.121, 0.153]
11	[0.009, 0.011]	[0.046, 0.052]	[0.001, 0.002]	[0.003, 0.004]
12	[0.022, 0.027]	[0.085, 0.098]	[0.010, 0.014]	[0.020, 0.028]

Table 3. Multiple comparisons between centres.

We also illustrate the multiple comparisons lower and upper probabilities using these data. Table 3 gives all these lower and upper probabilities for $m = 10$ and $m = 50$ future observations per centre. Centre 1 has the highest observed proportion of deaths, and we see that the predictive lower and upper probabilities clearly indicate that, on the basis of our NPI model assumptions, this centre is the most likely one to lead to the highest number of deaths in m future heart operations. For smaller m there is a higher chance of two or more centres leading to the same maximum number of such deaths, hence the differences between the lower and upper probabilities for events '$i > \max_{j \neq i} j'$ and '$i \geq \max_{j \neq i} j'$ tend to decrease for larger m. For $m = 50$, the imprecision is often greater than for $m = 10$, although this is not a general effect, mostly due to the fact that the upper probability for many of these centres gets closer to zero for larger m.

Table 2 presented the pairwise comparison between Centres 3 and 5, with data $(n_3, s_3) = (157, 26)$ and $(n_5, s_5) = (217, 36)$, giving observed proportions 0.1656 and 0.1659, respectively. The lower and upper probabilities for these two centres in Table 3 are very close, but notice that those

for Centre 3 are slightly greater than those for Centre 5, and imprecision is slightly larger for Centre 3. This is caused by the fact that there are fewer observations for Centre 3 than for Centre 5, so the evidence against Centre 5 leading to the most future deaths is stronger than for Centre 3, although of course there is sufficient evidence for both these centres to make it quite unlikely that they would lead to the highest future proportion, in particular due to the data from Centre 1.

The internal consistency of our lower and upper probabilities ('coherence' [21], or 'F-probability' [23], see [1]) implies that, if we consider a collection of events that form a partition of all possible outcomes, lower probabilities must sum up to less than 1, and upper probabilities to more than 1 (or, of course, both to 1 in case there is no imprecision, which does not apply here). However, the events considered in Table 3 (per column) do not form partitions, as the events '$i > \max_{j \neq i} j'$ do not take the outcomes with more than one centre having the maximum number of future deaths into account, whereas such outcomes appear more than once in the events '$i \geq \max_{j \neq i} j'$. This implies that our lower probabilities for events '$i > \max_{j \neq i} j'$ must sum up to less than one, and our upper probabilities for events '$i \geq \max_{j \neq i} j'$ must sum up to more than one, which is indeed confirmed in Table 3.

m :	1	3	5	10	50	250
$\overline{P}(1 > \max_{j \neq 1} j)$	0.051	0.094	0.129	0.197	0.482	0.778
$\underline{P}(1 > \max_{j \neq 1} j)$	0.047	0.087	0.118	0.177	0.426	0.705
$\Delta(1 > \max_{j \neq 1} j)$	0.004	0.007	0.011	0.020	0.056	0.073
$\overline{P}(1 \geq \max_{j \neq 1} j)$	0.400	0.382	0.371	0.397	0.583	0.802
$\underline{P}(1 \geq \max_{j \neq 1} j)$	0.387	0.365	0.350	0.369	0.526	0.733
$\Delta(1 \geq \max_{j \neq 1} j)$	0.013	0.017	0.021	0.028	0.057	0.069

Table 4. Lower and upper probabilities for Centre 1 having highest future proportion.

Table 3 shows that Centre 1, with the largest observed proportion of deaths, has the largest lower and upper probabilities in these multiple comparisons. Table 4 shows that Centre 1 distinguishes itself clearer for larger m, which is due to the implicit randomness in such predictive inferences. The imprecision Δ also increases with m, an effect that is particularly clear when comparing $m = 50$ with $m = 250$, as for $m = 250$ the lower and upper probabilities are further away from 0.5, so normally one would see a little less imprecision in this latter case due to the effect that we have discussed before, whereas here imprecision is greater even though the lower and upper probabilities are closer to 1. Clearly, also the chance of other centres having precisely the same future number of deaths decreases for larger m.

