Jury size and composition - a predictive approach

F.P.A. Coolen, B. Houlding, S.G. Parkinson
Durham University, UK

Abstract

We consider two basic aspects of juries that must decide on guilt verdicts, namely the size of juries and their composition in situations where society consists of sub-populations. We refer to the actual jury that needs to provide a verdict as the 'first jury', and as their judgement should reflect that of society, we consider an imaginary 'second jury' to represent society. The focus is mostly on a lower probability of a guilty verdict by the second jury, conditional on a guilty verdict by the first jury, under suitable exchangeability assumptions between this second jury and the first jury. Using a lower probability of a guilty verdict naturally provides a 'benefit of doubt to the defendant' robustness of the inference. By use of a predictive approach, no assumptions on the guilt of a defendant are required, which distinguishes this approach from those presented before. The statistical inferences used in this paper are relatively straightforward, as only cases are considered where the lower probabilities according to Coolen's Nonparametric Predictive Inference for Bernoulli random quantities [5] and Walley's Imprecise Beta Model [24, 25] coincide.

Keywords. Imprecise Beta Model, lower probability, Nonparametric Predictive Inference, representation of sub-populations.

1 Introduction

In law, the use of juries is often regarded as a natural manner for reaching a verdict, mostly used when a defendant is charged with a serious crime. In such situations, there is typically uncertainty about the guilt of the defendant, and most civilized societies only wish to convict the defendant if there is considered to be very strong evidence that the defendant committed the crime: in case there is remaining doubt, the defendant should normally be given the benefit of the doubt, and should not be convicted. Due to the presence of uncertainty, it is natural that probabilistic and statistical methods have been used to analyze several theoretical aspects of juries (e.g. [11]), and of uncertainty in law more generally (e.g. [12]). During a trial, an enormous amount of information is typically presented to a jury. Such information may consist of many facts brought alight, with different emphasis on their relevance and circumstances under which these facts did or might have occurred (or not), and the manner in which this all is presented can be very confusing to members of the jury. Clearly, this makes it difficult to translate all such information into suitable data for a statistical approach based on a full model, and the one-off nature of specific court cases appears to prevent a classical frequentist statistical approach to support jurors in reaching a verdict. From a Bayesian perspective, it would be extremely difficult to provide a detailed model a priori, as one would have to foresee all possible information that might appear in a court case, in the right order (as e.g. the defence will often adapt its strategy to counter arguments presented by the prosecutor), and based on detailed expert judgements (as, effectively, only one realization of the whole process is actually observed, so any prior information is likely to remain influential). Of course, some aspects of 'uncertainty in law' have been discussed frequently, e.g. the so-called 'prosecutor's fallacy', which is a mistake due to confusion of conditional probabilities [1]. If one would wish to use Bayesian statistical reasoning to decide on a defendant's guilt, one would also require prior probabilities on his guilt. It would not only be very difficult to assess such prior probabilities meaningfully, but any explicit quantification of a juror's prior beliefs that the defendant is guilty would be considered to be highly inappropriate. Jurors are typically not trained in law, statistics or probability, so such an approach would be deemed to fail even if suggested. It is, therefore, very difficult to even consider a suitable general way in which statistics could assist jurors with their possibly very difficult task, namely that of deducing whether or not the defendant is guilty on the basis of

all evidence presented.

In this paper, we are certainly not attempting the impossible. However, we emphasize the complexity of the use of statistical methods to support jurors on deciding their verdict, as any such use of statistics is explicitly absent in the approach presented in this paper. We do not propose a method for quantifying a 'level of certainty about guilt', and we do not require any prior thoughts about the defendant's guilt. We focus our attention on juries, and we study size of juries from a novel perspective, from which we also consider composition of juries if a population consists of recognized sub-populations. The main novelty in our approach is that nowhere any assumptions are made about the defendant's guilt, and also no attempt is made to model the complex stream of information jurors have to consider during a process. By considering a predictive criterion, which is introduced and explained in Section 2, we can still comment meaningfully on appropriateness of jury sizes from a theoretical perspective. It is important to emphasize here that we do not take practicalities of the processes used by juries to reach an overall verdict into account [14], we assume throughout this paper that each juror takes the evidence presented into account and reaches a decision without conferring with other jurors. Actually, our approach even allows the latter to take place, but as outcomes of such deliberations might depend on particular personality characteristics of individual jurors, it would make the appropriateness of the key exchangeability assumption underlying our approach (Section 2) less clear.

