Decision theory applied to selecting the winners, ranking, and classification

Nicholas T. Longford, Imperial College, London, United Kingdom

Abstract

We address the problem of selecting the best of a set of units based on a criterion variable, when its value is recorded for every unit subject to estimation, measurement, or another source of error. The solution is constructed in a decision-theoretical framework, incorporating the consequences (ramifications) of the various kinds of error that can be committed. The related problems of classifying the units to a small number of groups and ranking them are solved by a similar approach. An application is presented involving retention rates in the undergraduate courses of a university.

Keywords: classification, decision theory, empirical Bayes, latent score, league table, loss function.

N. T. Longford, Division of Medicine, Imperial College, 4th floor, 369 Fulham Road, London SW10 9NH, United Kingdom. Email: sntlnick@sntl.co.uk.
1 Introduction

Selecting the best of a set of units when information about them is subject to sampling variation or another source of error is a common task in many applications. These applications include selecting an excelling unit (school, school district, teacher or student) for a special honor or prize, awarding a contract to the bidder with the best proposal, and singling out a unit for detailed study of good practice and of antecedents of success. Solutions are often sought by applying a collection of hypothesis tests; see, for example, Goldstein and Healey (1995) and Spiegelhalter et al. (2012). The issue of multiplicity, combining the results of many hypothesis tests, is also important in the analysis of some clinical trials (Kong et al., 2005), in gene expression studies (Storey and Tibshirani, 2003), and in psychology and educational research (Shaffer, 1995). One general approach focuses on controlling the false discovery rate (Edwards and Hsu, 1983; Hsu, 1984; Benjamini, 2010).

We identify one profound weakness of hypothesis testing, namely, that it has no means of incorporating the consequences of the two kinds of error that can be committed: selecting a candidate that should not be selected (a false positive), and failing to select one that should be (a false negative). These errors occur in tandem when the selection is restricted to a single candidate. The assumption that there is a single unit that has the highest value of the relevant variable is usually quite plausible. However, multiple units or no unit may be selected. This may happen not only when there is a tie or no unit satisfies an additional constraint (implied or explicitly formulated), but also as a way of protecting the selection from excessive (expected) loss. If we are extremely averse to a false positive (inappropriate inclusion in a list), and no unit is clearly the best, then selecting no unit may be the wisest decision. In contrast, if a false negative (inappropriate exclusion from the list) is regarded as a calamity, then the best strategy is to select every unit that might conceivably be the best. In practice, the analyst’s or the client’s position (perspective, priority or value judgement) is somewhere between these two extremes. We regard it as essential to represent this position quantitatively, or at least to approximate it. It is an important input into the analysis developed in Section 2 and its extensions in the following sections.

The purpose of this paper is to present a set of methods for selecting the correct option from a finite list of available options. The options are assumed to cover all
possibilities and their pairwise overlaps are empty. One of the options has to be selected, and selecting more than one option is not possible. It is assumed that one of the options is correct and the others are all incorrect. The methods are sensitive to the consequences of the errors (choices of incorrect options). Through losses or loss functions, these consequences are ineluctable elements of the input to the analysis. On the one hand, the consequences matter; the analysis would be without a purpose, and thus inconsequential, if they did not. On the other hand, a modicum of uncertainty in specifying, and quantifying, the consequences is acceptable. Integrating them in the analysis is the main strength of our proposal over methods based on hypothesis testing. At the same time, having to establish these consequences introduces considerable difficulties and contention in some settings. In the example in Section 3 and in the application in Section 5, we show that quantifying the consequences is a worthwhile activity. It enables us to tailor the analysis to the perspectives, priorities and value judgements of the client, which a hypothesis test cannot, especially if the convention of the 5% level of significance is adopted rigidly. Inferential statements that conclude with an estimate (and the associated standard error) or a confidence interval have the same drawback — they are oblivious to the consequences of the estimation error. For a similar perspective, in the context of hypothesis testing, see Lehmann (1957).

The consequences are quantified in terms of loss, a quantity that can be treated like a monetary currency, of which we would like to spend as little as possible. A unit of loss is called a \textit{lossile}. The key properties of loss are additivity and linearity, defined by the following conditions. For any nonnegative constants $a$ and $b$, the loss of $a$ lossiles followed by the loss of $b$ lossiles is for all intended purposes equivalent to the loss of $a + b$ lossiles. Further, for any positive integer $c$, $c$ instances of loss of $a/c$ lossiles each are equivalent to a single instance of losing $a$ lossiles. These conditions make adding up losses and calculating their total and average over a set of instances (decisions) meaningful. If we replace one ‘old’ lossile (oL) as the unit of loss by a ‘new’ lossile (nL), at the rate of $r$ oL to 1 nL, as is done by a currency exchange, the nature of the problem of minimizing the loss is not altered in any way.

In everyday life and most research activities that involve uncertainty, the consequences of erroneous decisions are considered in earnest. People (decision makers) sometimes choose different options in identical settings, because they have different perspectives, priorities or value judgements. A statistical analysis that ignores these
consequences has to be supplemented by a decision (selection of a course of action). This is often left to be done by the client or sponsor of the analysis. We argue that making such a decision belongs to the remit of statistics because it entails evaluations that have a statistical essence. Also, making a decision ad hoc, after an analysis in which the consequences are ignored, is deficient (Lindley, 1985; DeGroot, 2004; Longford, 2013). In fact, most analyses cater for a perspective, but they are impoverished by not recognizing this perspective, and are made irrelevant when it differs substantially from the perspective of the client.

The problem of comparing a set of quantities $\theta_u$, $u = 1, \ldots, N$, when only their estimates $\hat{\theta}_u$ are available, has an affinity to the one-way analysis of variance. The quantities $\theta_u$ may be the within-group expectations, $\hat{\theta}_u$ the corresponding estimators (sample means) and $\sigma^2_u$ their sampling variances. In the textbook setting, $\sigma^2_u = \sigma^2 / n_u$, where $n_u$ is the sample size of group $u$ and $\sigma^2$ is the within-group variance, common to all the groups. The task is to identify the group (the winner) that has the greatest (or the smallest) expectation. The F-test for the hypothesis that the group means are identical is irrelevant for this task. Interpreting failure to reject the null hypothesis as evidence that the expectations $\theta_u$ are identical is a commonly made logical error, and selecting all units as winners in this case is very likely associated with excessive loss. If the null hypothesis is rejected, we have made no progress toward identifying the winner. The various extensions of the analysis of variance (e.g., to heteroscedastic and non-normally distributed outcomes) do not resolve this issue. So, whether this hypothesis is rejected or not, we would proceed, within the confines of hypothesis testing, by making pairwise comparisons.

If the test for comparing groups $u$ and $u'$ concludes with failure to reject the null hypothesis, then there is no obvious way to choose either $u$ or $u'$ as having the greater expectation. See Hsu (1984) for adaptations of hypothesis testing to address problems that include seeking the winner. They require considerable ingenuity and can accommodate only a simple structure of the consequences (losses); none of the loss functions introduced in Section 4.3 are allowed by Hsu (1984). See Lindley (1998) for a related criticism in the context of bioequivalence.

An approach similar to the one we apply here is presented in Longford (2014b), and most of the criticism of hypothesis testing expressed there carries over to this setting. In particular, the roles of hypothesis and alternative in a hypothesis test cannot be
interchanged, whereas in the setting of selection or comparison of two units, the two options we contemplate have equal status. Further, combining the results of several hypothesis tests is difficult because it involves a calculus of probabilities of dependent events. In contrast, expected losses are easy to combine because the calculus for both expectation and loss is linear: the relevant losses, or expected losses, are simply added up, even when the losses are correlated. See Longford (2014a) for an application to a gene expression study in which the problem of testing many hypotheses is avoided.