Table 4 shows that the actual choice of m is important for such multiple comparisons, as these lower and upper probabilities depend strongly on m. As before, we believe that such a table, reporting lower and upper probabilities for different values of m, provides the clearest picture of our NPI-based inferences, as it can be misleading to just pick a particular value of m. Of course, if a decision is to be based on a unique value of m, the relevant NPI-based lower and upper probabilities can directly be included in the decision making process.

Example 2: Efron and Morris [14] discuss data on toxoplasmosis in El Salvador, with information for 36 cities. Congdon [4] uses a subset of data from 10 of these cities to illustrate the use of hierarchical priors in Bayesian statistics. We use the same subset (Table 5) to illustrate the methods introduced in this paper. The data represent n_i people tested in City i, for $i = 1, \ldots, 10$, of whom s_i tested positive for toxoplasmosis.

City	(n_i, s_i)	s_i/n_i	order	City	(n_i, s_i)	s_i/n_i	order
1	(51,24)	0.4706	7	6	(75,53)	0.7067	1
2	(16,7)	0.4375	8	7	(13,8)	0.6154	4
3	(82,46)	0.5610	5	8	(10,3)	0.3000	9
4	(13,9)	0.6923	2	9	(6,1)	0.1667	10
5	(43,23)	0.5349	6	10	(37,23)	0.6216	3

Table 5. Toxoplasmosis data.

m :	1	3	5	10	50	100
$\overline{P}(6 > 9)$	0.609	0.873	0.938	0.979	0.997	0.998
$\underline{P}(6 > 9)$	0.498	0.743	0.829	0.908	0.972	0.979
$\Delta(6 > 9)$	0.111	0.130	0.109	0.071	0.025	0.019
$\overline{P}(6 \geq 9)$	0.959	0.971	0.981	0.991	0.998	0.999
$\underline{P}(6 \geq 9)$	0.914	0.918	0.932	0.952	0.978	0.982
$\Delta(6 \geq 9)$	0.045	0.053	0.049	0.039	0.020	0.017
$\overline{P}(10 > 7)$	0.271	0.402	0.454	0.514	0.606	0.625
$\underline{P}(10 > 7)$	0.216	0.309	0.337	0.363	0.386	0.388
$\Delta(10 > 7)$	0.055	0.093	0.117	0.151	0.220	0.237
$\overline{P}(10 \geq 7)$	0.789	0.705	0.680	0.660	0.648	0.648
$\underline{P}(10 \geq 7)$	0.746	0.617	0.568	0.513	0.429	0.412
$\Delta(10 \geq 7)$	0.043	0.088	0.112	0.147	0.219	0.236
$\overline{P}(8 > 9)$	0.312	0.555	0.656	0.757	0.857	0.871
$\underline{P}(8 > 9)$	0.195	0.321	0.368	0.416	0.471	0.480
$\Delta(8 > 9)$	0.117	0.234	0.288	0.341	0.386	0.391
$\overline{P}(8 \geq 9)$	0.909	0.867	0.859	0.861	0.877	0.880
$\underline{P}(8 \geq 9)$	0.792	0.662	0.613	0.563	0.507	0.499
$\Delta(8 \geq 9)$	0.117	0.205	0.246	0.298	0.370	0.381

Table 6. Some pairwise comparisons between cities.

Table 6 presents some pairwise comparisons for the toxoplasmosis data. It is interesting to compare these to the similar inferences in Example 1, as there are substantially less observations per group in this data set, and the observed proportions vary far more than the proportions in Example 1. The observed proportions of cities 6 and 9 differ substantially, even the small number of observations for City 9 does not prevent the lower and upper probabilities for '$6 > 9'$ to be large. Again, we see these values increase with m to become close to 1. When compared with the similar

inferences for Example 1 (Table 2), we have substantially more imprecision due to having fewer observations per group.

For pairwise comparison of Cities 7 and 10, with observed proportions close to each other, the intervals created by the lower and upper probabilities contain 0.5 for m not too small, and the values for '$10 > 7'$ and for '$7 > 10'$ are again pretty similar, as was the case for the pairwise comparison of Centres 3 and 5 in Table 2. Due to the relatively small numbers of observations for Cities 7 and 10, the corresponding imprecisions in Table 6 are large, which is particularly clear when comparing these Δ's with those for Centres 3 and 5 in Table 2. The fact that the lower and upper probabilities for '$10 \geq 7'$ become close to those for '$10 > 7'$ for larger m is again due to the decreasing chance of getting precisely the same number of people testing positive in two cities, reflecting randomness of such events.

