Variation of the rates of necrotising enterocolitis in the neonatal networks in England

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

Abstract

Necrotising enterocolitis is an oft-fatal disease of the intestinal tract in neonates born prematurely and with low birthweight. We study the variation of its rates across the neonatal networks (groups of hospital-based neonatal care units) in England. We illustrate the problematic nature of hypothesis testing for a key decision which an analysis is meant to inform, and apply an approach based on decision theory. We emphasise the role of sensitivity analysis in dealing with the ambiguity encountered in the process of eliciting information about the perspective of the client or sponsor for whom the analysis is conducted. In the application based on a fiducial distribution, the likelihood is approximated by a linearising transformation of the score function.

Keywords: *beta distribution; binary outcomes; decision theory; necrotising enterocolitis; neonatal care; verdict.*

N. T. Longford, Department of Medicine, Imperial College, London, UK.
Email: `sntlnick@sntl.co.uk`.

1 Introduction

Infants born preterm and with low birthweight are cared for in neonatal units of about 170 hospitals in England. The vital organs of these infants tend to be poorly developed and are prone to life-threatening failure. Feeding them presents a dilemma: the process of maturing depends on nutrition, but the nutrition provided enterally presents a stern test for the gastrointestinal system. Neither is parenteral feeding free of any risks. Mother's breast milk is universally regarded as the best form of nutrition, [1], but, in the first few postnatal days, some infants do not have the sucking instinct and some mothers do not express milk, especially after a delivery by caesarean section. Care for extremely preterm born (at gestational age of 28 weeks or earlier) entails a delicate balance of attending to a multitude of vulnerabilities.

Necrotising enterocolitis (NEC) is a disease of the gastrointestinal system that afflicts preterm born infants in the first few weeks of life, [2]. Its aetiology and antecendents are poorly understood. Its symptoms present themselves often suddenly, rendering effective treatment, including surgery, difficult or problematic, more so when the neonate has to be transferred for treatment to another unit (hospital).

Units differ a great deal in how they feed infants and how they adjust the feeding regimen in response to various threats to survival and development. On the one hand, there is unlikely to be a single optimal feeding protocol for infants in neonatal care; on the other hand, a vast variety of feeding practices is unlikely to be equally effective for preventing NEC. We note that the risk of NEC is not eliminated by exclusive breastfeeding, although the risk is smaller than with other types of feed or their combinations with maternal breast milk, [3]. There may be conflicts in the effort to prevent several diseases and disabilities that threaten the preterm-born neonate, and so balancing the various risks, with consequences taken into account, turns neonatal care into a medical 'art form'.

The UK National Neonatal Research Database (NNRD) is a register of daily activities in the neonatal units in Great Britain. It collects information about births (mother's background and prenatal treatments, details of the birth, daily feeding of the neonate, suspected or diagnosed illnesses, treatments, including medication and surgery, and discharge from the unit). We study an extract from NNRD related to 14 678 infants born at gestational ages under 32 weeks, who received care in neonatal

units in England in years 2013 and 2014.

Absence of any variation of the hospital- or network-level rates of NEC would support the hypothesis that the feeding practices or protocols used in the hospitals are equally effective (or, conceivably, ineffective) in preventing NEC, [3]. Several hospitals have no cases over the two years; some of them have cared for only a few dozen infants in our sample. It is more practical to study the incidence of NEC in neonatal networks. This is also more relevant because the hospitals within a network tend to provide similar care, and a lot of infants who are in care for a long time (several weeks) are transferred within the network from one hospital to another for treatments, such as surgery. There are 23 networks in England; most of them have 6 – 9 neonatal care units.

TABLE 1

Table 1 displays the counts of infants (n_k), cases of NEC among them (m_k) and the associated rates (percentages) $100\hat{p}_k$, where $\hat{p}_k = m_k/n_k$, for the networks $k = 1, \ldots, K = 23$. The national rate of NEC is 3.15%. It has to be quoted with some qualification because it is strongly affected by the selection (definition) of the studied population, and the range of prematurity in particular. The observed rates $\hat{p}_k$ are quite small (2.0 – 4.6%) and, even with the sample sizes $340 < n_k < 1020$, it is difficult to judge whether the underlying probabilities p_k are identical, and if not, then to what extent they vary. We address this issue in Sections 2 and 3 by some established methods on which we highlight their deficiencies. In particular, we dismiss any solution based on a null-hypothesis test for the network-level variance of the rates, because such a test has no means of incorporating the consequences of the two kinds of error that may be committed. Another drawback is that there are no circumstances in which such a test would conclude with a support for the null-hypothesis. Failure to reject a hypothesis is commonly treated as a licence for a further analysis or some other conduct that presumes that the null is valid. This we regard as inappropriate. For example, a key issue in meta-analysis is whether the study-level treatment effects are identical or not, [4]. The choice between these two options is often based on a test of the relevant hypothesis; see [5], Chapters 3 – 5. Failure to reject it is inappropriately regarded as a justification to proceed with an analysis in which the treatment effects are assumed to be identical.

In Section 4 we describe an approach based on decision theory, which takes into account the client's assessment of the consequences of the two kinds of error. As its input it requires specifying the threshold between negligible (small) and substantial (large) variance of the networks and the loss ratio, a characteristic of these consequences. In Section 5 we apply this method to the data extracted from NNRD to choose between two courses of action that are appropriate when the variance is negligible (not to search for causes of variation) and when it is substantial (to search for causes). The concluding section summarises our proposal and outlines its wider potential.