2 Selecting the winner and the runner-up

Suppose units $u = 1, 2, \ldots, N$ are associated with unknown values $\theta_u$ of a (latent) variable. We want to identify all the winners, defined in general as the units $u$ such that $\theta_u \geq \theta_{u'}$ for all $u' \neq u$. We make no assumptions about the distribution from which the $N$ values of $\theta_u$ are generated, other than it being continuous, so that ties (identical values of $\theta_u$ for two or more units $u$) are not a concern. In some contexts, the winner also has to satisfy an established criterion, such as $\theta_u > T$. This is implemented by regarding the threshold $T$ as an additional candidate, as a pseudo-unit $N + 1$, with its value $\theta_{N+1} = T$. If it is the winner in this extended field, then none of the original (genuine) units are.

For each unit $u = 1, \ldots, N$, we have an unbiased estimator $\hat{\theta}_u$ of $\theta_u$, with sampling variance $\sigma_u^2$. Instead of estimation error, the value $\hat{\theta}_u$ may be subject to measurement error, and then the error $\hat{\theta}_u - \theta_u$ also has zero expectation and variance $\sigma_u^2$. We assume first that each $\hat{\theta}_u$ is normally distributed and $\sigma_u^2$ is known. It is essential to distinguish between the process (and distribution) that generates the values $\theta_u$ and the sampling (estimator) distribution $\hat{\theta}_u$. The two processes can be combined in a hierarchical (two-level) model.

Treated as a variable defined for the units $u$, $\hat{\theta}_u$ is sometimes called a surrogate or proxy for $\theta_u$. It is interpreted as an observed (manifest) substitute for an unobserved (latent) variable. As in Longford (2014b), we apply the fiducial argument (Fisher, 1955; Lindley, 1958). Following the realization of the values of $\hat{\theta}_u$, we regard $\hat{\theta}_u$ as fixed values and $\theta_u$, originally unknown constants, are now random variables. The original ‘model’, $\hat{\theta}_u = \theta_u + \varepsilon_u$, is thus converted to $\theta_u = \hat{\theta}_u - \varepsilon_u$, and then $\theta_u \sim \mathcal{N}(\hat{\theta}_u, \sigma_u^2)$. All probabilistic statements (probabilities $P$ and expectations $E$) that follow
are with respect to such fiducial distributions. In the Bayesian paradigm, they are more naturally derived as posterior distributions that correspond to noninformative priors. See Zabell (1992) and Hannig (2009) for an extensive discussion of the fiducial argument.

The fiducial argument is without any controversy when \( \hat{\theta}_u \) is normally distributed and there is no restriction on \( \theta_u \) (Lindley, 1958). The argument is not valid for estimators with bounded distributions, such as the uniform or chi-square. It cannot be applied to any discrete distributions.

Denote by \( \phi \) and \( \Phi \) the density and distribution function of the standard normal distribution, \( \mathcal{N}(0,1) \). By \( W \) we denote the set of winners, with the understanding that the size of this set is much smaller than \( N \), or may even be equal to one; \( 1 \leq |W| \ll N \). By conditioning on the value of \( \theta_u \) for a given unit \( u \), we obtain the probability that \( u \) is the winner,

\[
P(u \in W) = \frac{1}{\sigma_u} \int_{-\infty}^{+\infty} \frac{\phi\{z_u(x)\}}{\Phi\{z_u(x)\}} \prod_{k=1}^{N} \Phi\{z_k(x)\} \, dx,
\]

where \( z_k(x) = (x - \hat{\theta}_k)/\sigma_k \). This integral can be evaluated numerically. Part of the integrand is a product of \( N \) terms, which may be very small when \( N \) is large. Evaluating it as

\[
\exp \left( \sum_{k=1}^{N} \log \left[ \Phi\{z_k(x)\} \right] \right)
\]

avoids working with very small (positive) values. However, modern computers and computational software implemented on them propagate errors much less in multiplication than in addition of numbers of disparate magnitudes, so this alternative is not particularly useful (Weihs, Mersmann and Ligges, 2014, Chapter 2). We evaluate the integral in (1) by simple numerical quadrature. More sophisticated alternatives, such as a weighted version of the quadrature (Abramowicz and Stegun, 1972), no longer confer any advantage with modern computing facilities. We check the precision by evaluating the integral in (1) both directly and by using the expression in (2) with 1000 and 10 000 regularly spaced quadrature points. The results obtained differ only slightly (by less than \( 10^{-5} \)). Another important check is that \( \sum_{u=1}^{N} P(u \in W) = 1 \). The evaluation is implemented in \( R \) (R Core Development Team, 2013) in the form of a user-defined function.
Suppose selecting a unit that is not the winner (a false positive) is associated with a penalty of one lossile, and not selecting the winner (a false negative) is associated with $R$ losses; $R \geq 0$. We refer to $R$ as the penalty ratio. Large $R$ ($R \gg 1$) can be interpreted as an inclusive or liberal selection, and $R \ll 1$ as a selection that is strict, cautious, or has a high standard. Note that these descriptions relate not to the quality of the unit, as characterized by $\theta_u$, but to the treatment of the uncertainty about this value ($\sigma_u$).

A unit with large $\sigma_u$ could plausibly have $\theta_u$ far greater than $\hat{\theta}_u$, so a more liberal selection, with greater $R$, would appear to treat it more favorably than a stricter selection, with smaller $R$. With smaller $R$, greater weight is given to the plausibility of a value of $\theta_u$ not much greater or even smaller than $\hat{\theta}_u$.

The expected loss if unit $u$ were included in the list is $Q_{u+} = 1 - P(u \in W)$, and it is $Q_{u-} = RP(u \in W)$ if it were not included. Thus, we include units for which the balance, defined as $B_u = Q_{u+} - Q_{u-}$, is negative — when we expect to lose less by including unit $u$ than by excluding it from the select list.

If ties of the values of $\theta_u$ can be ruled out, one might reasonably expect to end up with selecting a single unit. However, it may happen that two or more units have negative values of the balance $B_u$. This may be embarrassing at a trivial level, but denial of the uncertainty that caused this contravenes both logic and good statistical practice, and eventually undermines the integrity of the analysis. In some settings, uncertainty can be controlled by design, by arranging $\sigma^2_u$ to be smaller than a set benchmark, but $\sigma^2_u$ can usually be made very small only with an unattainably large sample size. Selecting more than one unit, even though we expect there to be only one winner, represents neither a contradiction nor a failure of the method. It may be the optimal response given the elicited penalty ratio.

A compromise can be struck by introducing a penalty for selecting more than one unit. In its general form, $S_n$ losses are incurred when the size of the select list is $n$; $|W| = n$. We set $S_1 = 0$, although the problem of minimizing the expected loss $Q_u + S_{|W|}$ is not altered by adding a constant to every element of the sequence $\{S_k\}$. A meaningful sequence $S_k$ is increasing, $0 \leq S_1 < S_2 < \ldots$, although a positive loss can be declared also when no unit is selected, $S_0 > 0$. An extreme solution is simply to select the unit with the smallest value of $B_u$. It corresponds to very large $S_k$ for all $k \neq 1$. 
As noted above, a pseudo-unit $N + 1$, with $\theta_{N+1} = T$ and $\sigma^2_{N+1} = 0$, may be introduced if we want the (worthy) winner to have a value of $\theta_u$ greater than $T$. In this case, no penalty should be set for selecting an empty set, $S_0 = 0$, even when the pseudo-unit $N + 1$ is not selected either.