Cities 8 and 9 together have only 16 observations, so we expect large imprecision in their pairwise comparison. Table 6 shows that even for such few observations, the observed difference in the proportions for these cities leads to lower and upper probabilities that may reflect that more future positive tests would be expected, out of the same number m of people tested, in City 8 than in City 9. However, the imprecision is large, which nicely reflects that only little information is available for this pairwise comparison.

i	$[\underline{P}, \overline{P}](i > \max_{j \neq i} j)$	$[\underline{P}, \overline{P}](i \geq \max_{j \neq i} j)$
1	[0.002, 0.007]	[0.003, 0.011]
2	[0.004, 0.024]	[0.007, 0.033]
3	[0.010, 0.032]	[0.017, 0.048]
4	[0.204, 0.435]	[0.245, 0.488]
5	[0.009, 0.031]	[0.014, 0.045]
6	[0.208, 0.404]	[0.262, 0.473]
7	[0.079, 0.231]	[0.101, 0.274]
8	[0.001, 0.010]	[0.001, 0.013]
9	[0.000, 0.008]	[0.001, 0.011]
10	[0.054, 0.142]	[0.075, 0.183]

Table 7. Multiple comparisons between cities for $m = 50$.

Table 7 presents our NPI-based lower and upper probabilities for multiple comparisons for the 10 cities in Example 2, for $m = 50$ future observations. There is again substantially more imprecision than in Example 1 (Table 3). It is also interesting to compare Cities 4 and 6, which have the two largest observed proportions with data $(n_4, s_4) = (13, 9)$ and $(n_6, s_6) = (75, 53)$, leading to proportions 0.6923 and 0.7067, respectively. These two cities have by far the largest lower and upper probabilities in Table 7, with more imprecision for City 4 reflecting that we have fewer data for this city.

5 Concluding remarks

This paper presented lower and upper probabilities for predictive pairwise and multiple comparisons of groups of proportions, which fit in the nonparametric predictive inferential framework [1, 5] and are internally consistent [1]. As such, they can be interpreted both from frequentist and from Bayesian perspective, in the sense that if you were to buy and sell gambles according to the prices related to the subjective interpretation of these lower and upper probabilities [21], no dutch-book could be made against you at any particular moment in time. It is important to emphasize that these lower and upper probabilities always bound the corresponding precise empirical probabilities, which is an intuitively attractive property for lower and upper probabilities. These results only apply in situations where the finite exchangeability assumptions underlying our NPI-based lower and upper probabilities [5] are satisfied, so not in situations where e.g. changes to the process are likely to lead to increased future proportions of successes.

We have illustrated our methods for comparisons via two examples, and used these to discuss certain features, in particular with regard to the influence of the choice of the number of future trials considered and the imprecision in our results. Throughout, the different groups considered were assumed to be fully independent. Bayesian hierarchical models (see e.g. [4, 16]) are currently often used to model information from several sources, where an assumed model structure typically implies positive dependence between different groups. We do not suggest our method as an alternative that is in any sense better than such Bayesian methods, but we think that our method provides interesting further insights and it may be useful to compare the outcomes of both Bayesian hierarchical methods and our nonparametric predictive inferences if one wishes to compare proportions data from several groups. If these methods lead to quite different conclusions, then this may indicate that these are mostly due to the further modelling assumptions underlying the Bayesian approach, so it enables one to see the effects of such modelling assumptions more clearly. Due to the explicit predictive nature of our inferences [15], they are most suitable in situations were interest is actually in a particular number of future trials, although of course studying the inferences for varying future numbers of trials, as in the examples in Section 4, may be more generally useful. Our lower and upper probabilities can be directly included in decision problems, e.g. if one has specified utilities for future successes or failures. Due to the imprecision, it is possible that no uniquely best decision would be strongly indicated, which tends to become less likely the more data are available. If the number of future trials m would become very large, our lower and upper probabilities would start to resemble results from likelihood theory, in the sense that ratios of different lower probabilities would converge to corresponding ratios of likelihood function values (and the same would apply for upper probabilities), which is considered in more detail by Coolen [5] and which is logical since, in the limiting situation for $m \to \infty$, we would get to the same setting as used by De Finetti [12] for his well-known Representation Theorem, which underlies his justification of updating in the Bayesian framework.