Section 2 provides a short discussion of a typical statistical method for inference on jury verdicts and jury size, as presented in the literature. Then it presents the main criterion and assumptions underlying our novel approach, as well as the results of our approach on jury size. In Section 3 we show how this approach can be used to decide on optimal representations of 'independent' sub-populations in a jury, our approach as presented here also has some attractive features when compared to e.g. statistical methods for stratified sampling, which we will discuss briefly in Section 4 together with some further comments. Throughout this paper, uncertainty is quantified via lower and upper probabilities, where it is particularly attractive to use lower probabilities as, for the events considered, these effectively 'give the benefit of doubt' to the defendant. As we only consider a relatively straightforward statistical model with lower and upper probabilities, we use these without many further comments. For the events considered, lower and upper probabilities from Coolen's Nonparametric Predictive Inference for Bernoulli random quantities [5] coincide with those from Walley's Imprecise Beta Model [24], which is the special case of Walley's Imprecise Dirichlet Model for the situation with only two categories [25][1].

2 Jury size

Friedman [11] discusses different jury sizes and criteria for convictions, focussing on 12 jurors, with either a 12-out-of-12 or 10-out-of-12 criterion (the latter leading to a guilty verdict if supported by at least 10 of the 12 jurors), and on 6 jurors (6-out-of-6). He emphasizes that his analysis is not based on whether or not a person is actually guilty, and he also does not make any assumptions about guilt. Instead, he focusses on the degree to which the person appears to be legally guilty or the inverse, the degree to which he can defend himself. Friedman suggests that this appearance of guilt may be considered as equivalent to the probability that an individual juror would consider the defendant guilty, and assumes that the defendant affects each of the jurors equally and independently. This allows the use of the Binomial distribution, for given number of jurors and given degree of apparent guilt, to calculate the probability of conviction. Friedman then considers the probability of conviction as a function of this degree of apparent guilt, and discusses some characteristics of several jury systems from this perspective. Clearly, the unanimous 12-out-of-12 system has a relatively low probability of conviction for values of the degree of apparent guilt which are not close to 1. Friedman's discussion is in well-known statistical terms of errors of Type I, i.e. conviction of innocent individuals, and errors of Type II, i.e. failure to convict guilty individuals. This discussion is somewhat informal due to the change from assumed (non-) guilt to degree of apparent guilt. Friedman mentions that this statistical model is based on the assumption that all jurors are unbiased and equivalent in their perception. He briefly discusses the possibility of an atypical juror, which may be a strong argument in favour of jury systems that do not require unanimity. Essential in this approach is the introduction of a parameter, ϕ say, which, although not directly observable, is assumed to have a meaningful and unambiguous interpretation, in Friedman's work it is the degree of apparent guilt and $\phi \in [0, 1]$, with $\phi = 0$ meaning that the defendant is certainly not guilty, in the sense that his innocense is absolutely certain to every juror, and $\phi = 1$ meaning that every juror is absolutely certain of the defendant's guilt.

[1] For Walley's model, the value of a further parameter s in the notation of [25] must be chosen: throughout this paper we set $s = 1$ without further mentioning, as this is the value for which the lower and upper probabilities for the events considered coincide with those from Coolen's NPI approach.

Bayesian methods in statistics provide a framework for dealing with uncertainty about parameters in a consistent manner, namely by expressing subjective beliefs about such parameters, for an assumed statistical model, via prior probability distributions, which are then combined with observed data to give the posterior probability distribution of the parameters. In many situations this seems highly sensible, although it does explicitly require information about the parameters to be taken into account. Clearly, with the parameter used by Friedman, representing the defendants degree of apparent guilt, it may be a far from trivial task to model subjective beliefs about this parameter via a probability distribution. Nevertheless, it might be considered attractive to attempt a Bayesian approach to problems on adequate jury size and composition, with a parameter representing either the defendant's guilt, or Friedman's 'appearance of guilt'. However, in addition to the need for a prior distribution on such a parameter, any such an approach would require further assumed probabilities, namely for the variety of events which can be summarized as 'juror gives correct judgement'. Not only is it extremely difficult to have meaningful information on such events, let alone to quantify the uncertainty about them, these events are also (normally) unobservable and any assigned probability values will be influential on the overall inferential results.

In this paper, we present a different approach to considerations of jury size, and jury composition (Section 3). Let us consider the main reason for the very existence of a jury: it is assumed to represent the population in the sense that its final verdict should, ideally, be in line with that of 'the population', if 'the population' were confronted with the same information from the whole process. Of course, it is difficult to formulate any such a 'verdict of an entire population', we propose the following solution. Throughout this paper, we will refer to the actual jury as JA, and we consider a second, imaginary jury JI, also selected from the general population in a similar manner as JA. We now study aspects of JA by making some suitable exchangeability assumptions, and considering predictive inferences on JI's verdict based on information from JA's verdict. In particular, we will consider the lower probability of a guilty verdict by JI, given a guilty verdict by JA. We discuss this idea in more detail at the end of this section, we first develop the idea further and consider its implications for jury size considerations.