As an aside, we mention that there is no universal definition of NEC, nor a protocol for its diagnosis. This introduces further ambiguities relevant to (international) comparisons of studies and their meta-analyses. We use a definition of NEC based on laparotomy (abdominal surgery) and death certificates, which has a low rate of false positives, but the rate of false negatives is inflated by missed cases that were cured or confused with other diseases of the gastrointestinal tract.

Testing the hypothesis that the cluster-level variance in a two-level random-effects model vanishes is a challenging problem because zero variance is at the boundary of the parameter space and the established results related to the chi-squared distribution of the likelihood ratio statistic do not apply. It was shown by [6] and [7] that the null-distribution in this problem is a linear combination of chi-squares, with the weights equal to certain eigenvalues that have to be estimated. Less progress has been made with other two-level models, such as the beta-binomial, which we use in Section 4, and with the distribution of the likelihood ratio statistic under the alternative of a positive variance. In Section 4.1, we approximate the distribution of a related statistic using linearising transformations of the score function.

2 Variation of the rates of NEC

We assume that the observed rates $\hat{p}_k$ are independent and unbiased estimators of the corresponding underlying rates (probabilities) p_k, with respective variances $v_k = p_k(1-p_k)/n_k$. Some births are multiple, with more than one sibling in care, and so the assumption of independence within networks is not valid. We believe that the associated error is negligible. Although a nurse usually cares for several infants, and

the care of every infant is shared by several nurses, the assumption that the outcomes within a network are mutually independent is quite credible.

Let $p = (p_1 + \cdots + p_K)/K$ be the average of the network-level probabilities. Note that it differs from the national rate $\bar{p} = \sum_k n_k p_k / n$, where $n = n_1 + \cdots + n_K$. Another rate to consider is $p^* = \sum_k \pi_k p_k$, where the weights π_k form the (multinomial) distribution of infants across the networks in a hypothetical superpopulation. To avoid some complexities, we assume that the counts n_k are fixed, so we do not distinguish between $\bar{p}$ and p^*. Apart from p and $\bar{p}$, a focal quantity is the variance of the rates,

$$\tau^2 \;=\; \frac{1}{K} \sum_{k=1}^{K} (p_k - p)^2 \;,$$

or its superpopulation version, which would be estimated without bias by $K\tau^2/(K-1)$ if the probabilities $p_1, \ldots, p_K$ were available.

With a mindset focused on hypothesis testing, we would test the null-hypothesis $\mathrm{H}_0 \colon\ \tau^2 = 0$. The rationale for it is that if it were established that $\tau^2 = 0$, we would not look for any causes of NEC in the feeding practices of the networks. The rationale would be weak for small values of this variance because the scope for improvement by identifying the cause(s) would be very limited.

For illustration, we apply the permutation test for H_0. In this test, the $n = 14\,678$ infants, with their NEC states (case or negative) intact, are reassigned completely at random to synthetic versions of the networks subject to the (fixed) caseloads n_k. This reassignment is replicated many ($H = 1000$) times. For the realised dataset and each replication, we evaluate the statistic

$$S \;=\; \frac{1}{K-1} \sum_{k=1}^{K} (\hat{p}_k - \hat{p})^2 \;.$$

Let the value of this statistic be S_0 for the realised dataset and S_h, $h = 1, \ldots, H$, for the replicate (synthetic) datasets. These replicates represent the sampling distribution of S under H_0. Therefore, we reject H_0 if S_0 is exceptional among $S_1, \ldots, S_H$. Figure 1 displays the histogram of 1000 replicate values S_h, with the realised value S_0 marked by vertical dashes. This value is at the percentile 94.4 of the replicate values, so we do not reject the null-hypothesis at 5% level, albeit by a narrow margin.

This result is not helpful for choosing one of the two contemplated courses of action, to explore the causes of NEC and not to do so. The test is insensitive to the perspectives

and value judgements of the client. They are influenced by the cost and effort required for the exploration as well as the disruption caused by it. Also, the exploration would not be justified if the variance τ^2 were known to be positive but negligible (small).

FIGURE 1

The first step in addressing this problem is to specify the threshold between negligible and substantial. The importance of this is transparent from the two extreme cases. If we set the threshold very close to zero we stack the choice in favour of 'substantial'. In contrast, if we set it very large we rule out this choice. In an elicitation exercise, the threshold could be set by presenting to the experts sets of rates $p_1, \ldots, p_K$ in suitable diagrams, and asking about each set whether they regard the dispersion of its K values as small enough or too large. The value of the threshold is influenced by the contemplated courses of action (the available options) as well as value judgements and perspectives; in brief, it is *subjective*. There may be some contention about its exact value, but the threshold is indisputably positive; improvements in the treatment of neonates would be negligible by uncovering causes of small network-level variation.

Suppose we settle on the threshold of $\tau_0 = 0.4\%$ for the standard deviation of the rates. Then a more relevant hypothesis might be $\tau \leq 0.4\%$ against the alternative $\tau > 0.4\%$. However, a test of this hypothesis could not arbitrate between the two options. Whilst rejection of the hypothesis would give support for $\tau > 0.4\%$, failure to reject it would not constitute evidence for $\tau \leq 0.4\%$. Even if we used failure to reject the hypothesis as a licence to assume that $\tau \leq 0.4\%$, insensitivity to the client's perspective and value judgements would remain a drawback of the test as a means of choosing one of the two options.

