International Journal of Statistics and Probability; Vol. 5, No. 3; May 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Making Better Decisions: Can Minimizing Frequentist Risk Help?

Rose D. Baker[1] & Ian G. McHale[1]

[1] Centre for Sports Business, Salford Business School, University of Salford, UK.

Correspondence: Ian G. McHale, E-mail: i.mchale@salford.ac.uk

Received: September 12, 2015 Accepted: October 9, 2015 Online Published: April 14, 2016

doi:10.5539/ijsp.v5n3p80 URL: http://dx.doi.org/10.5539/ijsp.v5n3p80

Abstract

The concept of shrinking bet size in Kelly betting to minimize estimated frequentist risk has recently been mooted. This rescaling appears to conflict with Bayesian decision theory through the likelihood principle and the complete class theorem; the Bayesian solution should already be optimal. We show theoretically and through examples that when the model determining the likelihood function is correct, the prior distribution (if not dominated by data) is 'correct' in a frequentist sense, and the posterior distribution is proper, then no further rescaling is required. However, if the model or the prior distribution is incorrect, or the posterior distribution improper, frequentist risk minimization can be a useful technique. We discuss how it might best be exploited. Another example, from maintenance, is used to show the wider applicability of the methodology; these conclusions apply generally to decision-making.

Keywords: Kelly betting criterion; frequentist risk; shrinkage; tennis betting, power-law process.

1. Introduction: Bet Shrinkage

Baker and McHale (2013) introduced the concept of bet shrinkage in Kelly betting, for the case where the probability of winning is not accurately known. They derived approximations to the amount of shrinkage required, and showed that shrinking bet size increased expected utility both in a simulated gambling situation, and in tennis betting. Their methodology is frequentist, and can be applied in contexts other than Kelly betting. It will however doubtless raise Bayesian hackles, because Bayesian decision theory would seem to state that no such rescaling is required, and it is underpinned by two important theorems, the likelihood principle and the complete class theorem (see e.g. Robert (2007)). If Bayesian decision theory could in general be improved by a frequentist tweak, this would strike at the very foundations of decision theory. Can the circle be squared, so that bet shrinkage based on frequentist criteria is admitted as a valid procedure, whilst leaving Bayesian decision theory intact?

This paper gives an answer to this question, and clarifies the circumstances under which the type of adjustment recommended by Baker and McHale might be useful, and how it could best be used in decision making. We first recap briefly on the logic of bet shrinkage and explain how this relates to Bayesian decision theory, before describing Monte-Carlo simulations and an analysis of a large tennis dataset. We also look at equipment replacement, deliberately choosing a quite different decision problem from Baker and McHale (2013), to see from a practical viewpoint whether the methodology could be of general use. Finally, we draw some conclusions as to how rescaling the decision variable might be used in decision theory.

The example that was used to motivate bet shrinkage is that of betting on tosses of a die. For example, in the 17th century, the Chevalier de Méré famously made money by making equal-odds bets that he could throw an 'ace' (a six) in 4 tosses of a die, for which the probability is $1-(5/6)^4 \simeq 0.5177$ (e.g., David, 1987). He consistently lost money in a subsequent bet that he could throw a double six in 24 tosses of two dice (probability is $1-(35/36)^{24} \simeq 0.4914$), leading him to seek help from Blaise Pascal. Suppose we (wisely) decide to accept this second bet, but, like the Chevalier, are unable to compute the probability of winning. We might try a simulation, tossing two dice 24 times, repeating this procedure n times and so estimating the probability θ of winning as $\hat{\theta} = m/n$, where m is the number of times that a double six does not come up. This would give some information about θ, but it would not be the very accurate knowledge we could gain from simple probability theory, where the only inexactness in our knowledge of θ arises from tiny inaccuracies in the manufacture of the dice. We focus on this example, although Baker and McHale (2013) also consider the general problem of deciding how much to bet, if the bettor has only an inexact estimate (maybe a guess) of the probability of winning.

The decision we as investors face is how much of our bankroll to invest, given our knowledge (belief) about θ and the odds offered. For guidance on this, we could use the Kelly betting formula (Kelly, 1963). This is based on a logarithmic utility. Given odds of b (here $b = 1$), after betting a proportion δ of our bankroll, the utility from a unit bankroll becomes $\ln(1+b\delta)$ with probability θ and otherwise becomes $\ln(1-\delta)$. Maximising the expected utility gives $\delta = \{(b+1)\theta - 1\}/b$. Maclean, Thorp and Ziemba (2012) include a comprehensive collection of papers on the Kelly criterion, and Poundstone (2005) also describes its use in finance.

If we knew the value of θ, we could make an optimum bet of $\delta = 2\theta - 1 \simeq 0.0172$ of our bankroll. Because we have only an estimate of θ, we will bet more or less than the optimum if we use the Kelly formula with our 'plug in' estimate $\hat{\theta}$ or with the mean of the posterior distribution of θ derived from using say a Jeffreys prior.

Intuitively, we could imagine many copies of ourselves deciding how much to bet, and we could generate the results that they might obtain through dice tossing by bootstrapping (e.g. Efron and Tibshirani, 1993) the actual results, or, equivalently, using the sampling distribution of θ. The finding is then that the average utility of these bootstrapped decisions can be increased by shrinking the bet size, and we apply this shrinkage factor to the bet we are about to make. This is the motivation behind bet shrinkage as described by Baker and McHale (2013). This concept of considering multiple copies has some affinity with the concept of the 'wisdom of many in one mind' proposed by Vul and Pashler (2008), and Herzog and Hertwig (2009), in that it seeks to improve an individual decision by splitting it into multiple decisions.