The runner-up can be selected using an identity similar to (1). Let $V$ be the set of runners-up, and denote by $U_h$ the set of all units $u$ with unit $h$ removed. The probability of being a runner-up is

$$P(u \in V) = \frac{1}{\sigma_u} \sum_{h \neq u} \int_{-\infty}^{+\infty} \left[ 1 - \Phi\{z_h(x)\} \right] \prod_{k \in U_h} \frac{\phi\{z_k(x)\}}{\Phi\{z_k(x)\}} \phi\{z_u(x)\} \Phi\{z_u(x)\} dx$$

$$= \frac{1}{\sigma_u} \sum_{h \neq u} \int_{-\infty}^{+\infty} \frac{1 - \Phi\{z_h(x)\} \phi\{z_u(x)\}}{\Phi\{z_h(x)\} \Phi\{z_u(x)\}} \prod_{k=1}^{N} \Phi\{z_k(x)\} dx$$

$$= \sum_{h \neq u} P(u \in V; U_h) - (N - 1)P(u \in V), \quad (3)$$

where $P(u \in V; U_h)$ denotes the probability of $u$ being the winner in the reduced set $U_h$. The integrands in (3) correspond to the probabilities of $h$ being the overall winner and $u$ being the winner in the subset $U_h$. The integration is over the distribution of $\theta_u$. Let $R^{(2)}$ be the penalty ratio for selecting a runner-up. Thus, inappropriate selection is associated with one lossile and failure to select a unit that is a runner-up with $R^{(2)}$ lossiles. A default choice is $R^{(2)} = R$, although there may be a case for setting $1 < R^{(2)} < R$ when $R > 1$, and $1 > R^{(2)} > R$ when $R < 1$. This corresponds to being more cautious in selecting the runner(s)-up $V$ than the winner(s) $W$ when we are liberal with the latter, and being more liberal in selecting $V$ when we are strict in selecting $W$. For example, $R^{(2)} = \sqrt{R}$ satisfies this condition.

The balance (of the expected losses) for unit $u$ being a runner-up is

$$B_u^{(2)} = 1 - (R^{(2)} + 1) P(u \in V),$$

and the unit is selected as a runner-up when $B_u^{(2)} < 0$; the zero can be replaced by another value common to all $u$. A unit may be selected as both a winner and a runner-up. In this case, we are certain to have committed an error. However, the other three options, selecting the unit to one list, $W$ or $V$, or to neither, may each be associated with greater expected loss.

Probabilities for ranks other than first and second have expressions similar to (1) and (3), with respectively one and $N - 1$ integrals to be evaluated (numerically). The
number of such integrals increases rapidly toward rank $N/2$ or $N/2 - 1/2$, although some of the integrals can be omitted because their values are negligible. An alternative approach is presented in Section 4.

3 Example

In this section, we explore an example constructed specifically to study the properties of the proposed decision rule. We generated the data displayed in Figure 1 by a simple process. Details of this process are given in Appendix A. The black discs mark the estimates $\hat{\theta}_u$, $u = 1, \ldots, 24$, sorted in the ascending order. The vertical segments are delimited at $\hat{\theta}_u \pm 2\sigma_u$. Their thickness is a function of the probability of being the winner, evaluated by (1). The probability is also indicated at the bottom of the diagram, with ‘0.’ removed, to save space. Thus, there appear to be two front-runners, units 23 and 24, with respective probabilities 0.28 and 0.41.

Unit 7 has a much greater standard error $\sigma_7$ than any other unit. Its probability of being a winner, 0.08, appears to be out of proportion with its value $\hat{\theta}_7$. However, a large value of $\theta_7$ cannot be ruled out, and it could exceed even both $\theta_{23}$ and $\theta_{24}$. We
Figure 2: Balances as functions of the penalty ratio $R$, on the linear and multiplicative scales for $R$. The winner(s) are listed at the bottom in the left-hand panel.

have to distinguish between inferences about ‘Who is the winner?’ and ‘How well has a unit performed?’ Our focus in this section is on the former — we seek a dichotomous classification to winner(s) and non-winners. Solutions for the latter question include 1, a league table; 2, classification to a set number of ordinal categories; and 3, a rule for comparing any two units, A and B, and issuing one of three verdicts, $A > B$, $A < B$, or $A = B$. Section 4 deals with forming league tables (ranking) and with classification.

The balances $B_u$ are plotted as functions of the penalty ratio $R$ in Figure 2. They are linear functions of $R$ (see the left-hand panel), but the multiplicative scale, used in the right-hand panel, is more natural for $R$. The balance changes its sign for unit 24 at $R = 1.44$, for unit 23 at 2.62 (marked by vertical gray ticks), and for unit 7 at 11.44, off the horizontal scale. For most other units $u$, $B_u$ decreases slowly and intersects zero at large values of $R$.

For penalty ratio $R < 1.44$, the balances $B_u = 1 - (R + 1)P(u \in W)$ are positive for all units $u$, so no unit would be selected. For $1.44 < R < 2.62$, the balance $B_u$ is negative only for $u = 24$, so this unit would be selected on its own. For $2.62 < R < 11.44$, units 23 and 24 would be selected together, and for higher values of $R$ further units would be added to the selection, starting with unit 7. Arguably, such a high value of $R$ is appropriate only in some unusual circumstances, but no computational
or conceptual problems would arise with it. This illustrates the adaptability of the method.

This discussion suggests that $R$ does not have to be pinned down to a single (exact) value. It suffices to specify an interval, such as $(5, 10)$; we refer to it as the **plausible range** of $R$. If the same set of units is selected for all values of $R$ in this range, then we have an unequivocal solution. Otherwise, one or several units may be selected for some plausible values of $R$ but not for others. The threat of such ambiguity (impasse) encourages specifying as narrow a range of $R$ as possible. See Longford (2014b) for another application of this idea. An intermediate class may be defined for the units associated with ambiguity.

It is instructive to explore how the probabilities change when we alter the estimate or the standard error (uncertainty) of unit 24. The left-hand panel of Figure 3 displays these probabilities as functions of the change $\Delta$, when $\hat{\theta}_{24}$ is altered to $\hat{\theta}_{24} + \Delta$, for $-1 < \Delta < 1$. The probability is increasing for unit 24, and is decreasing for the other units, most sharply for the closest contender, unit 23.

The right-hand panel of Figure 3 displays the probability functions for the same four units when the standard error $\sigma_{24}$ is multiplied by a factor $\lambda \in (0.5, 2)$ and all $\hat{\theta}_u$ are held fixed at their original values. The horizontal axis is on the multiplicative scale.
The vertical axes of the two panels have the same extent. With greater uncertainty about \( \theta_{24} \), while holding all estimates \( \hat{\theta}_u \) fixed, the probability of unit 24 being a winner increases; for the other units the probability decreases. This can be loosely interpreted as follows. By allowing more extreme values of \( \theta_{24} \) in a symmetric way, the extremely large values of the distribution of \( \theta_{24} \) exercise a greater influence on the selection of the winner.

In general, it is not meaningful to compare the gradients of the functions for a unit in the two panels, even though both functions have very little curvature. However, we can state that the change of the estimate \( \hat{\theta}_{24} \) by \( \Delta = 0.1 \) is associated with a change in the probability, by about 0.02, similar to the change brought on by inflating the standard error \( \sigma_{24} \) by about 40% or reducing it by 28.5%; that is, multiplying or dividing \( \sigma_{24}^2 \) by \( \lambda^2 = 1.4^2 = 2.0 \).