Walley [21] presents a similar model for Bernoulli data, using Bayesian updating with a set of

Beta prior distributions. To define the size of this set of priors, Walley uses a parameter s, with imprecision increasing as function of s. Coolen's [5] NPI-based lower and upper probabilities often coincide with the predictive results for Walley's model for $s = 1$, but not always. We have not managed to derive general results on the comparison of these two models, but we have not found examples were our model gives less imprecision (for $s = 1$ in Walley's model). The problem for a complete analytical comparison is the fact that predictive lower and upper probabilities for Walley's model often need to be computed by numerical optimisation. As Walley's approach generalizes the standard Bayesian setting with a parametric model, it implicitly assumes infinitely exchangeable sequences of trials, which is a stronger assumption than the exchangeability assumption in our approach and may well explain why Walley's model leads to less imprecision on some occassions. Walley's model clearly has the advantage that it is parametric, hence also allows inferences which are explicitly based on the long-run probability of success, which in our method can only be formulated by using large values of m, but may therefore still depend on the particular choice of m. For values of s larger than 1, Walley's model gives lower (upper) probabilities that are often smaller (greater) than those from our model.

Our method can be generalized to similar predictive lower and upper probabilities for more general multiple comparisons inferences, e.g. subset selection, where one can both study the lower and upper probabilities that a selected subset of groups contains the group that will give most future successes, and lower and upper probabilities that all selected groups in the subset will give more future successes than the not selected groups. Such inferences are, for example, useful in screening experiments, where one wishes to reduce the number of groups with which to continue an experiment [3].

It is also relatively straightforward to use our NPI-based comparisons in case of missing data, if one does not wish to make any additional assumptions. It is obvious in our method how lower and upper probabilities are affected by whether the missing data would have been successes or failures. For example, for pairwise comparison of two groups, with observations (n_i, s_i), for $i = 1, 2$, and k_i missing observations for group i, the lower probability for group 1 leading to more successes than group 2 is derived by use of the results in Section 3, with all k_1 missing observations for group 1 assumed to be failures and all k_2 missing observations for group 2 assumed to be successes, and the other way around for the corresponding upper probability. So using the results presented in Section 3, these would be derived as

$$\underline{P}(Y_{1,n_1+k_1+1}^{n_1+k_1+m} > Y_{2,n_2+k_2+1}^{n_2+k_2+m} | Y_{1,1}^{n_1+k_1} = s_1, Y_{2,1}^{n_2+k_2} = s_2 + k_2)$$

and

$$\overline{P}(Y_{1,n_1+k_1+1}^{n_1+k_1+m} > Y_{2,n_2+k_2+1}^{n_2+k_2+m} | Y_{1,1}^{n_1+k_1} = s_1 + k_1, Y_{2,1}^{n_2+k_2} = s_2).$$

There are important research challenges to make our theory more widely applicable. For example, NPI has not yet been extended to include covariates, nor has it been extended for multi-variate data, although Hill [19] gives some indication about how the underlying $A_{(n)}$ assumption could be generalized to more dimensions. Whether or not such generalizations are feasible without further

modelling assumptions to avoid too much imprecision (due to the so-called 'curse of dimensionality') is an interesting topic for future research.

Appendix

We give a general result that justifies the lower and upper probabilities used for pairwise and multiple comparisons in Section 3. Let X and Y be independent discrete random quantities on $\{0, 1, \ldots, m\}$, with specified lower and upper probabilities for the events $X < x$ and $Y \geq y$ for $x, y \in \{0, 1, \ldots, m\}$ which are F-probabilities (see [1, 22, 23] for a definition of F-probability). The minimum upper bound for the probability for the event $Y > X$, consistent with these specified lower and upper probabilities, which therefore is itself an upper probability, is

$$\overline{P}(Y > X) = \sum_{y=0}^{m} \overline{P}_X(X < y) \times [\overline{P}_Y(Y \geq y) - \overline{P}_Y(Y \geq y + 1)].$$

The justification of this upper probability is as follows. Let M_X and M_Y be the sets of classical probability distributions for X and Y on $\{0, 1, \ldots, m\}$ which are bounded by the specified lower and upper probabilities for events $X < x$ and $Y \geq y$, respectively, so

$$\begin{aligned} M_X &= \{P_X \,|\, \underline{P}_X(X < x) \leq P_X(X < x) \leq \overline{P}_X(X < x), x \in \{0, 1, \ldots, m\}\}, \\ M_Y &= \{P_Y \,|\, \underline{P}_Y(Y \geq y) \leq P_Y(Y \geq y) \leq \overline{P}_Y(Y \geq y), y \in \{0, 1, \ldots, m\}\}. \end{aligned}$$