That bet shrinkage is reasonable can be seen very easily; if we only carried out say two sets of tosses for the gambling example, we would obtain $\hat{\theta} = 1$ on roughly a quarter of occasions. Thus the Kelly formula with the estimated probability of winning plugged in would lead us to bet our entire bankroll, which we would lose with a probability of nearly a half. This illustrates the necessity of bet shrinkage.

2. Bayesian Decision Theory and Bet Shrinkage

We now examine what is implied in shrinking bet sizes from the viewpoint of Bayesian decision theory. This theory, like all Bayesian inference, is underpinned by the likelihood principle, which states that inference should depend only on the data actually observed, and on the complete class theorem, which states that Bayes estimators are admissible (no other estimator has smaller risk for all values of θ). Note however that there has always been some doubt over the likelihood principle, and Birnbaum's proof of it (see e.g. Robert, 2007) is currently disputed by Mayo (2010).

Lehmann (2008) gives an intuitive summary of the complete class theorem 'The most important link between the two approaches [Bayesian and frequentist] is the crucial result of Wald that, roughly speaking, any sensible statistical procedure is a Bayes procedure corresponding to some (proper or possibly improper) prior distribution.'

Following chapter 2 of Robert (2007), given a loss function $L(\theta, \delta)$ for carrying out action δ when the space of θ is Θ, parameter value(s) is/are θ, the prior distribution is $\pi(\theta)$ and the posterior distribution $\pi(\theta|x)$ after observing x, the expected posterior loss is

$$\rho(\pi, \delta|x) = \int_{\Theta} L(\theta, \delta)\pi(\theta|x)d\theta.$$

A decision rule $\delta(x)$ is optimal if it minimizes the posterior expected loss (maximises the posterior expected utility) and is then said to be a Bayes decision rule.

The frequentist risk is the integral over the sample space $\mathcal{X}$

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x))f(x|\theta)dx,$$

where f is the sampling distribution of x. This is what Baker and McHale (2013) sought to minimize once the decision rule δ had been determined, by rescaling $\delta \to \lambda\delta$. It was of course then necessary to plug in an estimate $\hat{\theta}$ of θ to make $R(\theta, \delta)$ computable.

The integrated risk $r(\pi, \delta)$ is

$$r(\pi, \delta) = \int_{\mathcal{X}} \rho(\pi, \delta(x)|x)m(x)dx, \tag{1}$$

where $m(x)$ is the marginal distribution of x. This risk in the gambling case is the average loss we would expect, on repeated gambles where the winning probability is drawn from the prior distribution, and so is precisely what regular bettors wish to minimize. From Bayes' theorem $\pi(\theta|x)m(x) = f(x|\theta)\pi(\theta)$ and from this, on using Fubini's theorem (see, for example, Thomas and Finney, 1996) to change the order of integration, the integrated risk can be rewritten as

$$r(\pi, \delta) = \int_{\Theta} R(\theta, \delta)\pi(\theta)d\theta. \tag{2}$$

From (1) any decision rule δ minimizing the expected posterior loss $\rho(\pi, \delta(x)|x)$ will also minimize the integrated risk, which from (2) can also be achieved by minimizing the frequentist risk. Yet if the integrated risk is precisely what we seek to minimize on repeated betting, and the Bayes rule already does this, how can bet shrinkage help?

Before answering this question, we quote a proof that shrinkage is required, essentially following Baker and McHale (2013), but recast in the standard notation of decision theory. We reduce x to $\hat{\theta}$, our (Bayesian or other) estimator of θ.

Theorem 1. *Given an unbiased estimator $\hat{\theta}$ of win probability, frequentist risk $R(\theta, \delta)$ can be decreased $\forall\ \theta$ by shrinking the bet size.*

Proof. To prove that risk can be decreased by shrinking the bet size, we show that $dR(\theta, \lambda\delta)/d\lambda|_{\lambda=1} > 0$ for non-degenerate $f(\hat{\theta}|\theta)$. The loss function (negative utility) is

$$L(\theta, \lambda\delta) = -\theta \ln(1 + b\lambda\delta(\hat{\theta})) - (1 - \theta)\ln(1 - \lambda\delta(\hat{\theta})).$$

The frequentist risk is then

$$R(\theta, \delta) = \int_{\theta} L(\theta, \delta(\hat{\theta}))f(\hat{\theta}|\theta)d\hat{\theta},$$

so that

$$dR(\theta, \lambda\delta)/d\lambda|_{\lambda=1} = -\int_0^1 f(\hat{\theta})|\theta)\delta(\hat{\theta})[\frac{\theta b}{1 + b\delta(\hat{\theta})} - \frac{1 - \theta}{1 - \delta(\hat{\theta})}]d\hat{\theta}.$$

which reduces to

$$dR(\theta, \lambda\delta)/d\lambda|_{\lambda=1} = (1/(b + 1))\mathrm{E}\{\theta/\hat{\theta} + (1 - \theta)b/(1 - \hat{\theta})\} - 1.$$

Since $1/\hat{\theta}$ and $1/(1 - \hat{\theta})$ are convex functions, Jensen's inequality, here that $\mathrm{E}(1/\hat{\theta}) > 1/\mathrm{E}(\hat{\theta})$ and $\mathrm{E}(1/(1 - \hat{\theta})) > 1/\mathrm{E}(1 - \hat{\theta})$ applies. Then the fact that $\mathrm{E}(\hat{\theta}) = \theta$ gives the required result. When there is no uncertainty in the probability of winning, the inequality becomes an identity. □

The curvature $d^2R(\theta, \lambda\delta)/d\lambda^2$ is positive and $dR(\theta, \lambda\delta)/d\lambda|_{\lambda=0} < 0$, so that there is a unique minimum loss for $0 < \lambda < 1$. The optimum amount of shrinkage is an estimate, but some shrinkage is required, and a decision rule that shrunk bet size δ by ϵ unless $f(\hat{\theta}|\theta)$ were degenerate would dominate the plug-in decision rule, i.e. give lower frequentist risk for all values of θ.