Fiducial probabilities do not have a straightforward interpretation in terms of replications because they are conditional on the outcomes. Their validity in the case of normally distributed outcomes is established by reference to the (Bayesian) posterior distribution with a noninformative prior. For the normal distribution, this prior, with a constant density on \((-\infty, +\infty)\), is improper, but the inferences based on it are in our context valid, in the Bayesian sense. This discussion is redundant for an analyst committed to the Bayesian paradigm, but we want to point out that the method can be applied also in the frequentist perspective.

That the sampling variance of the estimators \( \hat{\theta}_u \) is known is often a simplifying approximation to the reality. The consequences of the error made by such an approximation can be established by simulations. Difficulties arise when the departures from the assumption of known variances \( \sigma_u^2 \), \( u = 1, \ldots, N \), may be in a wide variety of directions. Such departures and their consequences have to be considered in a specific setting, linked to a particular application.

We present the results of a set of simulations for a dataset generated by the same process as the dataset given in Figure 1. We use the evaluations in (1), but generate sets of \( N \) estimates as independent draws from scaled \( t \) distributions with 10 degrees of freedom in one simulation and 30 in another. The \( t \) distributions are scaled so as to have the prescribed standard deviations \( \sigma_u^2 \). In another simulation, we retain the normal distributions for \( \hat{\theta}_u \), but generate the values of the variances from chi-square distributions with 10 degrees of freedom, scaled so that the expectations of the
standard deviations would coincide with $\sigma_u$. We repeat this simulation with 30 degrees of freedom. Each simulation uses 1000 replications.

The results for penalty ratio $R = 5$ are summarised in Figure 4. The simulated rates (probabilities) of being declared a winner are drawn by distinct symbols, as indicated in the caption. The probability associated with the normality assumption is represented by a black disc, the probabilities for the $t$ distributions are connected by vertical segments (with those for 30 degrees of freedom marked by gray discs) offset slightly to the left of the black disc, and the probabilities for the chi-square distributions are represented similarly to the right of the black disc, using gray squares. Only a few vertical segments are visible because the two rates they connect are too close to one another. The numerical values are given in Appendix B.

The diagram shows that the probabilities differ only slightly for the distributions with both 10 and 30 degrees of freedom. Even the largest observed changes are inconsequential; we therefore conclude that, in the studied context, selection of the winner
is not affected by small departures from the assumed model of normality of $\hat{\theta}_u$ with known variances.

4 League tables and classification

League tables have for a long time been regarded as an important element of informing the public about the performance of key services, such as education and health care. Forming league tables (ranking) is also a challenging statistical problem because the ranks are discontinuous functions of the underlying outcomes. Laird and Louis (1989) devised a solution with the property of minimum mean squared error. Some aspects of using estimated ranks and understanding the uncertainty about them are elaborated by Shen and Louis (1998). Leckie and Goldstein (2009) and the discussion of Goldstein and Spiegelhalter (1996) have concluded with a more balanced view of the strengths and weaknesses of such (estimated or exact) rankings of public institutions. We set aside the issue of the differing within-unit distributions of students’ or patients’ background (case-mix), which can in principle be addressed by adjustment for covariates. The sampling variation of the ranks, especially in the mid-range, is substantial, and so the detail, or informativeness, of such tables is largely illusory, especially when used for predicting the performance of the units in the (near) future. Their temporal (year-to-year) variation is an additional source of uncertainty.

4.1 Ranking as a set of selections of winners

Notwithstanding this criticism, we propose the following method for ranking units. If we identify any winners, they are assigned rank 1. If we find none, then we remove the unit with the smallest balance $B_u$ and assign it rank 2. Next, we identify winners in the remainder of the units, and assign them rank 2, with rank 1 possibly staying unassigned. If no winner is found, we remove the unit with the smallest balance $B_u$, assign it rank 3, and continue this process until only one unit is left. This unit is assigned the rank next to the lowest of those already assigned.

For the example in Figure 1, with $R = 2$, we obtain the league table 24, 1, 23, 2, 24, 7, 20, 21, 12, 5, 5, 20, 6, 21, 4, 22, 3, 13; the rank of the listed unit is given in the subscript. Rank 3 is not awarded because the lowest of the 22 values of $B_u$ for the units $u$ still in the contest is for $u = 22$, and $B_{22} > 0$. The promotion of unit 7 to
rank 5 may appear excessive, but it reflects our liberal attitude \((R = 2)\), which can be interpreted as granting the benefit of doubt that \(\theta_7\) may be far greater than \(\hat{\theta}_7\), given the substantial uncertainty about \(\theta_7\). For \(R = 5\), the unit is promoted even higher. The league table is \(23, 24, 22, 7, 20, 18, \ldots, 5, 10, 3, 4, 6, 1\). For \(R < 1\), unit 7 is ranked much lower; its large standard error now works to its disadvantage. The ranks can be shifted to cover the entire range from 1 to \(N\), with an appropriate arrangement for ties (multiple assignments of a rank). As an alternative, they can be linearly transformed to have the same mean and variance as the sequence 1, \ldots, N, equal to \((N + 1)/2\) and \(N(N + 1)/12\), respectively.

In each round, the top available rank is assigned to units with \(B_u < B_0 = 0\). This rule can be made more liberal by replacing zero with a positive value of \(B_0\), and stricter by replacing it with a negative value. For example, the criterion with \(B_0 = 0.25\) and \(R = 2\) happens to yield the same ranking as \(B_0 = 0\) and \(R = 5\). The classes of rules defined by setting \(B_0\) and \(R\) are in general different, but they tend to be similar. We prefer setting the value of \(R\) because it has a natural interpretation as the penalty ratio.

### 4.2 Ordinal classes

As an alternative to a league table, we propose classifying the units according to a small set of thresholds or cutpoints. Let these cutpoints be \(T_1 > T_2 > \ldots > T_{K-1}\), and set \(T_0 = +\infty\) and \(T_K = -\infty\). Then the probability that unit \(u\) belongs to class \(k\), \(\theta_u \in (T_k, T_{k-1})\), is

\[
t_{u,k} = \Phi\left(\frac{T_{k-1} - \hat{\mu}_u}{\sigma_u}\right) - \Phi\left(\frac{T_k - \hat{\mu}_u}{\sigma_u}\right),
\]

with the definition of \(\Phi\) supplemented by its limits at \(\pm\infty\). With \(K\) classes, there are \(K\) correct decisions (correct choices of class \(k = 1, \ldots, K\)) and \(K(K - 1)\) incorrect choices (true class \(k\) and selected class \(k' \neq k\)). The losses can be presented in a \(K \times K\) matrix; Table 1 gives an example. The diagonal has zeros (no loss for the correct decision), and the entries increase with distance from the diagonal. The matrix does not have to be symmetric, nor have any regularity features. For example, being liberal may be built into the assessment by assigning greater loss for confusing true class \(a\) with selected class \(b > a\) than vice versa, as in Table 1.
Table 1: An asymmetric loss matrix for five classes.

<table>
<thead>
<tr>
<th>Selected class</th>
<th>True class</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0 2 3 4 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10 0 1 2 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>15 5 0 1 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>20 10 5 0 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30 20 15 10 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the example, one lossile is incurred for every elementary error in selecting a class lower than the true class, with an additional lossile when the true class is 5 and when class 1 is selected incorrectly. Errors in the other direction, selecting a higher class than the true class earn a $R = 5$ times greater loss. For example, confusing class 5 with class 3 costs 3 lossiles, whereas confusing class 3 with class 5 costs 15 lossiles. The penalty ratio $R$ does not have to be constant, and various irregularities can be introduced in the loss matrix. For example, the ratio $R$ may be different for ‘small’ errors (confusion of neighbouring classes) and ‘large’ errors (confusion by several class boundaries).