In the next section we shed light on this problem by finding the range of standard deviations, or variances, that are compatible (not in a statistical contradiction) with the observed rates. We simulate sets of sample rates according to a model with its parameters set to a given value of τ and relate these rates to the realised rates by means of the statistic S. We repeat this exercise for values of τ on a fine grid.

3 Beta-binomial model for sets of binomial rates

The class of beta distributions, given by the density

$$g(x;\alpha,\beta) \;=\; \frac{1}{B(\alpha,\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

for $0 < x < 1$, is commonly adopted as a model for probabilities. It is the conjugate class for the binomial in the Bayesian setting. That is, if the prior for a probability r is beta distributed and the outcome, the number of events (or cases) in a set of trials, has a binomial distribution, then the posterior distribution of r is also beta. We consider beta distributions that have the same expectation as the network-level mean rate p, $\alpha/(\alpha+\beta) = p$, so that $\beta = \alpha(1-p)/p$. We prefer the parametrisation in terms of the variance

$$\begin{aligned} \tau^2 &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{ap^2(1-p)}{1+ap}\,, \end{aligned}$$

where $a = 1/\alpha$. With this parametrisation, $\tau^2(a)$ is an increasing function of a with limits of zero and $p(1-p)$ as $a \to 0$ and $a \to +\infty$, respectively.

For a given value of τ^2 which, together with p, determines the values of α and β, we generate a set of K synthetic probabilities p_k as a random sample from the beta distribution with $\alpha = 1/a$ and $\beta = (1-p)/(ap)$, and then generate synthetic binomial counts m_k and evaluate the statistic S using the sample rates $\hat{p}_k = m_k/n_k$, $k = 1, \ldots, K$. By replicating this process (H times), we obtain an empirical distribution of S under the assumption that the variance of the underlying rates p_k is τ^2.

We apply this process for the borderline variance τ_0^2 and relate the realised value S_0 to the distribution of the synthetic values of S. The variance τ_0^2 is plausible for the realised rates $\hat{p}_k$ if the realised value S_0 is not in the tail of the empirical distribution of S. In a more complete analysis, we find the quantile of the synthetic distribution that matches S_0, and establish the range of plausible values of τ^2, or of τ.

Figure 2 plots these quantiles for $0 < a < 0.20$. We adopt the conventional standard of regarding values in the 5% tail as not conforming with the variance used for simulations (horizontal dashes). The diagram shows that values of a up to 0.098 are plausible; $a = 0.098$ corresponds to the beta distribution $B(10.20, 313.98)$, which has

standard deviation $\tau = 0.00968$, that is, nearly 1%. Thus, standard deviations τ up to about 1% are plausible. This range includes (all) values regarded as negligible, [0.0, 0.4)%, but also some substantial values. The percentiles 2.5 and 97.5 of the fitted beta distribution are 1.54% and 5.30%, respectively. We can simulate random samples of size $K = 23$ and ask the experts to confirm that such variation is indeed substantial.

FIGURE 2

The upper bound of about 1% is contingent on the beta-binomial model, for which we have no support other than analytical convenience. In Supplementary Materials we repeat this exercise of finding the range of plausible values of τ by applying other realistic models. This serves as a sensitivity analysis; it confirms that the range of plausible values of τ depends on the model adopted for $\{p_k\}$ only slightly, and does not alter the conclusion that both negligible and substantial variation are plausible.

4 Negligible or substantial?

Let the two courses of action be A, appropriate if $\tau^2 \leq \tau_0^2$, and B, appropriate if $\tau^2 > \tau_0^2$. Our analysis will conclude with a data-dependent choice, a or b, referred to as the *verdict*. Choosing a when A holds (aA), and choosing b when B holds (bB), are correct verdicts, associated with no loss. The verdict-state pairs aB and bA are associated with respective losses L_{aB} and L_{bA}.

As a general concept, the loss quantifies the consequences (ramifications) of the incorrect decisions. We refer to its unit as a *lossile*. Its essential property is additivity: a loss of l_1 lossiles in one instance and of l_2 lossiles in another is for all intended purposes equivalent to a single loss of $l_1 + l_2$ lossiles. In brief, a lossile is like a unit of currency, of which we want to spend as little as possible. The nature of loss is not altered if the currently adopted lossile cL is replaced by a new lossile nL at the rate of $\rho : 1$, as is done in currency exchange. No generality is therefore lost by setting $L_{\mathrm{bA}} = 1$. In fact, all that matters is the loss ratio $R = L_{\mathrm{aB}}/L_{\mathrm{bA}}$.

We want to choose the verdict, a or b, that is associated with smaller expected loss. In the frequentist perspective the expected loss depends on the unknown parameter τ^2. This dependence can be removed by the fiducial argument, [8] and [9], which justifies the switch of the fixed and random status for the values of the parameters and the

data after the data is realised. This can be regarded as a frequentist adaptation of the Bayesian paradigm. See [10] for background and examples.

We consider the beta-binomial model for the rates p_k. The maximum likelihood (ML) estimate of the shape parameters α and β of the (marginal) distribution of these rates is found by the Newton-Raphson method described in the Appendix. It converges rapidly, requiring in our case eleven iterations to yield the estimates $\hat{\alpha} = 30.96$ and $\hat{\beta} = 959.48$. The (asymptotic) standard errors estimated from the observed information matrix are 22.41 and 693.42. The confidence intervals based on them are inappropriate because they would suggest that negative values of both α and β are plausible. This can be remedied by defining the confidence region as the points (α, β) for which $l(\alpha, \beta) \geq l(\hat{\alpha}, \hat{\beta}) - \frac{1}{2} q_{0.95,2}$, where $q_{p,j}$ is the p-quantile of the chi-squared distribution with j degrees of freedom (χ^2_j); $q_{0.95,2} = 5.99$.