Bet shrinking amounts to shrinking the mean of the posterior distribution towards $1/(b+1)$, the bookie's implied probability of a win. For the gambling example, for a shrinkage factor λ with the Jeffreys prior we could take $\hat{\theta} = \lambda(m + 1/2)/(n + 1) + (1 - \lambda)/(b + 1)$, when $\delta = \lambda\{(b + 1)\hat{\theta} - 1\}/b$.

The resolution of the apparent contradiction between the need for bet shrinkage and the optimality of δ from minimising the expected posterior risk is that the above proof hinges on $\hat{\theta}$ being an unbiased estimator, and a Bayes estimator cannot be unbiased except in the trivial case where the risk is zero (Lehmann and Casella 1998, ch. 4). Hence the decision rule that the proof shows can be dominated is not a Bayes rule, and there is no contradiction.

We can see in more detail how the solution to the paradox works in the dice example. Consider the prior distribution $\pi(\theta)$; taking the Haldane (improper) conjugate prior distribution $\pi(\theta) \propto \theta^{-1}(1 - \theta)^{-1}$ leads to the sample estimated probability $\mathrm{E}(\theta|x) = m/n$, while the commonly-used Jeffreys prior $\pi(\theta) = (1/\pi)\theta^{-1/2}(1 - \theta)^{-1/2}$ leads to $\mathrm{E}(\theta) = (m + 1/2)/(n + 1/2)$. Thus the Haldane prior leads to an unbiased estimator of θ, where the decision rule can be dominated by bet shrinkage. This is because the Haldane prior leads to improper posterior distributions where if $m/n = \hat{\theta} = 1$ or zero, the probability distribution collapses. The risk $R(\theta, \delta)$ is then infinite, and (1) and (2) do not apply. We may wonder what would happen using a reference prior, but in one dimension, the reference prior is the Jeffrey's prior (Bernardo, 2005).

It is useful to consider what happens using other priors. If we imagine a computer simulation in which values of θ are generated from the prior, and win/lose data are then generated with win probability θ, indeed no bet shrinkage is required. This can be seen from (2), but has been verified by computation. One could say that the Bayesian estimator $\mathrm{E}(\theta)$ is pre-shrunk towards the value $\theta = 1/2$ by just the right amount to minimize the integrated risk. However, on carrying out the same experiment where θ is generated from a more peaked prior distribution but where the Jeffreys prior is assumed in deriving the optimum betting amount δ, bet shrinkage is required. The frequentist risk in (2) is being averaged over the wrong prior distribution for θ. The problem with the Bayesian approach is that we are using a vague or uninformative prior distribution, to reflect our ignorance of the value of θ, but in repeated betting, what we need in order to minimize frequentist risk is the prior distribution of the win probabilities for the sort of bet we habitually make. The prior ignorance assumption amounts to assuming a very broad distribution of win probabilities (large variance). If we really had such a broad distribution, with peaks at $\theta = 0$ and 1, we could make good bets on the basis of very limited information.

The shrunken decision will have lower risk most of the time, but occasionally will not, so for priors other than the Haldane prior, the shrunken decision rule will not dominate the Bayes rule, in accordance with the complete class theorem.

Similarly we may suspect that a 'wrong' model which then yields a 'wrong' likelihood function will also cause the Bayesian approach to perform badly. In the case of a wrong model, one can attempt to rectify the situation by developing a more comprehensive model, but where there is little data and the prior distribution is important, the rescaling is useful.

In general, minimizing frequentist risk starts with a Bayesian choice for δ, based on maximising expected utility. For Kelly betting, this also corresponds to using a plug-in estimate for θ, but this will not be true in general. Of course, one could also simply use a plug-in estimator, and seek to improve that through scaling. Although a frequentist transformation of some kind is applied to δ, rescaling is merely the simplest possibility. In Kelly betting, one could instead take $\delta \to \delta^\gamma$, giving shrinkage when $\gamma > 1$. This choice however would not shrink large bets where $\delta \simeq 1$. There is no way to find the optimal transformation of δ. In general, positive decision variables can be shrunk or expanded, and variables defined on the whole real line could have a constant added.

The minimization of frequentist risk occurs occasionally in statistical theory, e.g. Beran, (2005), where minimizing risk serves to select the best model from the class of candidate Bayes estimators. Entirely frequentist procedures such as Mallows' C_p (Mallows, 1973) also minimize risk under quadratic loss. There are two minimizations to be done: fitting a least-squares model and then minimizing C_p over the discrete range of model choices. This is also the case in bet shrinkage: one minimizes expected posterior loss to obtain $\delta(x)$ and then minimises frequentist risk to obtain λ.

Müller (2012) suggests modifying the posterior distribution in linear regression using an 'artificial posterior' with the sandwich covariance matrix. This reduces asymptotic frequentist risk. Thus the concept of minimizing frequentist risk has been used successfully in statistics, albeit sporadically. In following sections, we consider some examples to gain practical experience with rescaling.