The expected loss is a linear combination of the probabilities $t_{u,k}$. The vector of expected losses for the $K$ classes is $At_u$, where $A$ is the loss matrix and $t_u$ the vector of probabilities $(t_{u,1}, t_{u,2}, \ldots, t_{u,K})^\top$. Suppose we place unit $u$ into class $h$; then the expected loss is the element $h$ of vector $At_u$. In general, we select the class for which the expected loss is smallest, given the data.

4.3 Loss functions

Instead of using a loss matrix, we can declare the loss as a function of the magnitude and sign (direction) of the error. If unit $u$ is assigned to class $h$, then no error is committed if $T_h < \theta_u < T_{h-1}$. The error is $e_u = \theta_u - T_h < 0$ if $\theta_u < T_h$ and it is $e_u = \theta_u - T_{h-1} > 0$ if $\theta_u > T_{h-1}$. A piecewise constant loss function is defined as $L(e) = 1$ if $e < 0$ and $L(e) = R$ if $e > 0$. A piecewise linear loss function is defined
as $L(e) = -e$ if $e < 0$ and $L(e) = Re$ if $e > 0$. A piecewise quadratic loss function is
defined as $L(e) = e^2$ if $e < 0$ and $L(e) = Re^2$ if $e > 0$. In all three cases, $L(0) = 0$.

Denote $z_{h,u} = (T_h - \hat{\theta}_u)/\sigma_u$. Below, we use the convention that $\int_{-\infty}^{\infty} f(x)dx = f_{+\infty} f(x)dx = 0$ for any integrable function $f$. If we assign unit $u$ to class $h$, the expected piecewise quadratic loss is

$$Q_{u,h} = \frac{1}{\sigma_u} \int_{-\infty}^{T_h} (T_h - x)^2 \phi\left(\frac{x - \hat{\theta}_u}{\sigma_u}\right) dx + \frac{R}{\sigma_u} \int_{T_h-1}^{+\infty} (x - T_{h-1})^2 \phi\left(\frac{x - \hat{\theta}_u}{\sigma_u}\right) dx$$

$$= \sigma_u^2 \left\{ \int_{-\infty}^{z_{h,u}} (z_{h,u} - z)^2 \phi(z) dz + R \int_{z_{h-1,u}}^{+\infty} (z - z_{h-1,u})^2 \phi(z) dz \right\}$$

$$= \sigma_u^2 \left[ \Phi_2(z_{h,u}) + R \left\{ 1 + z_{h-1,u}^2 - \Phi_2(z_{h-1,u}) \right\} \right], \quad (4)$$

where $\Phi_2(z) = (1 + z^2)\Phi(z) + z\phi(z)$ is equal to twice the primitive function of $\Phi_1(z) = z\Phi(z) + \phi(z)$, which in turn is the primitive function of the distribution function $\Phi(z)$. The result in (4) is obtained by a linear transformation of the integrand, followed by two-fold integration by parts. For $h = 1$, we set the value of $1 + z_{h-1,u}^2 - \Phi_2(z_{h-1,u})$ to zero, its limit as $z \to +\infty$. For each unit $u$ we select the class $h$ for which $Q_{u,h}$ is the smallest. The factor $\sigma_u^2$ is not relevant for the comparisons of the expected losses $Q_{u,h}$, $h = 1, \ldots, K - 1$.

For piecewise linear loss, we obtain an expression similar to (4):

$$Q_{u,h} = \sigma_u \left[ \Phi_1(z_{h,u}) + R \left\{ -z_{h-1,u} + \Phi_1(z_{h-1,u}) \right\} \right].$$

The loss matrix can be interpreted as a coarse version of a piecewise continuous loss function. The two kinds of loss function can be combined, introducing losses for the magnitude and sign of the error as well as for every skipped cutpoint. In a limited set of applications, we have not found such combinations of loss functions useful.

Table 2 displays the expected losses for the classification to three groups, delimited by the thresholds 18 and 20, for the example used in Section 3, with piecewise quadratic loss and $R = 2$. The table shows that only units 23 and 24 are assigned to class 1, with unit 22 failing by a narrow margin (0.97 vs. 0.71). Units 10 and 12 and 14–19 are assigned to class 2 with very small expected losses. Even though unit 7 has an appreciable probability of being a winner, it is assigned to class 2. However, the expected losses for the three choices are in a relatively narrow range (1.60 – 3.33). Only four units, 1 – 4, are assigned to class 3.
Table 2: Expected losses in classification to three groups, delimited by 18 and 20 points, with piecewise quadratic loss and $R = 2$.  

<table>
<thead>
<tr>
<th>Unit</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
<th>Class 6</th>
<th>Class 7</th>
<th>Class 8</th>
<th>Class 9</th>
<th>Class 10</th>
<th>Class 11</th>
<th>Class 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.40</td>
<td>7.63</td>
<td>5.75</td>
<td>6.06</td>
<td>4.91</td>
<td>5.89</td>
<td>3.33</td>
<td>3.82</td>
<td>3.74</td>
<td>5.09</td>
<td>2.79</td>
<td>3.57</td>
</tr>
<tr>
<td>2</td>
<td>2.82</td>
<td>1.07</td>
<td>1.23</td>
<td>0.74</td>
<td>0.64</td>
<td>0.31</td>
<td>1.60</td>
<td>0.40</td>
<td>0.42</td>
<td>0.10</td>
<td>0.43</td>
<td>0.08</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
<td>0.36</td>
<td>0.80</td>
<td>0.66</td>
<td>1.02</td>
<td>0.84</td>
<td>2.79</td>
<td>1.59</td>
<td>1.64</td>
<td>1.45</td>
<td>2.48</td>
<td>2.25</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The values for the selected class (minimum expected loss) are printed in italics.

The risk associated with classification is summarized by the total of the expected losses for the assignments of the units. For the setting of Table 2, it is equal to 9.03 losses, 0.38 per unit on average. The contributions to the total range from 0.03 (unit 15) to 1.60 losses (unit 7), both for assigning to class 2. Another characteristic of the assignment is its stability, described by the range of values of $R$, and of other settings (parameters) if applicable, for which the assignments are not altered, or are altered only slightly. Note that we use the neutral version of the term ‘stability’, referring to an impartial assessment, and not to any claim that the decision made with the proposed method is more stable than with any competitor. The assessment is essential for informing how narrow should be the plausible range for $R$, or another elicited parameter, for the decision to be unequivocal (without ambiguity). Greater effort is required in elicitation to arrive at a narrower plausible range.

Thus, at around $R = 2.24$, the assignment with minimum expected loss switches for unit 4 from class 3 to 2. In the direction of strictness, the first change takes place at $R = 1.245$, when the assignment of unit 5 switches from class 2 to 3. In the direction of inclusiveness, the first switch after $R = 2.24$ takes place at $R = 2.76$, when the assignment of unit 22 changes from class 2 to 1. For $R = 2$, the selection has by far the highest expected loss for unit 7, even though the selection is quite stable. It would switch from class 2 to 1 at $R = 7.6$ and to 3 at $R = 0.9$.