We prefer the parametrisation in terms of the expectation $p = \alpha/(\alpha + \beta)$ and variance $\tau^2 = p(1-p)/(\alpha+\beta+1)$, because the latter is the parameter of interest. Also, asymptotic normality implies that the ML estimators $\hat{p}$ and $\hat{\tau}^2$ are weakly correlated and have (marginally) normal and scaled chi-squared distributions.

4.1 Approximating loglikelihood by linearisation

We assess the proximity of the fiducial distributions of p and τ^2 to the respective (univariate) normal and inverse chi-squared distributions by the following method. Let the loglikelihood be $l(p)$. If the corresponding distribution is normal, then l is a quadratic function and its differential (the score) l' is linear. A loglikelihood without a tractable analytical form, such as the marginal loglikelihood for τ in the beta-binomial model, is evaluated pointwise on a regular grid of values of τ. If the score l' does not have a closed form it is approximated by the scaled finite differences $g(\tau; \Delta) = \{l(\tau + \Delta) - l(\tau)\}/\Delta$ for small Δ. Normal likelihoods are characterised by linear score. We assess whether l' is linear informally from the plot of g, or by the deviation of g from a linear approximation. Proximity to some other familiar distributions is assessed by transformations of the score to linearity.

The Student t distribution with M degrees of freedom is given by the density

$$f_M(x) = \frac{\Gamma_2(M+1)}{\sqrt{M\pi}\,\Gamma_2(M)} \left(1 + \frac{x^2}{M}\right)^{-(M+1)/2},$$

where $\Gamma_2(x) = \Gamma(\frac{1}{2}x)$. The differential of its log-density $l_M(x) = f_M(x)$ is

$$\frac{\partial l_M}{\partial x} \;=\; -\frac{M+1}{2}\,\frac{2x}{M+x^2}\,,$$

so $x\,\partial l_M/\partial x$ is a linear function of $1/(M+x^2)$:

$$x\,\frac{\partial l_M}{\partial x} \;=\; -(M+1) + \frac{M(M+1)}{M+x^2}\,.$$

In practice, M is not known but can be fitted by trial and error, by plotting the values of $1/(M+x^2)$ against the x-multiple of the score. The linearisation is not unique; suitable expressions can be derived also for $x^{-1}\partial l_M/\partial x$ and $x(\partial l_M/\partial x)^{-1}$, but reciprocals of x or of the score may introduce instability at $x = 0$ or at the ML estimate. In a typical setting, we work with a scaled noncentral t distribution. The linear transformation involved is incorporated in the derivations straightforwardly.

The χ^2_M distribution is given by the density

$$f_M(x) \;=\; \frac{1}{2^{M/2}\,\Gamma_2(M)}\,x^{M/2-1}\exp\left(-\frac{x}{2}\right)$$

$(x > 0)$. The differential of the log-density,

$$\frac{l_M}{\partial x} \;=\; \frac{M-2}{2x} - \frac{1}{2}\,,$$

is a linear function of $1/x$ for $x > 0$. This provides a simple way of assessing whether a distribution (or likelihood) is close to a χ^2 — by proximity of the x-multiple of the score to a linear function. If confirmed, then the degrees of freedom can be inferred from the intercept of the linear approximation. Further, if $(x+c)\partial l/\partial x$ is a linear function of x for a constant c, then the likelihood is related to a linear transformation of a χ^2 distribution. Such distributions and their generalisations arise in estimation of the cluster-level variance in random-effects analysis of variance, [6] and [7]. The number of degrees of freedom can be approximated from the intercept of the $(x+c)$-multiple of the score.

For the density of the inverse chi-squared distribution with M degrees of freedom (χ^{-2}_M),

$$f(x) \;=\; \frac{1}{2^{M/2}\,\Gamma_2(M)\,x^{1+M/2}}\exp\left(-\frac{1}{2x}\right),$$

we obtain by similar operations the identity

$$x^2\,\frac{\partial l}{\partial x} \;=\; -\left(\frac{M}{2}+1\right)x - 1\,. \tag{1}$$

The χ^{-2} distribution arises in the application of the fiducial argument to a typical variance estimator. Suppose $\hat{\tau}^2$ is such that $X = M\hat{\tau}^2/\tau^2$ has χ^2_M distribution. Then the fiducial distribution of τ^2, derived from $M\hat{\tau}^2/X$, is the $M\hat{\tau}^2$-multiple of the χ^{-2}_M distribution.

When the expectation p and variance τ^2 are estimated simultaneously, we apply these approximations to the marginal likelihood obtained by (numerically) integrating over the nuisance parameter p or τ^2. In the application in Section 5 we find that the loglikelihood for variance τ^2 is well approximated by a χ^{-2} density after a linear transformation.