A first possible approach would be to assume exchangeability at the level of the juries, which may be most natural if JA and JI consist of the same number of jurors and the same conviction rule (required number of jurors' guilty votes to provide an overall jury guilty verdict) applies for both. In this setting, the precise conviction rule is of no actual relevance. We consider the JA verdict as one observation of a Bernoulli random quantity, and the JI verdict as a second Bernoulli random quantity which we wish to predict, and which we assume to be exchangeable with the JA verdict. Let us denote a guilty verdict of JA (JI) by $JA{-}G$ ($JI{-}G$). Both Coolen's NPI approach for Bernoulli random quantities [5], and Walley's IBM [24, 25] give $\underline{P}(JI-G|JA-G) = 1/2$, which does not provide much useful insight in this setting, and is certainly not very strong evidence that 'the population' would consider the guilty verdict appropriate. Of course, by conjugacy the corresponding upper probability of a not-guilty verdict by JI is $1/2$, so one could argue that this would support a guilty verdict as a fair representation of the population's judgement in such a case, but as it is generally accepted (in societies that like to consider themselves 'civilized') that a defendant is only convicted in case of strong evidence, and hence that the defendant should get the benefit of the doubt, this result based on assumed exchangeability at the jury level does not appear to be strong enough as a basis for decisions. For completeness, let us also mention the corresponding upper probability $\overline{P}(JI-G|JA-G) = 1$, which seems logical in such cases where there is no evidence in the available data (here the single observation $JA-G$) that there has to be any level of doubt about the defendants guilt.

A logical alternative approach to this problem is by focussing on the votes of individual jurors, and to assume exchangeability between jurors in JA and jurors in JI. From here on, we assume such exchangeability at the level of individual jurors. Focussing on individual jurors' votes, it becomes important to consider the conviction rule applied. From a mathematical perspective, it might be of interest to study all conviction rules that can be defined, in relation to real-world law scenarios it makes sense to restrict attention to k-out-of-K rules (with $k > K/2$), where the jury verdict is 'guilty' if at least k of the K jurors vote 'guilty'. Actually, we will focus on the unanimity conviction rule ($k = K$) for guilty verdicts of JA. It will be relevant, however, to consider more general k-out-of-K rules for JI, as we use JI to reflect the population at large, and as such it might for example be of interest to know the lower probability that JI would reach a guilty verdict under a specific conviction rule, given that the jurors in JA voted 'guilty' unanimously. For even wider flexibility, we will consider scenarios under which JA and JI are not required to consist of the same number of jurors, with n jurors in JA and m jurors in JI. It should be emphasized here that $n = 12$ is the present situation in many jury systems,

although studies of effectiveness of juries consisting of 6 or 8 jurors have been reported [9, 20, 23]. We will discuss below what unanimous guilty verdicts of juries JA of some sizes other than 12 imply for juries JI.

Coolen [5] derived and justified the following general results for nonparametric predictive inference (NPI) for $m+n$ exchangeable Bernoulli random quantities. 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 for the event $Y_{n+1}^{n+m} \in R_t$, given data $Y_1^n = s$, for $s \in \{0, \ldots, n\}$, is

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

The corresponding lower probability is derived via the conjugacy property

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

where $R_t^c = \{0, 1, \ldots, m\} \backslash R_t$.

For the setting in the current paper, we are only considering data consisting of a unanimous guilty verdict of JA, so $s = n$, in which case we will denote $Y_1^n = n$ by (n, n), and the event that there are at least y successes in the following m observations, so $Y_{n+1}^{n+m} \geq y$ for which we will use the notation $Y_m \geq y$, leading to NPI lower probability, for $y = 0, 1, \ldots, m$

$$\underline{P}(Y_m \geq y | (n, n)) = 1 - \frac{(n+y-1)!m!}{(y-1)!(n+m)!} \qquad (1)$$

The corresponding upper probability is equal to 1, which is fully in line with intuition yet of little relevance for the rest of this paper. For Walley's imprecise Beta model [24], which is the special case of his Imprecise Dirichlet Model with only 2 categories [25], the lower probability for this event is identical to (1). This also coincides with the 'cautious' or 'conservative' Bayesian inference advocated by Hartigan [17] for such cases. It should be emphasized that Coolen's NPI and Walley's imprecise Beta model do not generally give identical upper and lower probabilities, with Coolen's NPI leading to slightly more imprecision for many events, due to the fact that it only assumes exchangeability of the $m+n$ random quantities involved whereas Walley's approach stays close to the robust Bayes framework [3], using a Binomial model which requires assumed embedding in an infinite sequence of such exchangeable random quantities [10].

For our jury problem, with the n jurors in JA all voting guilty and m jurors in JI, (1) provides the lower probability (according to NPI and Walley's model) that JI would also reach a guilty verdict under a y-out-of-m rule. If $m = n$ and $y = m$, so JI also requires a unanimous guilty vote to reach a guilty verdict, with JI and JA having the same number of jurors, then this lower probability is 1/2. This is naturally in agreement with the situation, briefly described above, where only exchangeability at jury level is assumed, which holds of course here due to the assumed exchangeability of jurors, and the same number of jurors and conviction rules for JA and JI in this situation[2]. More generally, if we assume that unanimity is also required for a guilty verdict of JI (so $y = m$), but we do not restrict the value of m, then the lower probability (1) is equal to $n/(n+m)$, which is increasing in n and decreasing in m, also in line with intuition.