3. Examples

3.1 Gambling

First, we study a gambling problem where results of between 8 and 20 trials of a gamble were simulated, and the bet was then made using the information gleaned from these trials about winning probability. The probability of winning was in fact a random variate from a beta prior distribution. The even-odds Kelly bet is computed using the Jeffreys prior, and is shrunk using the approximate formula given in Baker and McHale (2013). In our notation, the shrinkage factor is

$$\lambda = \frac{((b+1)\hat{\theta} - 1)^2}{((b+1)\hat{\theta} - 1)^2 + (b+1)^2\sigma^2} = \frac{\delta(\hat{\theta})^2}{\delta(\hat{\theta})^2 + ((b+1)/b)^2\sigma^2}, \tag{3}$$

where σ^2 is the estimated variance of the sampling distribution of $\hat{\theta}$. Table 1 shows the results. When win probabilities are simulated from the Jeffreys prior, so that the prior is 'correct', from (2) bet shrinkage can only increase the frequentist risk (decrease the expected utility). However, it does so by a tiny amount. Once the distribution of win probabilities diverges from the assumed prior, bet shrinkage starts to increase the expected utility. The shrinkage applied was not the optimal amount, but even the simple approximation in (3) is worthwhile. This bears out the point that shrinkage is useful if the prior distribution is 'wrong'. The evaluation of a decision rule for a bettor with specified risk aversion by whether it would have 'made money' on betting against the market or against a proxy bookmaker is precisely what is recommended by Johnstone *et al* (2013).

Table 1. Expected utilities from 10000 simulations, computing bet size using Kelly betting and the approximate shrunken Kelly bet from Baker and McHale (2013). The Jeffreys prior is used, and bet win probabilities are generated from the distributions shown.

Win prob. dist.	E(U) (Kelly)	E(U) (shrinkage)
Jeffreys ($\alpha = 1/2, \beta = 1/2$)	29.23	29.17
Uniform ($\alpha = 1, \beta = 1$)	17.43	17.47
Beta ($\alpha = 2, \beta = 2$)	8.86	9.06
Beta ($\alpha = 3, \beta = 3$)	5.63	5.93

3.2 Tennis Betting

Next, we revisit tennis betting, where Baker and McHale (2013) found that bet shrinkage increased the expected utility. This conclusion can be refined. We obtained more comprehensive data on the results of matches from the top tier of men's professional tennis, the ATP Tour, for 2002-2013 (31530 matches) from www.tennis-data.co.uk. The data included the participants' names and ATP world rankings points at the time of the match, the match result and up to six bookmaker's odds for each game, of which we used the one that was present most often. Following Boulier and Stekler (1999) and Clarke and Dyte (2000), we adopt a simple logistic regression model for estimating the probability of victory for the higher

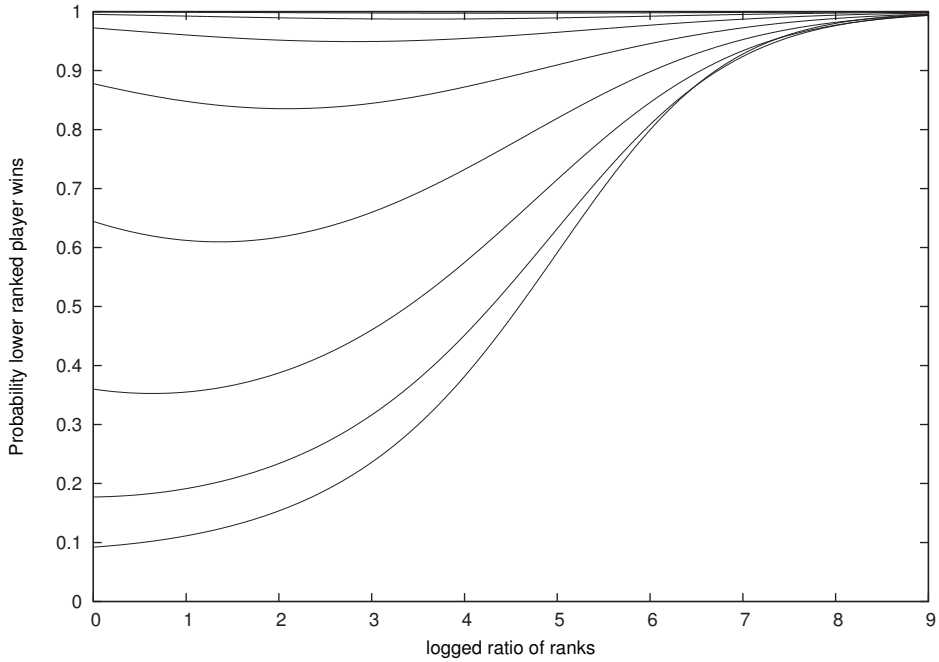

Figure 1. The logistic model used in tennis betting. The lines correspond to log odds of -2.5,...4.5, going from the lowest to highest line.

ranked (better) player using the natural logarithm of the ratio of the two players' world rankings points as a covariate. We also include the logarithm of the ratio of bookmaker's odds for the two players. Including these two variables, and their squares and the cross (interaction) term gives a model with 6 parameters that is good enough to enable betting to be worthwhile. Of course, much better models can be constructed using other covariates. We divided the data into 10 time slices, with the fit to all previous data being used to make bets for the current time slice. The details of the model used are not the point of this paper, but for completeness table 2 shows the fitted model parameters, when the model is fitted to the whole dataset. Figure 1 shows the model graphically.

Table 2. Tennis betting model parameters, values, standard errors, Z scores and p-values for a Wald test that the parameter is zero.

Predictor	Coef	SE Coef	Z	P
constant	-0.0165	0.02987	-0.5513	0.581396
logrank ratio	-0.1601	0.04761	-3.3630	0.000771
logrank ratio squared	0.0803	0.01677	4.7865	0.000002
logodds	1.1692	0.03695	31.6414	< 0.000001
logodds squared	0.1043	0.02678	3.8941	0.000099
log rank ratio times log odds	-0.1161	0.03165	-3.6684	0.000244

The focus here is not how good the model is, but whether bet shrinkage is required. We first explore the advantage of a Bayesian formulation. With the 6 covariates $x_1 \cdots x_6$ and parameter values $\beta_1 \cdots \beta_6$ the model for the win probability θ is

$$\theta = \frac{1}{1 + \exp(-\sum_{i=1}^{6} \beta_i x_i)} = 1/(1 + \exp(-\boldsymbol{\beta}^T \mathbf{x})).$$