4.2 Decision theory

Let A be the course of action appropriate if $\tau^2 \leq \tau_0^2$ and B the other contemplated course, appropriate if $\tau^2 > \tau_0^2$. Suppose the loss associated with the verdict-state aB, L_{aB}, is constant, and $L_{\mathrm{bA}} = 1$, so that $L_{\mathrm{aB}} = R$. Denote $\theta_\mu(x) = (x+c)/\{\mu(\hat{\tau}^2 + c)\}$ as a function of $x > -c$, with $\mu > 0$ a parameter. If for a known constant c the fiducial distribution of $\tau^2 + c$ is a scaled χ^{-2},

$$f_M\Big(x; \hat{\tau}^2, c\Big) \;=\; \frac{\{M\,(\hat{\tau}^2 + c)\}^{M/2}}{2^{M/2}\,\Gamma_2(M)\,(x+c)^{1+M/2}}\;\exp\left\{-\frac{1}{2\theta_M(x)}\right\}$$

$(x > -c)$, then the expected loss associated with verdict b is

$$Q_{\mathrm{bA}} \;=\; \int_{-c}^{\tau_0^2} f_M\Big(x; \hat{\tau}^2 + c, c\Big)\;\mathrm{d}x \;=\; F_M\Big\{\theta_M\Big(\tau_0^2\Big)\Big\}\;,$$

where $F_M(y)$, with a single argument, is the distribution function for the density f_M and $F_M(x; \hat{\tau}^2, c) \;=\; F_M\{\theta_M(x)\}$; we omit the arguments when $c = 0$ and $\hat{\tau}^2 = 1$. Similarly, we obtain the identity

$$Q_{\mathrm{aB}} \;=\; \int_{\tau_0^2}^{+\infty} f_M\left(x; \hat{\tau}^2, c\right)\;\mathrm{d}x \;=\; L_{\mathrm{aB}}\left[1 - F_M\Big\{\theta_M\Big(\tau_0^2\Big)\Big\}\right]\;.$$

We choose verdict a when $Q_{\mathrm{aB}} < Q_{\mathrm{bA}}$, that is, when

$$\hat{\tau}^2 \;<\; \frac{\tau_0^2 + c}{M\,F_M^{-1}\left(\frac{R}{1+R}\right)} - c\,,$$

and choose b otherwise. Denote the right-hand side of this inequality by τ_*^2; we refer to it as the borderline (between the two verdicts). If the verdict is chosen for which the expected loss is smaller, then the largest possible expected loss is $F_M\{\theta_M(\tau_*^2)\}$.

Piecewise constant loss, given by constants L_{aB} and $L_{\mathrm{bA}} = 1$, often does not adequately capture the consequences of the two kinds of erroneous verdicts. In the context of studying the network-level variation of NEC, it is difficult to argue against constant loss L_{bA} when action B would be equally futile for any negligible variance $\tau^2 \leq \tau_0^2$. However, ignoring greater (more substantial) variation, by verdict a when the state is B, may result in greater loss. This motivates the piecewise linear loss function, $L_{\mathrm{aB}} = R_0 + R_1(\tau^2 - \tau_0^2)$ for $\tau^2 > \tau_0^2$. The slope R_1 quantifies the increase of gravity with the increasing misjudgement of the variance being substantial.

Below we apply the identity

$$xF_M\Big(x; \hat{\tau}^2, c\Big) \;=\; \frac{M\,(\hat{\tau}^2 + c)}{M-2}\, F_{M-2}(x)$$

for $M > 2$. It is derived by reference to the densities f_M and f_{M-2}. For the piecewise linear loss we have the identities

$$\begin{aligned}
Q_{\mathrm{aB}} &= \int_{\tau_0^2}^{+\infty} \Big\{R_0 + R_1\left(x - \tau_0^2\right)\Big\}\, f_M\{\theta_M(x)\}\,\mathrm{d}x \\
&= \Big\{R_0 - R_1\left(\tau_0^2 + c\right)\Big\}\Big[1 - F_M\Big\{\theta_M\Big(\tau_0^2\Big)\Big\}\Big] \\
&\quad + R_1 \int_{\tau_0^2}^{+\infty} (x + c)\, f_M\{\theta_M(x)\}\,\mathrm{d}x \\
&= \Big\{R_0 - R_1\left(\tau_0^2 + c\right)\Big\}\Big[1 - F_M\Big\{\theta_M\Big(\tau_0^2\Big)\Big\}\Big] \\
&\quad + \frac{MR_1\,(\hat{\tau}^2 + c)}{M-2}\Big[1 - F_{M-2}\Big\{\theta_M\Big(\tau_0^2\Big)\Big\}\Big]\,,
\end{aligned}$$

so long as $M > 2$. The expected loss Q_{bA} is a decreasing function of $\hat{\tau}^2$ and Q_{aB} is increasing. This can be shown by differentiating Q_{aB} and Q_{bA}. Further, $Q_{\mathrm{aB}} \to +\infty$ and $Q_{\mathrm{bA}} \to 0$ as $\hat{\tau}^2 \to +\infty$. Therefore the borderline equation $Q_{\mathrm{aB}} = Q_{\mathrm{bA}}$ has at most one solution. A nonnegative solution, denoted by τ_*^2, exists when

$$F_M\{\theta_M(\tau_0^2)\} \;\geq\; \frac{R_0}{R_0 + 1}\,.$$

The solution depends on the loss function (its coefficients R_0 and R_1), the degrees of freedom M, the constant c and the threshold τ_0^2. Verdict a is issued when $\hat{\tau}^2 < \tau_*^2$ and verdict b otherwise. Verdict b is issued for all values of $\hat{\tau}^2$ when there is no nonnegative solution.