Let us consider some numerical values of (1) for situations of relevance to our discussion on jury size. These values are explicitly given in the text below, and to aid the discussion they are also presented in Table 1.

$n = 12,\ m = 12$:					
$y =$	11	10	9	8	7
$\underline{P} =$	.761	.891	.953	.981	.993
$n = 6,\ m = 6$:					
$y =$	5	4			
$\underline{P} =$	.773	.909			
$n = 6,\ m = 12$:					
$y =$	8	7			
$\underline{P} =$	.908	.950			
$n = 24,\ m = 24$:					
$y =$	23	20	13		
$\underline{P} =$	.755	.975	.99996		
$n = 24,\ m = 12$:					
$y =$	7				
$\underline{P} =$	.9995				
$n = 12,\ m = 100$:					
$y =$	100	99	95	51	
$\underline{P} =$	.107	.204	.502	.9995	

Table 1: Some values of $\underline{P} = \underline{P}(Y_m \geq y | (n, n)))$

[2]For simplicity, we assume here that the unanimity rule actually applied for JA: if this is not the case, then one might not learn the exact number of jurors in JA that voted 'guilty' - both NPI and Walley's method can, of course, also deal with information that would appear in such cases, but we do not discuss this explicitly in this paper.

First of all, for $m = n = 12$, these lower probabilities for the values $y = 11, 10, 9, 8, 7$ are 0.761, 0.891, 0.953, 0.981, 0.993, respectively. This means that, if a 12-person jury JA reaches a unanimous guilty verdict, then the lower probability that the majority of members of a second 12-person jury, with all 24 individual jurors involved assumed to be exchangeable (with regard to their individual votes - a further discussion of this assumption is provided near the end of this section), would have reached the same guilty verdict, is very high indeed (0.993). One can interpret this as a reflection of the strength of evidence of the information in the JA guilty verdict. However, one could argue that, ideally, a substantial majority of the population should (be expected to) agree with the guilty verdict, so perhaps the values 0.891 (for $y = 10$) or 0.953 ($y = 9$) are more natural to focus on. As mentioned above, several studies have focussed, both from theoretical and practical perspectives, on juries of smaller sizes, in particular 6-person juries have been considered [9, 20, 23]. For the case where both JA and JI are 6-person juries, so with $m = n = 6$, the lower probability (1) is equal to $0.773, 0.909$ for $y = 5, 4$, respectively. So, the unanimous guilty vote of the 6-person jury JA now only implies a lower probability of 0.909 for the event that a majority of the 6-person jury JI would agree with this verdict, which is a substantial reduction from the 0.993 for the corresponding lower probability if both JA and JI consisted of 12 persons. Another method that could be used to compare actual jury sizes 12 and 6, is by considering the lower probability (1) for $n = 6$, but with $m = 12$ and $y = 7$, which is equal to 0.950. However, due to the discrete nature of these events comparisons are slightly complicated, as $m = 6$ and $y = 4$ more naturally relates to the case with $m = 12$ and $y = 8$, for the latter (still with $n = 6$) the lower probability (1) is equal to 0.908, which is very close to the 0.909 for the former case. Studies of the performance of juries of size 6 are mostly initiated by practical aspects of 12-person juries, unfortunately also including considerations of costs. From such a perspective, the increased risk of getting a JA guilty verdict under the unanimity rule for a 6-person jury, which would not be in line with the verdict of the majority of the larger population, and when compared to a 12-person jury, might need to be balanced with such cost considerations, although this would involve consideration of utilities at a level that many might find 'unethical' to do explicitly as it would require balancing between utilities of an individual (the defendant) and of society at large.

It is also interesting to see what could be gained, in terms of the lower probability (1), by doubling the JA size to $n = 24$. For $m = 24$, (1) is equal to 0.755 for $y = 23$, 0.975 for $y = 20$ and 0.99996 for $y = 13$, while for $m = 12$ the lower probability of a majority of JI jurors agreeing with the guilty verdict (so $y = 7$) is equal to 0.9995 (which one may wish to compare to the corresponding values, as mentioned above, of 0.993 and 0.950 for $n = 12$ and $n = 6$, respectively).