The expected losses for the units discussed above are drawn as functions of $R$ in
Figure 5. The class associated with the smallest expected loss is indicated at the top margin of each panel, and the borderlines at which the selection switches from one class to another are marked by vertical dashes. The plotted functions are constant for selecting class 3, and are close to linearity for classes 1 and 2. The vertical axes of the panels have identical scales. For units 4 and 5, the expected losses are in narrow ranges throughout $R \in (0,9)$. However, class 3 would be selected only for extremely large values of $R$ because the expected loss for the choice of class 2 increases very slowly with $R$. For units 7 and 22, the expected losses for class 1 increase steeply; class 1 would be selected only for very small values of $R$.

5 Application

Concerns have arisen at a university that too many students fail exams or drop out for other reasons in the first semester of undergraduate study. The left-hand panel of Figure 6 displays the numbers of students enrolled in the university’s courses and the numbers who failed to proceed to the second semester in a recent academic year. The size of the plotting symbol indicates the estimated percentage of loss (failures and dropouts), and the shading reflects the sampling variation of this estimate, assuming an independent binomial distribution for each unit;

$$(X_u \mid n_u) \sim \text{Binom}(p_u, n_u),$$

where $X_u$ is the number of students who failed in or dropped out from course $u$; $n_u$ is the number of students who enrolled. In a Bayesian analysis, this model is supplemented by a noninformative prior for $p_u$. The conjugate beta distribution $B(a, b)$ is the natural choice. It corresponds to adding $a$ successes and $b$ failures to the original data. We use the uniform prior $B(1,1)$; the results are insensitive to the choice of $a$ and $b$ much smaller than the smallest counts of failures in the data.

The largest course had enrollment of 305 students, but 83 of them (27.2%) left by the end of the first semester. The two courses that are apparently most successful in retaining students are marked by circles with dots in their centers. The right-hand panel reproduces this information in a more familiar layout, as percentages of failures and their nominal 95% confidence bands derived by the normal approximation to the
Figure 5: Expected losses for selection of classes 1, 2 and 3 for units 4, 5, 7, and 22, as functions of the penalty ratio. The classes are indicated at the right-hand margin and the choice for which the minimum expected loss is attained is printed at the top margin of each panel.
Figure 6: Enrollments and losses by the end of the first semester in the 37 departments of a university. In the left-hand panel, the size of the points is related to the estimated probability of loss and the shading to the standard error of the percentage of losses. The vertical segments in the right-hand panel cover the nominal 95% confidence limits of the percentages. The circled dots are explained in the text.

Although the departments have entrance exams and interviews, most of them with nontrivial failure or rejection rates, some students enrol in courses only to seek transfer to another course, or another university, as soon as the semester starts. Some students fail to obtain a grant and are forced to abandon their studies, or postpone them for a year. Other students change their minds about studying and employment. The university management is concerned about the resources wasted by failing to retain students, or to screen them by the entrance exam, whereas the national office in charge of standards is concerned that the departments would respond to the problem by lowering their academic standards. Up to 10% of failures and dropouts in the first semester is regarded as acceptable, while a rate of 20% or higher is regarded as calamitous.

The department with the largest enrollment appears to have by far the highest failure rate or, in any case, accounts for a large fraction of the failures, 10.5%. In contrast, two courses with much smaller enrollment have lost only a handful of students.
each (2 out of 71 and 4 out of 143; encircled dots in Figure 6). We study the results for these departments and determine the range of perspectives in which they would be judged as exceptional. Also, we classify the departments according to three threshold rates, 10%, 15% and 20%, marked in the diagram by dashes. Throughout we rely on the normal approximation to the binomial distribution; it is problematic only for one of the departments, with data 2/71.

The fiducial probability that the largest department (83/305) has the highest rate of failures and dropouts is 0.752, so it is selected as a winner for $R > 1/0.752 - 1 = 0.33$. The department with the second highest probability, 0.084, has the results 38/171. It would be selected as a winner only for $R > 10.9$, an extremely liberal rule. Therefore we conclude that the department with results 83/305 has a failure rate higher than any other department. The level of exceptionality of this department can be assessed by the small expected loss associated with its selection, $1 - 0.752 = 0.248$; the corresponding expected losses for all other departures are far greater.

An alternative that may be easier to motivate finds the borderline number of students leaving for which the department would still be selected as the worst on its own. For the respective penalty ratios 0.5, 1 and 2, these counts are 81, 77 and 72. The closest contender, the department with figures 38/171, would be declared as (jointly) worst if its numbers of losses were 52, 48 and 45 with $R = 0.5, 1$ and 2, respectively, assuming the original figures, 83/305, of the ‘worst’ department.

The departments with results 2/71 and 4/143 have the highest probabilities of having the lowest rates of failure, 0.461 and 0.467, respectively. They are selected both as winners for $R > 1.17$ and the former is selected on its own only for $1.14 < R < 1.17$. Thus, both are selected as winners in almost any liberal perspective, and neither is selected in a strict perspective ($R < 1$).

We classify the departments to four groups according to the thresholds of 10%, 15% and 20%, using the penalty ratios $R = 0.5, 1$ and 2. With increasing $R$, the classification becomes more liberal and the assignments drift toward the ‘worst’ group (above 20%). Among the 37 departments, 27 have the same classification for the three values of $R$, with 5, 11, 7 and 4 departments in the respective classes $-10$, 10 – 15, 15 – 20, and 20+, referred to as classes 1 – 4. The classification of the other ten departments changes once each, to a neighboring class, from 1 to 2 between $R = 0.5$ and $R = 1$ in three instances (pattern 122), in two instances each from 1 to 2 between $R = 1$ and
$R = 2$ (pattern 112) and from 2 to 3 between $R = 0.5$ and $R = 1$ (pattern 233), and in one instance each the classification has pattern 223, 334 and 344. For $R = 0.5, 10, 14, 9$ and 4 departments are in the respective classes 1–4, and for $R = 2$, this division changes to 5, 16, 10 and 6. Figure 7 presents these classifications, with the patterns printed at an angle. The ‘constant’ patterns, such as 444, are abbreviated to the class (4) selected with all three values of $R$. The classification is equivocal for departments with estimated rates close to the thresholds.

6 Discussion

In the previous sections, we worked with mutually independent unbiased direct estimators of $\theta_u$, which use information only from the unit $u$. Empirical Bayes methods, as implemented in multilevel analysis, are commonly applied with data that comprise many aggregate units. Borrowing strength across the units, exploiting their similarity (Robbins, 1955; Efron and Morris, 1973; and Laird and Louis, 1989), is a widely acknowledged strength of the methods. Their key assumption is that the quantities $\theta_u$ are associated with a normal or another distribution. This assumption, particularly attractive when there are many units, has been subjected to scrutiny by Longford (2012). His argument centers on considering hypothetical replications of the study in which the same purpose is to be served. In the context of selecting the winner, inference is sought about a specific set of units. Therefore, prior to realising the data, $\theta_u$ are fixed. Appendix C derives the properties of the estimators of the fixed values $\theta_u$ when they are (inappropriately) assumed to be random.

Although normality is often a plausible assumption for an estimator, some estimators have distinctly non-normal distributions. The methods we propose can be extended to such distributions straightforwardly, because we do not rely on any properties that are specific to the normal distribution; the appropriate density and distribution function are simply substituted for $\phi$ and $\Phi$, respectively. In particular, when $\sigma_u$ is not known and is estimated, it gives rise to a scaled $t$ distribution, with the number of degrees of freedom assumed to be known. As illustrated in Figure 3, the expected loss depends on $\sigma_u$ less strongly than on $\hat{\theta}_u$, although the additive and multiplicative scales are difficult to compare in general. When a modicum of error in $\sigma_u$ has negligible consequences, a $t$ distribution even with a moderate number of degrees of freedom,
Figure 7: Classification of the departments of a university to four groups with penalty ratios $R = 0.5, 1$ and 2. Sets of three integers (classes) are printed at an angle for departments for which the classification is not the same for the three penalty ratios. Two departments happen to have identical counts 32/154.
such as 30, can be replaced by a normal distribution (see Figure 4); the associated uncertainty about $\sigma_u$ can be ignored. On the other hand, $\hat{\theta}_u$ based on a small sample size (of students or patients) would not be useful because of the substantial uncertainty about $\theta_u$.