The piecewise quadratic loss function is defined as $L_{\mathrm{aB}} = R_0 + R_2(\tau^2 - \tau_0^2)^2$ for $\tau^2 > \tau_0^2$. We do not regard it as useful in practice for a variance because τ^2 already

involves squaring. The piecewise quadratic loss may be useful for τ. Note that this loss differs from the piecewise linear loss for τ^2.

5 Application

To arrive at a verdict about the network-level variance of the rates of NEC, whether it is negligible or substantial, we proceed by the following steps. First we fit a beta-binomial (hierarchical) model to the observed rates. Then we approximate the fiducial distribution of the network-level variance by a χ^{-2} distribution. And finally, we compare the expected losses Q_{aB} and Q_{bA}, and issue the verdict that has smaller expected loss. We deal with the ambiguity about some of the settings of the analysis, the threshold variance τ_0^2 and the loss ratio R in particular, by a sensitivity analysis.

5.1 Model fit and approximations

The parameters of the (marginal) beta distribution are estimated by ML using the Newton-Raphson algorithm. The estimates are $\hat{\alpha} = 30.96$ and $\hat{\beta} = 959.48$, implying the estimated marginal rate 3.13% and standard deviation 0.55%.

Although we are concerned primarily with the network-level variance τ^2, we discuss briefly making decisions about the expectation p, which may be of interest in similar applications. A set of profile likelihoods for p is drawn by thin solid lines in the left-hand panel of Figure 3, with the values of τ printed at the right-hand margin. It is difficult to judge how close these profile loglikelihoods are to quadratics. It is much easier to judge whether their differentials, plotted in the right-hand panel, are linear. For greater values of τ they are close to linearity, but for small values their curvature is substantial, so the fiducial conditional distribution of p given τ^2 is not normal. From the diagram we also conclude that the estimators of p and τ^2 are not independent, since the profiles have maxima at different probabilities p.

FIGURE 3

The marginal loglikelihood, plotted by full solid line in the left-hand panel, is distinctly not normal. This is clear from the pronounced curvature of the score function (full solid line in the right-hand panel). No noncentral t distribution offers a good approximation either. We omit the details and focus on τ^2.

The marginal loglikelihood for τ^2 is approximated very well by a linearly transformed χ^{-2} density. That is, it corresponds to $1/(\kappa X + c)$ for suitable constants c and κ, where X is a random variable with a χ^2 distribution. Denote $\Lambda_c(x) = (x+c)^2 \partial l/\partial x$. If the fiducial distribution of $(\tau^2 + c)/\kappa$ is χ_M^{-2}, then $\Lambda_c = -(1 + \frac{1}{2}M)(x + c) + 1/(2\kappa^2)$. The constant $c = 0.395$ is found by trial and error as the value for which Λ_c is closest to a linear function. The process of finding c is illustrated in Figure 4. The left-hand panel presents the (marginal) loglikelihood l for τ^2, from which it would be difficult to assess the proximity to a linearly transformed χ^{-2} density. The right-hand panel displays the function $(c+x)^2 \partial l/\partial x$ for a selection of constants c. They include $c = 0.395$, for which the function has only slight curvature. As an aid to judge this, an approximating straight line is drawn by wide gray dashes. The curvature of the transformed loglikelihood changes smoothly, so finding the linearising constant is straightforward. The thin dashes mark the ML estimator, $\hat{\tau}^2 = 0.335$.

FIGURE 4

The slope of the linear fit is 9.818 and so, according to equation (1), the estimator is associated with $M = 17.64$ degrees of freedom. Further, the intercept, equal to 3.17, implies that $\kappa = 0.397$. We conclude that the fiducial distribution of τ^2 is such that $(v + 0.395)/0.397$ is well approximated by the $\chi^{-2}_{17.64}$ distribution. In the model of one-way analysis of variance with normally distributed outcomes and random effects (of $K = 23$ clusters), the cluster-level variance would have 22 degrees of freedom if every cluster was very large. When some of the clusters are small or of moderate size, there is bound to be fewer degrees of freedom. Thus $M = 17.64$ is in accord with intuition. The constant κ^2 is the fiducial variance of τ^2. It can be (loosely) interpreted as the sampling variance of $\hat{\tau}^2$, and with greater rigour as the posterior variance of τ^2 assuming a constant (improper) prior for τ^2.

5.2 Verdict

Suppose first that the loss is piecewise constant. We do not specify the value of the loss ratio R, but find the borderline value R^*; for $R < R^*$ verdict a and for $R > R^*$ verdict b is issued. The value of R^* is obtained by solving the balance equation $Q_{\mathrm{aB}} = Q_{\mathrm{bA}}$,

see Section 4.2. It yields

$$R^* \;=\; \frac{F_M(\tau_0^2;\hat{\tau}^2,c)}{1-F_M(\tau_0^2;\hat{\tau}^2,c)}\,, \tag{2}$$

where F_M is the distribution function of χ_M^{-2}. The solution for $\tau_0^2 = 0.16$ is $R^* = 0.205$. If the entire plausible range of loss ratios, $(R_{\rm L}\,, R_{\rm H})$, is on one side of R^*, then the verdict is unequivocal — a if $R_{\rm H} < R^*$, and b if $R_{\rm L} > R^*$.