When considering the role of JI in representing the society at large, one can also argue that a substantially larger value of m would be appropriate. In Section 3, when considering jury composition in case of a population consisting of subgroups, we will find the use of size $m = 100$ for JI convenient. For the current scenario, $m = 100$ leads, for $n = 12$, to the following values for the lower probability (1): for $y = 100, 99, 95, 51$ we get $0.107, 0.204, 0.502, 0.9995$, respectively. Notice that this lower probability for $y = 51$, reflecting that a majority of JI will vote guilty given the unanimous guilty vote of the 12 jurors in JA, is greater than the corresponding lower probabilities for a majority of guilty votes in JI in the situations discussed above, with smaller values of m. Of course, for increasing m, NPI [5] requires an exchangeability assumption over an increasing number of random quantities. This raises the question of what happens if $m \to \infty$. For a meaningful answer, let us consider the limit of the right-hand side of (1) with $y = \theta m$ for $0 < \theta \leq 1$:

$$\lim_{m \to \infty} \left(1 - \frac{(n + \theta m - 1)! m!}{(\theta m - 1)! (n + m)!} \right) = 1 - \theta^n \qquad (2)$$

In this limiting situation (see [5] for a similar argument), the exchangeability assumption in NPI becomes 'infinite exchangeability', for which case De Finetti's Representation Theorem [10] shows that one could represent the random quantities involved as conditionally independent given a parameter, where the parameter is also a random quantity. One might see parallels between such a parameter and the above θ, but they are different, as our θ only has a meaning in the predictive inference considered, that is to specify events of interest, and is not considered to be an unknown property of the infinite sequence of future observations considered in this inference. Our inferences do not require the use of a prior distribution for θ, which would necessarily have required additional assumptions which we try to avoid. This limit $1 - \theta^n$ of $\underline{P}(Y_m \geq \theta m | (n, n))$ is decreasing in θ, which makes immediately clear that θ should not be interpreted as a limit for the proportion of guilty votes for the m jurors considered in JI. For illustration, this limiting lower probability (2) is given in Table 2, for some values of n and θ. Although these limiting values provide some insight, we find the actual inferences quite confusing as populations from which juries are selected will never be of infinite size, so restricting attention

to JI of smaller sizes, as discussed above, seems more in line with intuition.

n	$\theta = 0.50$	0.75	0.90	0.95	0.99
6	0.9844	0.8220	0.4686	0.2649	0.0585
12	0.9998	0.9683	0.7176	0.4596	0.1136
24	1.0000	0.9990	0.9292	0.7080	0.2143

Table 2: Some limiting lower probabilities (2)

Before we consider corresponding inferences on appropriate representative subgroups of different subpopulations (Section 3), we discuss the underlying exchangeability assumption between jurors in a bit more detail, also from the perspective of NPI [5] and Hill's assumption $A_{(n)}$ which is implicit in NPI.

It seems sensible to assume exchangeability of the individual jurors in JA and JI, as we did above (apart from the first considerations, when we only assumed exchangeability of the two juries JA and JI). For NPI [5], this exchangeability is actually assumed with regard to an assumed underlying representation of the Bernoulli random quantities which is very similar to the representation used by Thomas Bayes [2]. It is assumed that, corresponding to the Bernoulli random quantities, there are real-valued random quantities which are not observable, but which are so that if they exceed an unknown threshold they are 'successes', else they are 'failures'. Coolen's NPI for Bernoulli data [5] uses this representation together with Hill's assumption $A_{(n)}$[3], which effectively for this real-valued setting is a 'post-data exchangeability' assumption, meaning that the exchangeability assumption on $n+m$ random quantities still holds, for as far as prediction of m random quantities is concerned, once the values of the first n are known. This representation might be quite appropriate in a jury setting, as one could consider an underlying process where each individual juror reaches a conclusion on the strength of their believe in the guilt of the convicted person, and compares this strength to an individual 'guilt threshold value' to reach the individual vote. For the exchangeability assumption used in our approach, one could assume that the differences between each individual's strength of believe in guilt and corresponding individual guilt threshold value would be the unobservable real-valued random quantity in the assumed representation underlying NPI. Hence, we do not need to assume that all jurors would actually have the same guilt threshold value, nor that the strengths of their beliefs of guilt must be comparable. The fact that such concepts are not measurable in a meaningful manner supports the appropriateness of $A_{(n)}$ in this setting [6, 18, 19], as one never gets information that could be used to counter the underlying exchangeability assumption.

The question whether or not the exchangeability assumption is really appropriate here is quite subtle. It is again important to emphasize that we only assume exchangeability of the m and n jurors in JA and JI, which is reasonable if we have no specific information on these individuals and if we would assume that jurors in JI would be selected from the large population by the same process as used to select the jurors in JA. This, however, might not imply that these jurors are exchangeable with all members of the population, as the selection process is likely to favour or exclude some in the population. However, we believe that this issue is inherent to any selection procedure for juries, and therefore to any legal system that uses juries, and we consider it an advantage that our method does not actually need to assume such exchangeability between all members of society (the above discussion involving the limit for $m \to \infty$ was included more for its theoretical value than for its real-world relevance).