Then

$$\overline{P}(Y > X) = \sup_{P_X \in M_X, P_Y \in M_Y} P(Y > X).$$

Independence of X and Y and the theorem of total probability for classical probabilities give

$$P(Y > X) = \sum_{y=0}^{m} P_X(X < y) P_Y(Y = y),$$

and the upper probability for this event is derived by taking the minimum upper bound of the right-hand side over M_X and M_Y. For $P_X \in M_X$ and $P_Y \in M_Y$,

$$\begin{aligned} \sum_{y=0}^{m} P_X(X < y) P_Y(Y = y) &\leq \sum_{y=0}^{m} \overline{P}_X(X < y) P_Y(Y = y) \\ &\leq \sum_{y=0}^{m} \overline{P}_X(X < y) \times [\overline{P}_Y(Y \geq y) - \overline{P}_Y(Y \geq y + 1)], \end{aligned}$$

where the second inequality follows from the fact that $\overline{P}_X(X < y)$ is increasing in y as this specification is F-probability, hence we put the maximum possible Y-probability mass, consistent with M_Y, at the event $Y \geq m$, followed by putting the maximum possible remaining Y-probability mass, consistent with M_Y, at the event $Y \geq m - 1$, *et cetera*, which is also possible as the specification of these upper probabilities for $Y \geq y$ is F-probability. Both these inequalities are sharp and

this upper probability is clearly obtained for classical probability distributions in M_X and M_Y, so indeed the right-hand side is an upper probability $\overline{P}(Y > X)$ which is fully consistent with the specified F-probabilities for the events $X < x$ and $Y \geq y$ for $x, y \in \{0, 1, \ldots, m\}$.

The corresponding lower probability

$$\underline{P}(Y > X) = \sum_{y=0}^{m} \underline{P}(X < y) \times [\underline{P}(Y \geq y) - \underline{P}(Y \geq y + 1)]$$

is derived by similar arguments, and it is easily verified that these lower and upper probabilities for $Y > X$ are again F-probabilities. Corresponding lower and upper probabilities for the events $Y \geq X$, $X > Y$ and $X \geq Y$ are derived in the same manner.

References

[1] Augustin, T. and Coolen, F.P.A. (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, **124**, 251-272.

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

[3] Bechhofer, R.E., Santner, T.J. and Goldsman, D.M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*. Wiley, New York.

[4] Congdon, P. (2001). *Bayesian Statistical Modelling*. Wiley, Chichester.

[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 Augustin, T. (2005). Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. In: *ISIPTA'05 - Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, T. Seidenfeld, R. Nau and F.G. Cozman (eds.): http://www.sipta.org/isipta05/proceedings/index.html

[7] Coolen, F.P.A. and Coolen-Schrijner, P. (2003). A nonparametric predictive method for queues. *European Journal of Operational Research*, **145**, 425-442.

[8] Coolen, F.P.A. and van der Laan, P. (2001). Imprecise predictive selection based on low structure assumptions. *Journal of Statistical Planning and Inference*, **98**, 259-277.

[9] Coolen, F.P.A. and Yan, K.J. (2003). Nonparametric predictive comparison of two groups of lifetime data. In: *ISIPTA'03 - Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications*, J.M. Bernard, T. Seidenfeld and M. Zaffalon (eds.). Proceedings in Informatics 18, Carlton Scientific, 148-161:
http://www.carleton-scientific.com/isipta/2003-toc.html

[10] Coolen, F.P.A. and Yan, K.J. (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference 126*, 25-54.

[11] Coolen-Schrijner, P. and Coolen, F.P.A. (2004). Adaptive age replacement based on nonparametric predictive inference. *Journal of the Operational Reseach Society 55*, 1281-1297.

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

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

[14] Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, **70**, 311-319.

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

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

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

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

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

[20] Spiegelhalter, D.J., Aylin, P., Best, N.G., Evans, S.J.W. and Murray, G.D. (2002). Commissioned analysis of surgical performance using routine data: lessons from the Bristol inquiry. *Journal of the Royal Statistical Society, Series A*, **165**, 191-231.

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

[22] Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, **24**, 149-170.

[23] Weichselberger, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervalwahrscheinlichkeit as umfassendes Konzept*. Physika, Heidelberg.