The loss function is declared after elicitation (discussion or negotiation) with the client or the sponsor of the analysis. This process can be quite contentious, especially if the analyst insists on a single function, that is, on a single penalty ratio (or loss matrix). A constructive solution is to declare a range of plausible penalty ratios, or values of another parameter that has to be set. The problem is then solved for the values within this (plausible) range. Monotonicity properties of the solution reduce the task to having to solve the problem only for the two limits of the plausible range or the four vertices of the Cartesian product if two plausible ranges have to be set. Ambiguity arises when the solutions for the two limits (or four vertices) do not coincide. The risk of such a deadlock is reduced by declaring narrower plausible ranges.

The variety of loss functions can be greatly extended by linear combinations. If $L_1$ and $L_2$ are loss functions, then so is their combination $a_1L_1 + a_2L_2$ for any $a_1 > 0$ and $a_2 > 0$. Loss matrices can be combined similarly, so long as they have identical dimensions. The losses for the two kinds of error may have different structures. For example, understatement may be associated with a constant loss and overstatement with a linear loss. Despite the flexibility, simple loss functions are of greater importance in practice because they have to be accepted by all the parties that constitute the client.

To avoid capitalization on chance, the penalty ratio and the thresholds (for classification or ranking) should be selected prior to inspecting the data. Elicitation of their values may be informed by data simulated with the purpose of imitating the relevant features of the dataset to be analyzed: the number of units, the distribution of the standard errors and the distribution of the estimates. As an alternative, a dataset from the same context but a different occasion (e.g., past year) may be used for this purpose.

Ranks from consecutive years are bound to be compared. Even when measures of the uncertainty of such ranks are available, it is difficult to attach a level of confidence to a comparison selected after data inspection, especially when the units compared or some other context are selected post hoc. Bootstrap methods are a solution in principle, replicating the (improvised) selection process many times. In any case, we regard such
a comparison as misplaced, because the two values of $\hat{\theta}_u$ should be compared directly, or after an adjustment for the secular (year-to-year) trend. This can be accomplished by fitting a longitudinal model (Verbeke and Molenberghs, 2000).

We have not provided any general guidance for setting the loss or loss function(s). We, and other parties, have too little experience to codify the process of eliciting information from a client in a wide range of settings. Some invaluable advice is provided by Kadane and Wolfson (1998) and Garthwaite, Kadane and O’Hagan (2005), who review elicitation of prior distributions for Bayesian analysis. The principles they elucidate carry over to elicitation of the consequences of errors in decisions.

The main goal of this article is to sensitize readers to the relevance of the client’s perspective and to the value of integrating this perspective in any analysis. We hope that this will contribute to a change in the culture of statistical analysis, which at present is often detached from the client and does not provide the clients with a direct prescription for how to adjust their conduct in the light of the conclusions of the analysis. The established forms of analysis seek to be objective, catering for a distant and impersonal client. Our proposal implies a subjective analysis, tailored to a particular client’s perspective. This perspective has to be elicited and appropriately quantified. The additional burden this imposes on an analysis is rewarded by a solution — the choice of one of the available courses of action — that has immediate practical relevance.

7 Conclusion

Identifying the winner, ranking, classification, and similar tasks are often carried out by substituting estimates of the relevant quantities for the targets. We present here a formal framework for executing these tasks in which the uncertainty in estimation is incorporated in the inferences. The only additional input into the analysis is the loss function, or a plausible range of its parameter(s), usually the penalty ratio. The development acknowledges that these tasks are within the remit of decision theory and, indeed, the choice of the loss function can have a nontrivial impact on the outcome.
Acknowledgements

The dataset analyzed in Section 5 was obtained from a party (representative of a university) that wishes to remain anonymous. The text of the section, and the information disclosed in it, were approved by the university administration. Comments and suggestions of its representative are acknowledged. Detailed reviews by referees and a joint editor were instrumental in substantial improvement of the originally submitted manuscript.

References


### Appendices

#### A. Data for the example in Figure 1

The values of $\sigma_u, u = 1, 2, \ldots, 24$, used as an example in Section 3, were generated as $\sigma = 0.57 + \frac{1}{7}X$, where $X$ is a random draw from the chi-square distribution with 3.5 degrees of freedom. The corresponding sample sizes $n_u = \sigma^2 / \sigma_u^2$ (rounded) for $\sigma^2 = 100$ are printed at the top margin of Figure 1. The value of each $\hat{\theta}_u$ was generated as a random draw from the normal distribution with expectation $18.72 + |Z| \sigma_u$ and standard deviation $\sigma_u$, where $Z$ is a random draw from the standard normal distribution. The constants used in this algorithm were set by trial and error. The rounded sample sizes that correspond to the realised values of $\sigma_u, n_u = \sigma^2 / \sigma_u^2$, assuming $\sigma^2 = 100$, are printed at the top margin of Figure 1.
B. Results of the simulations in Section 3

This appendix gives in Table 3 the empirical rates of being declared a winner for the example presented in Section 3. The rates are plotted in Figure 4.

C. Empirical Bayes estimator with a fixed set of units

In this appendix, we show that the empirical Bayes estimator is less efficient than the trivial (unbiased) estimator for a small fraction of the units with extreme (the highest and the lowest) values of the outcome variable $\theta_u$, when the units are regarded as fixed.

Suppose $\theta_u, u = 1, \ldots, N$ are a random sample from a normal distribution, $\theta_u \sim N(\theta, \nu^2)$, and for each $\theta_u$ we have a conditionally unbiased estimator $\hat{\theta}_u$;

$$
\left( \hat{\theta}_u \mid \theta_u \right) \sim N(\theta_u, \sigma^2_u).
$$

These estimators are mutually independent. As an aside, we note that conditional independence of $\hat{\theta}_u, u = 1, \ldots, N$, is not a necessary assumption for the methods based on minimizing the expected loss; in contrast, it is an essential assumption for empirical Bayes methods. However, settings in which $\hat{\theta}_u$ are correlated are not common.

We assume that $\theta$ and $\nu^2$ are known and, to simplify the discussion, suppose $\nu^2 > 0$. This variance can be interpreted as the expected squared bias of $\theta$ when it is regarded as an estimator of $\theta_u$. In practice, $\theta$ and $\nu^2$ are estimated, $\theta$ usually with high precision. Dealing with the uncertainty about $\nu^2$, especially when there are only a moderate number of units $u$, is a problem that does not have a satisfactory solution. It can be partly relieved by using an appropriate prior distribution, but there is no effective remedy for estimating $\nu^2$ with only a few degrees of freedom, unless the prior is highly informative.

The empirical Bayes predictor of $\theta_u$ is

$$
\tilde{\theta}_u = \frac{\hat{\theta}_u \nu^2 + \theta \sigma^2_u}{\nu^2 + \sigma^2_u};
$$

it is conditionally unbiased and its conditional variance, given $\hat{\theta}_u$, is

$$
\text{var}\left( \hat{\theta}_u \right) = \frac{\nu^2 \sigma^2_u}{\nu^2 + \sigma^2_u}, \tag{5}
$$

which is smaller than both $\sigma^2_u$ and $\nu^2$. 