At the beginning of this section, we reviewed the approach by Friedman [11], which in a classical statistical manner focusses on errors of Type I and Type II for jury verdicts, and which makes clear the inherent difficulty when representing the defendant's guilt, or a corresponding 'degree of apparent guilt', in the statistical reasoning. The method presented in this section does not make use of any of these concepts, and only looks at jury verdicts under assumed exchangeability of jurors, so it explicitly does not add any assumption or inference on whether or not the jury is correct. It is important to emphasize this, as many might consider this a disadvantage. However, in most individual situations it will by the nature of court cases not be known whether or not the defendant is guilty, and avoiding any attempt to quantify beliefs about actual (or apparent) guilt seems to simplify the discussion in a straightforward and fair manner. Of course, methods such as Friedman [11] presented have their merits, but we believe that our method provides useful additional insights and possible arguments on appropriate jury sizes. We have only considered our approach under assumed unanimous guilty verdicts by JA. The approach is easily extended to also consider more general k-out-of-K conviction rules for JA, but as we have no ambition to propose, or even consider, an optimal rule, we do not address such different rules for JA further in this paper.

[3] We use the notation $A_{(n)}$ here generically, for inference on m future observations the actual assumption made is, in notation of Hill [18, 19] $A_{(n+m-1)}$, which also implies $A_{(l)}$ for all $l < n+m-1$ [5].

3 Jury composition

In this section, we briefly consider the interesting question of how to select representative juries from populations that consist of known separate sub-populations, where we assume independence of these sub-populations with regard to the individual verdicts of jurors from different sub-populations. We assume that the number and (relative) sizes of the sub-populations are known, and also that for each member of the population the sub-population to which they belong is known. We use the same general approach as in Section 2, with the actual jury JA and the imaginary jury JI, where the use of JI provides a convenient way for taking the relative sizes of the sub-populations into account. We assume that the individual verdicts of jurors belonging to the same sub-population are exchangeable, as before, and per sub-population we use the same lower (and upper) probabilities as in Section 2. In most of this section, we consider only two sub-populations. For more sub-populations, the general conclusion remains valid.

Let the two sub-populations be denoted by A and B, with $p_A \in (0,1)$ the proportion of the whole population that belongs to A. Let jury JA consist of n_A jurors from A and n_B from B, with $n_A + n_B = n$, and jury JI of m_A jurors from A and m_B from B, with $m_A + m_B = m$. An intuitive way to choose the numbers of jurors from each sub-population in JA, assuming that n has already been chosen, is by taking n_A as close as possible to $p_A n$, so to achieve proportional representation of the sub-populations in JA. However, if again we consider the jurors as representatives of the population, and hence of the sub-populations, this choice might not be optimal from a similar perspective as used in Section 2, namely when considering the lower probability that a second jury JI would also provide a guilty verdict if JA does so. A natural manner in which to reflect the relative sizes of the sub-populations is by choosing (approximately) the same proportions for the numbers of representatives in JI, as throughout our approach the role of the imaginary jury JI is to reflect the larger population. We saw in Section 2 that the actual choice of the size m of JI affects the predictive inferences of interest, but as we just want to introduce our approach for this setting, we will use $m = 100$ for illustrations in this section. So JI will be assumed to consist of $100p_A$ (rounded to nearest integer to give m_A) jurors from A, and $m_B = 100 - m_A$ jurors from B. For this JI, which clearly reflects the sub-populations, we now wish to choose n_A and n_B, under the assumption that $n_A + n_B = n$ and n is predetermined, such that a verdict of guilty by JA leads to maximum lower probability of a guilty verdict by JI. In this paper, we only consider unanimity conviction rules for both JA and JI in this situation, the approach is easily generalized to more general conviction rules for JA, JI or both. Due to the assumed independence of individual jurors' verdicts between jurors from JA and from JI, the lower probability of the event that all $m_A + m_B$ jurors in JI vote guilty, given all $n_A + n_B$ jurors in JA voted guilty (and under the same exchangeability assumptions per sub-population as used throughout this paper), is equal to

$$\frac{n_A}{n_A + m_A} \times \frac{n_B}{n_B + m_B}$$

By a basic exercise one can derive a general expression for the optimal choices of n_A and n_B which achieve the maximum value for this lower probability, but these do not provide much general insight, apart from the fact that the optimal fraction n_A/n is equal to $1/2$ if $p_A = 1/2$ (this is of course logical by symmetry), but will be closer to $1/2$ than p_A is in all other cases. In other words, the smaller of the two sub-populations will relatively be over-represented in JA, of course with this all under the constraint due to the discrete nature of n_A and the fact that n is likely to be small. For example, the optimal number n_A in a $n = 20$ person jury JA, for $m = 100$ (under the unanimity conviction rule for both juries), is equal to 8 for $p_A = 0.1$, 9 for $p_A = 0.2$ and for $p_A = 0.3$, and 10 for $p_A = 0.4$ and for $p_A = 0.5$. The optimal values of n_A for p_A greater than $1/2$ follow by symmetry. It might be considered to be remarkable that, for $p_A = 0.1$ and the imaginary jury JI consisting of 100 jurors (so 10 from A and 90 from B), $n_A = 8$ and $n_B = 12$ would give the optimal 20-person jury according to this predictive criterion. The lower probability optimised here is actually pretty robust if one varies n_A a little from this optimum, but it is substantially larger than if one would only select 2 jurors from A and 18 from B ('proportional representation'), namely 0.0523 versus 0.0278 for the latter case. Of course, these lower probabilities are pretty small as m is quite large, but if one relaxes the conviction rule for JI, similar results are achieved. Overall, this over-representation of smaller sub-groups is not really surprising, as the additional information from an extra juror added to a small number of jurors for a particular subgroup, in terms of the predictive power of the total information, is stronger than the corresponding information lost by reducing a larger number of jurors for the different subgroup accordingly.