Let the maximum-likelihood estimators be $\tilde{\boldsymbol{\beta}}$, where asymptotically $\tilde{\boldsymbol{\beta}} \sim N[\boldsymbol{\beta}, \mathbf{V}]$, where $\mathbf{V}$ is the negative inverse of the Hessian. Then using the delta-method (e.g. Oehlert, 1992), to $O(1/n)$ under a vague prior for $\boldsymbol{\beta}$ we have that the expectation of the posterior distribution is $\mathrm{E}(\theta) \simeq \tilde{\theta} - (\mathbf{x}^T \mathbf{V} \mathbf{x})\tilde{\theta}(1 - \tilde{\theta})(\tilde{\theta} - 1/2)$, showing that the Bayesian estimate of win probability is shrunk slightly towards the centre of its range. The effect on winnings from Kelly betting is of course small, but figure 2 shows that the Bayesian adjustment improves the winnings slightly.

If the model is adequate, given the large sample size the vague prior on $\boldsymbol{\beta}$ should pose no problem, and bet shrinkage should not be required. We examined the expected utility from the bets after shrinking all bets down by the same factor, and found

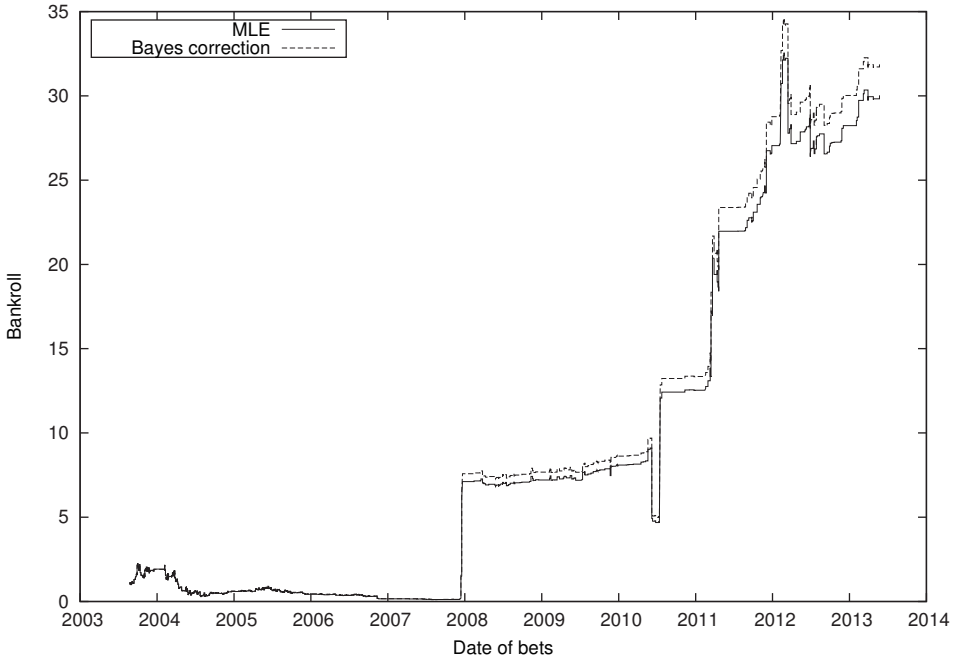

Figure 2. Cumulative bankroll from tennis betting, using the maximum-likelihood estimators of model parameters, and also calculating the win probability as the expectation over the posterior distribution.

that indeed any shrinkage only decreased expected utility. When the model was forced to be inadequate, the finding was different. We took the model described in table 2 and discarded the cross term. Now (surprisingly) performance was poor: the final bankroll was only 0.462 units of money, giving an expected utility of -0.772, so betting using the model actually decreases expected utility. Shrinking all bets down to 25% of their value however gives the best result, a final bankroll of 1.092 units with expected utility 0.088. Although we are winning or losing tiny amounts of money, bet shrinkage is needed with this inadequate model, and it converts a negative outcome to a positive one. As expected, bet shrinkage can only improve results when something is wrong in the problem formulation, i.e. a wrong prior distribution or as here an inadequate model.

3.3 The Power-Law Process

Our aim here was to apply the method of rescaling the decision variable to minimize the frequentist risk to another problem in decision theory, quite different from the Kelly betting application considered in the previous section. This allows us to explore the question of whether the method is of general applicability.

The nonhomogeneous Poisson process where the expected number of events to time t is $(t/\alpha)^\beta$ where $\alpha > 0, \beta > 0$ is widely used in deciding when to replace ageing machinery. For example, Kumar and Klefsjö (1992) applied it to the replacement of load-haul dump trucks. Given a cost c_r of repairing the system and a cost $c_a > c_r$ of replacing it, we wish to replace at age δ to minimize the cost per unit time $L = \{c_r(\delta/\alpha)^\beta + c_a\}/\delta$. For $\beta > 1$ we should replace at age

$$\delta = \alpha\{\frac{c_a}{c_r(\beta - 1)}\}^{1/\beta}, \tag{4}$$

and if $\beta \leq 1$ reliability does not deteriorate with time and the system should never be replaced. Since a wrong decision here predicts $L = 0$ but in fact results in infinite L , we impose the condition that the system be replaced at age $T \gg 1/\alpha$ in any case. This would likely happen in practice because of technological obsolescence. We assume $c_a/c_r = 100$ and without loss of generality set $c_a = 1$.

Of course, we do not know α and β and so will make a suboptimal decision if we estimate these parameters from data. We consider three methodologies: plugging the maximum-likelihood estimators (MLEs) into (4), the Bayesian method of replacing the expected number of failures by age δ by the expectation of this quantity over the posterior distribution, and minimization of the frequentist risk by rescaling the optimum value of δ found using the Bayesian method.