29
Table 3: Empirical rates (proportions) of being declared a winner, based on the standard errors plotted in Figure 1, but generated according to the process described in text.

<table>
<thead>
<tr>
<th>Unit</th>
<th>$\theta_u$</th>
<th>$\sigma_u$</th>
<th>$\mathcal{N}$</th>
<th>$t_{10}$</th>
<th>$t_{30}$</th>
<th>$\chi^2_{10}$</th>
<th>$\chi^2_{30}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18.49</td>
<td>1.02</td>
<td>0.036</td>
<td>0.041</td>
<td>0.046</td>
<td>0.046</td>
<td>0.049</td>
</tr>
<tr>
<td>2</td>
<td>18.56</td>
<td>1.15</td>
<td>0.081</td>
<td>0.071</td>
<td>0.076</td>
<td>0.088</td>
<td>0.077</td>
</tr>
<tr>
<td>3</td>
<td>18.61</td>
<td>0.84</td>
<td>0.014</td>
<td>0.015</td>
<td>0.023</td>
<td>0.023</td>
<td>0.010</td>
</tr>
<tr>
<td>4</td>
<td>18.62</td>
<td>0.63</td>
<td>0.000</td>
<td>0.006</td>
<td>0.001</td>
<td>0.007</td>
<td>0.004</td>
</tr>
<tr>
<td>5</td>
<td>18.62</td>
<td>0.60</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.003</td>
</tr>
<tr>
<td>6</td>
<td>18.66</td>
<td>0.73</td>
<td>0.006</td>
<td>0.009</td>
<td>0.008</td>
<td>0.010</td>
<td>0.008</td>
</tr>
<tr>
<td>7</td>
<td>18.66</td>
<td>0.79</td>
<td>0.017</td>
<td>0.012</td>
<td>0.014</td>
<td>0.014</td>
<td>0.015</td>
</tr>
<tr>
<td>8</td>
<td>18.73</td>
<td>1.49</td>
<td>0.210</td>
<td>0.203</td>
<td>0.223</td>
<td>0.180</td>
<td>0.221</td>
</tr>
<tr>
<td>9</td>
<td>18.76</td>
<td>0.91</td>
<td>0.028</td>
<td>0.023</td>
<td>0.029</td>
<td>0.032</td>
<td>0.024</td>
</tr>
<tr>
<td>10</td>
<td>18.80</td>
<td>0.91</td>
<td>0.044</td>
<td>0.045</td>
<td>0.045</td>
<td>0.033</td>
<td>0.038</td>
</tr>
<tr>
<td>11</td>
<td>18.83</td>
<td>0.87</td>
<td>0.031</td>
<td>0.041</td>
<td>0.045</td>
<td>0.027</td>
<td>0.031</td>
</tr>
<tr>
<td>12</td>
<td>18.84</td>
<td>1.02</td>
<td>0.086</td>
<td>0.081</td>
<td>0.059</td>
<td>0.077</td>
<td>0.076</td>
</tr>
<tr>
<td>13</td>
<td>18.85</td>
<td>0.62</td>
<td>0.002</td>
<td>0.005</td>
<td>0.006</td>
<td>0.006</td>
<td>0.007</td>
</tr>
<tr>
<td>14</td>
<td>18.85</td>
<td>0.86</td>
<td>0.033</td>
<td>0.035</td>
<td>0.037</td>
<td>0.037</td>
<td>0.029</td>
</tr>
<tr>
<td>15</td>
<td>18.85</td>
<td>0.62</td>
<td>0.004</td>
<td>0.006</td>
<td>0.006</td>
<td>0.009</td>
<td>0.004</td>
</tr>
<tr>
<td>16</td>
<td>18.86</td>
<td>0.63</td>
<td>0.022</td>
<td>0.034</td>
<td>0.036</td>
<td>0.029</td>
<td>0.032</td>
</tr>
<tr>
<td>17</td>
<td>18.86</td>
<td>0.63</td>
<td>0.004</td>
<td>0.010</td>
<td>0.009</td>
<td>0.007</td>
<td>0.006</td>
</tr>
<tr>
<td>18</td>
<td>18.90</td>
<td>0.88</td>
<td>0.037</td>
<td>0.048</td>
<td>0.050</td>
<td>0.039</td>
<td>0.044</td>
</tr>
<tr>
<td>19</td>
<td>18.92</td>
<td>1.18</td>
<td>0.137</td>
<td>0.116</td>
<td>0.133</td>
<td>0.139</td>
<td>0.130</td>
</tr>
<tr>
<td>20</td>
<td>18.93</td>
<td>0.85</td>
<td>0.026</td>
<td>0.036</td>
<td>0.037</td>
<td>0.037</td>
<td>0.043</td>
</tr>
<tr>
<td>21</td>
<td>18.95</td>
<td>0.81</td>
<td>0.040</td>
<td>0.032</td>
<td>0.029</td>
<td>0.034</td>
<td>0.031</td>
</tr>
<tr>
<td>22</td>
<td>18.98</td>
<td>1.05</td>
<td>0.107</td>
<td>0.093</td>
<td>0.092</td>
<td>0.090</td>
<td>0.101</td>
</tr>
<tr>
<td>23</td>
<td>19.36</td>
<td>1.40</td>
<td>0.336</td>
<td>0.329</td>
<td>0.319</td>
<td>0.310</td>
<td>0.321</td>
</tr>
<tr>
<td>24</td>
<td>19.45</td>
<td>1.55</td>
<td>0.428</td>
<td>0.399</td>
<td>0.406</td>
<td>0.413</td>
<td>0.394</td>
</tr>
</tbody>
</table>

Note: $\mathcal{N}$ — the originally assumed model, $\hat{\theta}_u \sim \mathcal{N}(\theta_u, \sigma_u^2)$, independently.
When we focus on a particular unit \( u \), we treat it and its value \( \theta_u \) as fixed, because we consider only replications in which this unit is represented, with constant sample size. Then \( \tilde{\theta}_u \) is biased,

\[
E(\tilde{\theta}_u | \theta_u) - \theta_u = -\frac{\sigma_u^2 (\theta_u - \theta)}{\nu^2 + \sigma_u^2},
\]

and its mean squared error is

\[
MSE(\tilde{\theta}_u | \theta_u) = \sigma_u^2 \frac{\nu^4 + \sigma_u^4 (\theta_u - \theta)^2}{(\nu^2 + \sigma_u^2)^2}.
\]

This expression coincides with (5) only when \( (\theta_u - \theta)^2 = \nu^2 \) or, equivalently, when \( \theta_u = \theta \pm \nu \), that is, when \( \theta_u \) is in the typical distance from \( \theta \). For \( \theta_u \in (\theta - \nu, \theta + \nu) \), the MSE in (6) is smaller than the variance in (5), whereas for \( \theta_u \notin (\theta - \nu, \theta + \nu) \), the variance in (5) is smaller. It is easy to show that \( MSE(\tilde{\theta}_u | \theta_u) \) is even greater than \( \sigma_u^2 = var(\hat{\theta}_u) \) when

\[
(\theta_u - \theta)^2 > 2\nu^2 + \sigma_u^2.
\]

So, the empirical Bayes predictor is less efficient than the direct estimator \( \hat{\theta}_u \) for the units with the largest and smallest values of \( \theta_u \). Depending on the values of \( \sigma_u^2 \) and assuming that the distribution underlying \( \theta_u \) is normal, this occurs for up to \( 200\{1 - \Phi(\sqrt{2})\} \approx 16\% \) of the units.