We do not wish to provide a more detailed study of this approach to decisions on jury composition, as the main goal here is the introduction of this criterion using the predictive lower probability of a guilty verdict by JI, given a guilty verdict by JA, and to emphasize

the attractive role of JI in representing the population. Naturally, there are many related topics that can be studied, and for some of these we did some preliminary analyses and calculations. For example, in the situation of two sub-populations, the influence of particular choices of m and n can be considered (the over-representation of the smaller sub-population to achieve optimality holds generally), and more general conviction rules can also be studied. We calculated several cases, only relaxing the conviction rule of JI, and the over-representation of the smaller sub-population was always present, be it to a lesser extent than for the unanimity rule for JI. For example, corresponding to the case discussed above with $m = 100$ and $n = 20$, if we use the 97-out-of-100 rule for conviction by JI, then for $p_A = 0.1$ the optimal n_A is equal to 6 (instead of 8 for unanimity as discussed above). This effect seems logical, as the loss of detailed information about the sub-population A is less likely to have a substantial influence on JI's overall verdict in the latter situation. We also performed some calculations for three sub-populations, in which case also the smallest (largest) sub-population is over-represented (under-represented) in the optimal jury composition. For example, again with $n = 20$ and $m = 100$, if sub-populations A, B and C consist of 10, 10 and 80 percent of the population, then the optimal representations are 6, 6 and 8, respectively, under the unanimity conviction rule for both juries JA and JI.

4 Concluding remarks

There is a considerable literature on the use of statistical methods in relation to aspects of law, including attention to specific problems involving juries which particularly received much attention in the seventies [9, 11, 13, 14, 15, 16, 23]. In addition to these mostly theoretical studies on jury size and conviction rules, there are also many studies of actual jury behaviour, see for example Ellsworth [8] who reports on a detailed observational study with attention to a variety of practical aspects, consideration of which goes far beyond the theoretical goals of the current paper. However, the use of lower and upper probabilities [24, 26, 27] in law scenarios is, unfortunately, still pretty rare, whereas it provides an attractive method to deal with the 'benefit of doubt to the defendant' issue which in law seems to be quite generally accepted, and more appealing than perhaps in many other areas where uncertainty is quantified to enable inference and decision making. In the discussion to Walley's paper which introduced the Imprecise Dirichlet Model [25], one discussant remarked that the first ever recorded use of lower and upper probability was actually in a law problem, by Ostrogradsky. The current authors have not been able to verify this claim, yet it is of interest to mention that Ostogradsky [21, 22] did consider two types of judge ('juror' in our terminology), namely 'condemning judges' and 'acquitting judges', and assumed different probability distributions for these, considering the propensity to render a guilty verdict when the person on trial is actually innocent. He then proceeded to calculate the probability of erroneous majority judgement, and using the 'principle of insufficient reason' for the prior probability of guilt, he showed that this probability of erroneous majority judgement only depends on the difference between the numbers of condemning and acquitting judges involved. Although this does not involve, neither explicitly nor in its nature, lower and upper probabilities, the idea to study the influence on different-natured jurors would be of interest to also study from our perspective, although it could not be embedded naturally in an NPI approach as such juror characteristics would typically not be observable.

The major contributions of this paper are the novel use of an imaginary 'second' jury JI to represent the larger population in a predictive statistical framework, with the corresponding opportunity to study appropriateness of real jury (JA) sizes and conviction rules, and the fact that the inferences do not make any assumptions on actual (or apparent) guilt of the defendant and also do not even attempt to conclude on such guilt. This work can be extended in many ways, most clearly of course by studying other conviction rules for JA, JI or both. In Section 3, the predictive approach was suggested for decisions on appropriate representations of sub-populations. This problem can also be considered from the classical perspective of 'stratified sampling' [4], where one often uses criteria considering the overall variance of a random outcome. The predictive approach presented here is an attractive alternative to classical stratified sampling, and could be studied in detail for more general sampling scenarios.