Bain and Englehardt (1991) give the MLEs $\hat{\alpha}, \hat{\beta}$ and their sampling distribution, which is needed for the rescaling method, and Bar-Lev *et al* (1992) describe Bayesian inference. Data can be time-truncated or failure truncated; we assumed failure

truncation, so that observation stops after the nth failure. Failure times are $t_1 \cdots t_n$. With the commonly used (improper) prior $\pi(\alpha,\beta) = 1/(\alpha\beta)$, from the posterior distribution one can obtain the expectation

$$\mathrm{E}(\delta/\alpha)^\beta = n\{\frac{\sum_{i=1}^{n-1} t_n/t_i}{\sum_{i=1}^{n-1} t_n/t_i + \ln(t_n/\delta)}\}^{n-1}. \tag{5}$$

This can be rewritten in terms of the MLEs $\hat{\alpha}$ and $\hat{\beta}$ as

$$\mathrm{E}(\delta/\alpha)^\beta = n\{1 + \frac{\hat{\beta}\ln\hat{\alpha}/\delta + \ln(n)}{n}\}^{1-n}, \tag{6}$$

from which, using the product-limit form of the exponential function, it can be seen that the Bayesian expectation converges to the MLE $(\delta/\hat{\alpha})^{\hat{\beta}}$ as $n \to \infty$. One cannot now solve analytically for the optimum value δ as in the MLE plug-in case (4), so a function minimizer must be used.

The approach of rescaling the decision variable, the replacement age δ, was studied using Monte-Carlo data that was representative of real data, e.g. the load-haul dump truck data of Kumar and Klefsjö and data on failures of marine engines from USS *Grampus* and *Halfbeak*, cited by Ascher and Feingold (1984). Without loss of generality, the scale factor α was set to $\alpha = 1$, and β was chosen to be from 0.8 to 2.0; the region around $\beta = 1$ is particularly interesting. Replacement was made no later than $T = 1000$.

It is possible to improve the Bayesian estimator, by adopting a more realistic prior distribution. For load-haul dump trucks, from the data $1 < \beta < 2$, whereas $\pi(\alpha,\beta)$ assigns the same probability for β to lie within the range $(10000, 20000)$ as for it to lie within the range $(1, 2)$. Clearly, in general β must be small, for who would tolerate equipment where the failure rate increased so dramatically with age? Also, if β is nearly zero, surely the manufacturers would run the equipment for a burn-in period, before unleashing it on the public.

It is straightforward to give the prior distribution a mean of unity (constant failure rate) and a realistic variance. For mathematical tractability, we used a gamma distribution for β with unit mean, so that $\pi(\alpha,\beta) = \frac{\gamma^\gamma}{\Gamma(\gamma)}\beta^{\gamma-1}\exp(-\gamma\beta)/\alpha$. It is unlikely that $\beta > 2$, and with $\gamma = 4$ the probability of this is only 0.042. The probability that $\beta < 0.75 = 0.352$, which is rather high, as from the datasets examined it is unlikely that $\beta < 0.75$, and for realism we would really need to use a different distributional form, such as a truncated normal distribution. However, with the gamma distribution prior for β, the result

$$\mathrm{E}(\delta/\alpha)^\beta = n\{\frac{\sum_{i=1}^{n-1} t_n/t_i + \gamma}{\sum_{i=1}^{n-1} t_n/t_i + \ln(t_n/\delta) + \gamma}\}^{n+\gamma-1}. \tag{7}$$

can be derived analytically; the use of the gamma prior for β and this result appear to be new. The algebra is not shown as it is routine. The effect of introducing γ is to lengthen the optimum replacement time, as huge values of β that would necessitate frequent replacements occur with lower probability.

We simulated 10000 realizations of possible data for each value of β for various sample sizes n, computed the MLEs $\hat{\alpha}$ and $\hat{\beta}$ and the MLE replacement age from (4), and the replacement age from the Bayesian estimator (5). This was done using a robust function minimizer. From the simulated realization, 8000 values of estimated α and β were computed using the formulae from Bain and Englehardt (1991), and the corresponding 8000 replacement ages derived from (6). The loss function (cost per unit time) was computed for a rescaling $\delta \to \lambda\delta$ using (5). A function minimizer was then used to find the scaling that minimized the loss function, and applying this rescaling factor to the Bayesian optimal δ gave our third estimate of replacement age. Finally, the true costs per unit time were evaluated for each of the 10000 realizations, using the values of α and β from which the realizations were simulated. We take the excess cost per unit time over that at the true optimum value of δ as the loss function, and this is reported in table 3. The results are correct to the accuracy shown, as was demonstrated by computing the variance of the loss functions, and also by scaling up the numbers of simulations until the computed table did not change.

It can be seen from table 3 firstly that the Bayesian solution is in general, but not always, somewhat superior to the MLE. Second, the rescaling to minimize frequentist risk performs fairly well, often giving a slightly lower cost than the Bayesian solution, but performing worse for higher values of β. Table 4 shows the performance of the methods with the $\gamma = 0$ prior. Here the Bayes method often performs worse than the MLE method, and the rescaling, based on the Bayes solution, also performs poorly.

From equations (1) and (2) it is clear that the Bayesian decision rule minimizes a weighted mean of the frequentist risks, so if rescaling reduces loss for some values of θ, it must increase it for others. The average rescaling factor $\lambda > 1$, so that replacement is postponed longer if minimizing frequentist risk.

Table 3. Shape parameter β, sample size n, Maximum-likelihood, Bayes and rescaled average loss functions, with the average rescaling factor λ. The Bayesian loss has $\gamma = 4$.