Setting the values of τ_0^2 and R_0, by eliciting them from experts (clients, or sponsors of the analysis), is a contentious exercise, often perceived to be out of character with other interaction between the analyst and the expert. In our study, we failed to convene a meeting in which such elicitation would be conducted; this issue is discussed in the next section. From the background to the general problem, we concluded that a plausible range of values of τ_0^2 is $(0.09, 0.25)$, corresponding to plausible standard deviations in the range $(0.3, 0.5)$. We also concluded that $R_{\rm L} > 1$ because the failure to discover that the networks' rates differ (verdict-state aB) would discourage any research that might lead to distinguishing good and poor practice, whereas verdict-state bA would lead to futile research. With this perspective, the verdict is b, that the variation is substantial. The borderline R^*, as a function of τ_0^2, increases from $R^*(0.09) = 0.09$ to $R^*(0.25) = 0.45$.

As an illustration, we discuss a different perspective in which the verdict would be equivocal. Figure 5 presents the borderline function $R^*(\tau_0^2)$ for $0 \le \tau_0^2 \le 0.36$. Points (τ_0^2, R) above the line correspond to verdict b, and points underneath to verdict a. The outer rectangle represents the plausible range $(0.09, 0.25) \times (0.30, 0.50)$ for τ_0^2 and R. The verdict is ambiguous because the borderline function $R^*(\tau_0^2)$ intersects the plausible rectangle — for some plausible pairs (τ_0^2, R) one verdict and for other plausible pairs the other verdict is appropriate.

FIGURE 5

Suppose next that the plausible range is narrower, namely $(0.12, 0.20) \times (0.30, 0.45)$, marked by darker shading. Now the verdict is b for all plausible values of τ_0^2 and R, owing to the reduction of the upper limit of the plausible range for τ_0^2. This shows how important it is to declare as narrow a plausible range as possible. Integrity of the declaration of the plausible region may be undermined by setting or re-setting it after data collection when the function $R^*(\tau_0^2)$ is already established. Thus, specifying

a narrower plausible region at the planning stage is rewarded by greater integrity and transparency of the analysis, as well as by reduced chance of an ambiguous outcome (impasse). However, it is imperative that all pairs (τ_0^2, R) outside the plausible rectangle can be ruled out.

Some of the effort to reduce the plausible range may be wasteful. The reduction of R_{H} has no impact, and neither has the increase of the lower bound for τ_0^2. The plausible range does not have to be a rectangle, but a rationale for any other shape, with dependence of the plausible range for R on the value of τ_0^2, is difficult to find.

Suppose next that the loss function L_{aB} is linear; $L_{\mathrm{aB}} = R_0 + R_1(\tau^2 - \tau_0^2)$ for $\tau^2 > \tau_0^2$. The borderline value of the linear coefficient R_1 is

$$R_1^* \;=\; \frac{R_0 - (R_0+1)\, F_M\{\theta_M(\tau_0^2)\}}{[1 - F_M\{\theta(\tau_0^2)\}] \;-\; \dfrac{M\,(\hat{\tau}^2 + c)}{M-2}\,[1 - F_{M-2}\{\theta_M(\tau_0^2)\}]}\,,$$

when $R_1^* > 0$. When the solution R_1^* is negative, the verdict is b for all $R_1 > 0$.

Figure 6 displays the borderline coefficient as a function of τ_0^2 for a set of values of R_0. It shows that, for a given R_0, R_1^* is positive for τ_0^2 greater than an initial value $\tau_0^2(R_0)$. For example, $\tau_0^2(0.3) \doteq 0.20$. For $\tau_0^2 < \tau_0^2(R_0)$, verdict b is appropriate for all $R_1 > 0$. The function $\tau_0^2(R_0)$ is increasing — verdict a is more appealing for greater τ_0^2.

FIGURE 6

The functions $R_1^*(R_0, \tau_0^2)$ are increasing, and their gradients become steeper with increasing τ_0^2. For large R_0, R_1^* has a steep gradient even at $\tau_0^2(R_0)$. In the perspective introduced with piecewise constant loss with $R > 1$, there is very little scope for linear loss because $\tau_0^2(R_0)$ would be smaller than a plausible value of τ_0^2 only for very small $R_0 \ll 1$. That would contradict the presumed aversion to the verdict-state pair aB.

6 Discussion

In everyday life, we often have to choose one out of a small number of alternative courses of action (options). When we are not certain about which option is optimal, we consider what is at stake, and weigh carefully the consequences of the suboptimal choices. Hypothesis testing is deficient for such a task (involving two options), especially if we rigidly adhere to the ubiquitous convention of the 5% level of significance.

Further problems arise when we want to combine several hypothesis tests because of the complex calculus of the probabilities involved. The results of hypothesis tests are often misinterpreted because such a test is poorly suited for deciding whether the hypothesis or the alternative holds, [11]. Our main objection to hypothesis testing in general is that it is oblivious to the consequences of the two kinds of error that are committed with nontrivial probabilities. These consequences may be difficult to assess, but an analysis disregarding them is inconsequential, or caters for a default perspective that may be in discord with the client's perspective. For an example of a transparent failure of hypothesis testing that was followed by a reanalysis informed by an elicitation exercise, see [12].

In the approach we propose, loss is introduced as a currency for error and the goal is to choose one of the alternative courses of action that, based on the recorded data, is associated with the smallest expected loss. This is implemented more naturally in the Bayesian paradigm, by operating with posterior distributions, [13] – [15]. Making several decisions introduces no difficulties in addition to making each decision separately because the losses involved, and their expectations, are additive. In the application presented in Section 5 we prefer the frequentist paradigm because it is more familiar to the community of neonatologists. Also, we want to demonstrate that decision theory is not in the exclusive domain of the Bayesian.