This approach could also allow an alternative to traditional Type I and Type II errors, with the former formulated as the event that JI would not convict the defendant when JA does reach a guilty verdict, and the latter as the event that JI would convict the defendant when JA does not. One would be particularly interested in the upper probabilities for these events. In this paper we have focussed on the lower probability of a guilty verdict by JI, given a guilty verdict by JA, which would correspond via the conjugacy property to the upper probability of a Type I error, if the latter was defined as suggested. We have not considered the Type II error, but we acknowledge that detailed study of its upper probability could provide

useful insights into this predictive approach to issues related to juries. The goal of this paper was not to present such a detailed study, but to propose a new approach to a classical theoretical problem. The paper was also not aimed at specific real-world jury scenarios, where far more complicated issues often play a role. Nevertheless, we believe that the results from this theoretical exercise can provide new insights into practical issues related to the use of juries.

In a study of jury size and composition, one might expect a general conclusion on 'best choices'. We do not pretend to be well placed to give such advice, as our only ambition has been to introduce a novel manner for study of jury size and composition that has the advantages described above. Practical limitations make it unlikely that jury sizes in law would increase, and of course from the perspective of the defendant it seems best (under the jurors' exchangeability assumptions) to have the maximum possible number of jurors and the strictest conviction rule. However, although we addressed this problem from the perspective of juries in law, a similar approach can be used for other decision problems involving representative groups. If there is not such a clear direction in which 'benefit of doubt' should be applied, one may wish to take both lower and upper probabilities into account, but even then the predictive approach proposed in this paper appears to provide sufficient promise to warrant further study.

Acknowledgements

We are grateful to Colin Aitken, Minh Ha-Duong and Eugene Seneta for providing useful suggestions on relevant literature and copies of relevant papers, and to referees for helpful suggestions on presentation. Steven Parkinson's contribution to this research was supported by an Undergraduate Research Bursary from The Nuffield Foundation.

References

[1] C.G.G. Aitken. Interpretation of evidence, and sample size determination. In: Gastwirth, J.L. (Ed.). *Statistical Science in the Courtroom*. Springer, New York, pp. 1-24, 2000.

[2] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370-418; 54: 296-325, 1763.

[3] J.O. Berger. Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25: 303-328, 1990.

[4] W.G. Cochran. *Sampling Techniques* (3rd Ed.). Wiley, New York, 1977.

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

[6] F.P.A. Coolen. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15: 21-47, 2006.

[7] F.P.A. Coolen and P. Coolen-Schrijner. Nonparametric predictive comparison of proportions. *Journal of Statistical Planning and Inference*, 137: 23-33, 2007.

[8] P.C. Ellsworth. Are twelve heads better than one? *Law and Comtemporary Problems*, 52: 205-224, 1989.

[9] V. Fabian. On the effect of jury size. *Journal of the American Statistical Association*, 72: 535-536, 1977.

[10] B. De Finetti. *Theory of Probability*. Wiley, Chichester, 1974.

[11] H. Friedman. Trial by jury: criteria for convictions, jury size and type I and type II errors. *The American Statistician*, 26: 21-23, 1972.

[12] J.L. Gastwirth (Ed.). *Statistical Science in the Courtroom*. Springer, New York, 2000.

[13] A.E. Gelfand and H. Solomon. A study of Poisson's models for jury verdicts in criminal and civil trials. *Journal of the American Statistical Association*, 68: 271-278, 1973.

[14] A.E. Gelfand and H. Solomon. Modeling jury verdicts in the American legal system. *Journal of the American Statistical Association*, 69: 32-37, 1974.

[15] A.E. Gelfand and H. Solomon. Analyzing the decision-making process of the American jury. *Journal of the American Statistical Association*, 70: 305-310, 1975.

[16] A.E. Gelfand and H. Solomon. Comments on 'On the effect of jury size'. *Journal of the American Statistical Association*, 72: 536-537, 1977.

[17] J.A. Hartigan. *Bayes Theory*. Springer, New York, 1983.

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

[19] B.M. Hill. De Finetti's Theorem, Induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In: *Bayesian Statistics 3*, J.M. Bernardo, et al. (eds.), Oxford University Press, pp. 211-241, 1988.

[20] M. Hunter. Improving the jury system: reducing jury size. *Public Law Research Institute*, 1996. (w3.uchastings.edu/ plri/spr96tex/jurysiz.html)

[21] M.V. Ostrogradsky. Extrait d'un mémoire sur la probabilité des erreurs des tribunaux. Bulletin Scientifique, No. 3. Sciences Mathématiques et Physiques. L'Académie Impériale des Sciences de Saint-Pétersbourg, 1, pp. xix-xxv, 1834 (published in 1838).

[22] E. Seneta. M.V. Ostrogradsky as probabilist. In: *Mikhail Ostrogradsky - Honoring his Bicentenary*, A. Samoilenko and H. Syta (eds.), Institute of Mathematics, National Academy of Sciences of Ukraine, pp. 69-81, 2001.

[23] D.A. Vollrath and J.H. Davis. Jury size and decision rule. In: *The Jury: Its Role in American Society*, R.J. Simon (ed.), 1980.

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

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

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

[27] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*. Physika, Heidelberg, 2001.