β	n	MLE	Bayes	rescaled	λ
0.8	5	0.018	0.012	0.011	1.84
0.9	5	0.026	0.016	0.015	1.38
1.0	5	0.036	0.020	0.020	1.14
1.1	5	0.045	0.022	0.022	1.09
1.5	5	0.119	0.045	0.052	1.03
2.0	5	0.381	0.151	0.164	1.06
0.8	10	0.003	0.003	0.003	47.26
0.9	10	0.006	0.006	0.004	3.09
1.0	10	0.010	0.008	0.007	1.80
1.1	10	0.014	0.010	0.009	1.40
1.5	10	0.070	0.042	0.048	1.08
2.0	10	0.194	0.140	0.160	1.06
0.8	25	0.001	0.001	0.001	8.28
0.9	25	0.002	0.002	0.001	14.18
1.0	25	0.003	0.003	0.002	3.93
1.1	25	0.005	0.005	0.004	1.71
1.5	25	0.043	0.037	0.041	1.07
2.0	25	0.114	0.115	0.118	1.03

As in the Kelly betting case, the loss function is asymmetric. In Kelly betting, betting too much can wipe out nearly the whole bankroll, and betting it all and losing would give an infinitely negative utility; refraining from betting merely gives zero utility. For equipment replacement similarly, replacing very often gives a huge disutility, but if β is not too large, not replacing often enough incurs only a small disutility. Hence in both cases rescaling δ makes sense, in the Kelly case scaling down the bet, and here scaling up the replacement age. As β increases, the asymmetry becomes smaller, and rescaling starts to perform badly. Estimating a further parameter δ from the data of course introduces noise and so can increase loss rather than decreasing it if the rescaling required is not large.

4. Conclusions

We have studied the technique of rescaling decision variables to minimize the frequentist risk, using the results of Bayesian decision theory, and also through Monte-Carlo simulations and analyses of real data. Besides examining Kelly betting, the quite different area of equipment replacement was studied, to address the question of the general applicability of the technique.

In drawing some conclusions for practical decision-making, it is first necessary to restate a remark about prior distributions. If, using our examples, we are betting repeatedly or replacing equipment on a regular basis, the frequentist properties of our (Bayesian) decision are important. The complete class theorem guarantees that they will be, but then the prior distribution must be right in a frequentist sense. If the prior is $\pi(\theta)$, the value θ should crop up a proportion $\pi(\theta)$ of times in the bets we make. This is of course likely not to be the case when we have no prior knowledge, and a 'vague prior' is used. For example, the Jeffreys prior $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ implies that the probability of winning the bet is most likely to be either one or zero; with a little data, we can surely guess which type of bet we are facing, and so make a nice profit. This is insanely optimistic, which is why the Kelly betting strategy performs poorly when we do not know θ. This shows that unless we have a large sample size, there are risks in using a vague prior. We also require the model to be correct, but we assume that incorrect models can be refined until adequate, so this is not a major problem.

The Kelly betting example is an unusual case, where because the estimate $\hat{\theta}$ is assumed unbiased, there is no Bayes estimator, and the generalized Bayes estimator that is used instead and which minimizes posterior expected loss for every x is not admissible (Robert, page 64). In general, minimizing frequentist risk by rescaling the decision variable is effective only when the prior distribution is 'wrong' and strongly affects inference, or when the model is wrong but cannot be easily rectified.

With this understanding, how can the technique of rescaling decision variables to minimize frequentist risk best be exploited in practice? There are several options:

1. routinely use frequentist risk minimization, based on rescaling a naive Bayesian decision;

Table 4. Shape parameter β, sample size n, Maximum-likelihood, Bayes and rescaled average loss functions, with the average rescaling factor λ. The Bayesian loss uses the unimproved prior with $\gamma = 0$.

β	n	MLE	Bayes	rescaled	λ
0.8	5	0.018	0.029	0.018	2.60
0.9	5	0.026	0.041	0.026	2.40
1.0	5	0.036	0.054	0.033	2.27
1.1	5	0.045	0.064	0.039	2.20
1.5	5	0.119	0.118	0.080	1.90
2.0	5	0.381	0.202	0.186	1.89
0.8	10	0.003	0.005	0.005	16.29
0.9	10	0.006	0.009	0.011	2.74
1.0	10	0.010	0.014	0.018	1.63
1.1	10	0.014	0.018	0.026	1.18
1.5	10	0.070	0.058	0.075	0.70
2.0	10	0.194	0.140	0.169	1.05
0.8	25	0.001	0.001	0.001	12.00
0.9	25	0.002	0.002	0.002	14.64
1.0	25	0.003	0.004	0.003	3.68
1.1	25	0.005	0.006	0.005	1.68
1.5	25	0.043	0.038	0.042	1.03
2.0	25	0.114	0.112	0.112	1.01

2. use in one-off situations, but in general seek to improve the prior distribution $\pi(\theta)$ with experience, and then take a purely Bayesian approach;
3. use purely as a diagnostic tool to check decisions;
4. never use.

Our conclusion is that 2 and 3 are appropriate. For a one-off bet or decision, we can probably make a better decision by rescaling the decision variable, unless we have a prior distribution with good frequentist properties. But a habitual bettor could adopt a prior distribution for win probability θ such as a beta distribution with mean $(1/(b+1))^\gamma$ and unknown variance. When $\gamma = 1$ this shrinks the bet size to zero, and when $\gamma < 1$ the expected probability of winning exceeds the probability implied by the bookmaker and a bet is made. An empirical Bayes analysis of data used in making past bets, such as the dice bets described in table 1, would maximise the marginal probability $m(x)$ to estimate the prior distribution to be used in future betting. A fully Bayesian analysis would use a hierarchical prior distribution.

Turning to equipment replacement, based on properties of the type of machinery for which replacement policies must be devised, the same method could be used to find a prior distribution $\pi(\alpha,\beta)$ that was more representative of the values of the shape parameter β typically arising. To make better decisions, it is vital to improve the prior distribution, through experience or through attempts to elicit the experience of experts, e.g. Percy (2002). Table 3 shows that replacement age should be increased over the range of β likely to be encountered. A realistic prior distribution with small probability that (say) $\beta < 1/2$ or $\beta > 2$ would improve the Bayesian results further.

Finally, model criticism is an area where Bayesian statistics routinely seeks help from the frequentist techniques used in model-checking. This method of rescaling can be used as a tool to check how good our decisions are. It would be a little daunting to find that one would have made more money if bets had been routinely shrunk to 25%, or if equipment had been replaced at double the replacement age used. Hence the mean value λ of the optimal scaling factor from a series of decisions is a statistic that is a measure of model/prior adequacy for the task at hand.

In conclusion, minimizing frequentist risk has been sporadically used in statistics, for example in Mallows' C_p, and we think that the related concept of rescaling the decision variable to minimize risk can sometimes be useful in decision making, either directly, or as a diagnostic tool to help improve Bayesian methods. It should however be used with caution, and the more reliable approach of making better decisions by improving Bayesian prior distributions with experience should be the norm.

Acknowledgements

We would like to thank an anonymous referee for helpful comments.

References

Ascher, H., & Feingold, H. (1984). *Repairable Systems Reliability; modeling, inference, misconceptions and their causes*, Marcel Dekker, New York.

Bain, L. J., & Englehardt, M. (1991). *Statistical analysis of reliability and life-testing models*, 2nd. Ed., Marcel Dekker, New York.

Baker, R. D., & McHale I. G. (2013). Optimal betting under parameter uncertainty: improving the Kelly criterion. *Decision Analysis, 10*(3), 189-199. http://dx.doi.org/10.1287/deca.2013.0271

Bar-Lev, S.K., Lavi, I., & Reiser, B. (1992). Bayesian Inference for the Power-Law process. *Ann. Inst. Statist. Math., 44*, 623-639. http://dx.doi.org/10.1007/BF00053394

Beran, R. (2005). ASP fits to multiway layouts. *Annals of the Institute of Statistical Mathematics, 57*(2), 201–220. http://dx.doi.org/10.1007/BF02507022

Bernardo, J. (2005). Reference analysis. Handbook of Statistics, Dey D. K., Rao C. R. eds, 25, 17C60, Elsevier, Amsterdam. http://dx.doi.org/10.1016/S0169-7161(05)25002-2

Boulier, B. L., & Stekler, H. O. (1999) Are sports seedings good predictors? An evaluation. *International Journal of Forecasting, 15*(1), 83-91. http://dx.doi.org/10.1016/S0169-2070(98)00067-3

Clarke, S. R., & Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research, 7*(6), 585-594. http://dx.doi.org/10.1111/j.1475-3995.2000.tb00218.x

David, F. N. (1987). Games, gods and gambling. Griffin, London.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York. http://dx.doi.org/10.1007/978-1-4899-4541-9

Kelly J. L. Jr. (1956). A new interpretation of information rate. *Information Theory, IRE Transactions on, 2*(3), 185-189. http://dx.doi.org/10.1109/TIT.1956.1056803

Herzog, S. M., & Hertwig, R. (2009). The Wisdom of Many in One Mind: Improving Individual Judgments With Dialectical Bootstrapping. *Psychological Science, 20*(2), 231-237. http://dx.doi.org/10.1111/j.1467-9280.2009.02271.x

Johnstone, D. J., Jopnes, S., Jose, V. R. R., & Peat, M. (2013). Measures of the economic value of probabilities of bankruptcy. *Journal of the Royal Statistical Society series A, 176*, 635-653. http://dx.doi.org/10.1111/j.1467-985X.2012.01085.x

Kumar, U., & Klefsjö, B. (1992). Reliability analysis of hydraulic systems of LHD machines using the power law process model. *Reliability and system safety, 35*, 217-224. http://dx.doi.org/10.1016/0951-8320(92)90080-5

Lehmann, E. L. (2008). Reminiscences of a Statistician: the company I kept, Springer, New York. http://dx.doi.org/10.1007/978-0-387-71597-1

Lehmann, E. L., & Casella, G. (1998). Theory of Point Estimation (2nd ed.), Springer Verlag, New York.

MacLean L. C., Thorp E. O., & Ziemba W. T. (2012). Eds. The Kelly Capital Growth Criterion: Theory and Practice. World Scientific, Singapore.

Mallows, C. L. (1973). Some Comments on C_p, *Technometrics, 15*(4), 661-675.

Mayo, B. (2010). An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle *in* Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science (D Mayo and A. Spanos eds.), Cambridge: Cambridge University Press: 305-314.

Müller, U. (2012). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica, 81*(5), 1805-1849.

Oehlert, G. W. (1992) , A Note on the Delta Method. *The American Statistician, 46*(1), 27-29.

Percy, D. F. (2002). Bayesian enhanced strategic decision making for reliability. *European Journal of Operational Research, 139*(1), 133-145. http://dx.doi.org/10.1016/S0377-2217(01)00177-1

Poundstone W. (2005). Fortune's Formula: The Untold Story of the Scientific Betting System that Beat the Casinos and Wall Street. Hill and Wang, New York.

Robert, C. P. (2007). The Bayesian Choice: from decision-theoretic foundations to computational implementation, Springer, new York.

Thomas, G. B., Jr., & Finney, R. L. (1996). *Calculus and Analytic Geometry,* 8th Ed. Reading, MA: Addison-Wesley, p. 919.

Vul. E., & Pashler, H. (2008). Measuring the crowd within: probabilistic representations within individuals. *Psychological Science, 19*(7), 645–647. http://dx.doi.org/10.1111/j.1467-9280.2008.02136.x