Our approach is more demanding on input: a, the parameter space has to be split into the subspaces in which each contemplated course of action is preferred (associated with no loss); b, the losses (or loss functions) associated with the discordant verdict-state pairs (kinds of bad decision) have to be declared. The first element is a counterpart of partitioning the parameter space to the hypothesis and alternative. The essential difference is that the courses of action are treated symmetrically, unlike a hypothesis and its alternative in hypothesis testing. In another aspect, the actions are treated asymmetrically — the consequences of choosing them inappropriately are in general uneven. When the parameter space is a real interval that contains zero in its interior and the fiducial distribution is continuous at zero, associating zero on its own with a course of action would result in never selecting it because the innumerably many alternatives in arbitrarily close distance from zero are a safer statistical bet.

We see the main difficulty in adopting a decision-theoretical approach not in the

technical complexity, but in the mainstream culture of statistical analysis and its interpretation which is wedded to the deeply ingrained framework of hypothesis testing. It delegates the selection of the post-analysis agenda (for production, management or further research) to parties outside the profession of statistics. We have argued by an example that this selection is a statistical task because it entails nontrivial statistical evaluations. The terms and mode of cooperation of the analyst and the expert have to be adapted to share the perspective, purpose and value judgements and convert them to a format suitable for separating the parameter space and for declaring the loss function. In many applications they are disregarded and not integrated in the analysis.

With no uncertainty about the threshold variance τ_0^2 and the loss ratio R, the application presented in Section 5 would conclude unequivocally with the choice of substantial variation of the network-level rates of NEC. Uncertainty about the key inputs is dealt with by sensitivity analysis, exploring the verdicts that would be issued for all the plausible settings. If both verdicts are plausible, then we reach an impasse. So, investment in a meticulous elicitation exercise, which concludes with a narrow (but credible) plausible range, is rewarded by reducing the chances of an impasse.

The subjectivity of the analysis might at first appear as a drawback. However, it is an essential feature of any analysis that is finely tuned to the perspectives, judgements and values of the client.

Acknowledgements

The data for this analysis were extracted from the National Neonatal Research Database (NNRD) by Dr. Cheryl Battersby who, together with Professor Neena Modi, introduced the author to the problem analysed in this paper. Data is contributed to NNRD by the clinical staff from all neonatal units in England.

References

[1] Sullivan S, Schanler RJ, Kim JH, Patel AL, Trawöger R, Kiechl-Kohlendorfer U, Chan GM, Blanco CL, Abrams S, Cotten CM, Laroia N, Ehrenkranz RA, Dudell G, Cristofalo EA, Meier P, Lee ML, Rechtman DJ, Lucas A. An exclusively human milk-based diet is associated with a lower rate of necrotizing enterocolitis than a

diet of human milk and bovine milk-based products. *Journal of Pediatrics* 2010; 156:562–567.

[2] Neu J. Preterm infant nutrition, gut bacteria, and necrotizing enterocolitis. *Current Opinion in Clinical Nutrition and Metabolic Care* 2015; **18**:285–288.

[3] Battersby C, Longford N, Mandalia S, Costeloe K, Modi N, on behalf of the UK Neonatal Collaborative Necrotising (UKNC-NEC) study group. Incidence of severe neonatal necrotising enterocolitis across neonatal networks and with different enteral feeding approaches in England, 2012–13: a whole-population surveillance study. *The Lancet, Gastroenterology and Hepatology* 2017; **2**; 43–51.

[4] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.

[5] Sutton, AJ, Abrams, KR, Jones, DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research.* Wiley: Chichester, UK, 2000.

[6] Crainiceanu CM, Ruppert D. (2004) Likelihood ratio tests in linear mixed models with one variance component *Journal of the Royal Statistical Society* Series B 2004; **66**:165–185.

[7] Greven S, Crainiceanu CM, Küchenhoff H, Peters H. Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics* 2008; **17**:870–891.

[8] Fisher RA. Statistical methods and scientific induction. *Journal of the Royal Statistical Society* Series B 1955; **17**:69–78.

[9] Seidenfeld T. R. A. Fisher's fiducial argument and Bayes' theorem. *Statistical Science* 1992; **7**, 358–368.

[10] Longford NT. *Statistical Decision Theory.* Springer-Verlag: Heidelberg, Germany, 2013.

[11] Lindley DV. Decision analysis and bioequivalence trials. *Statistical Science* 1998; **13**:136–141.

[12] Longford, NT. Policy-related small-area estimation. *South African Statistical Journal* 2015; **49**:105–119.

[13] Berger JO. *Statistical Decision Theory and Bayesian Analysis.* 2nd Ed. Springer-Verlag: New York, 1985.

[14] Lindley DV. *Making Decisions.* 2nd Ed. Wiley: Chichester, UK, 1985.

[15] DeGroot MH. *Optimal Statistical Decisions.* Wiley: New York, 2004.

Appendix. Newton-Raphson algorithm for the beta-binomial model

The likelihood for the beta-binomial model is

$$L(\alpha,\beta) \;=\; C\left\{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\right\}^K \prod_{k=1}^K \int_0^1 p^{m_k+\alpha-1}(1-p)^{n_k-m_k+\beta-1}\,\mathrm{d}p\,,$$

where α and β are the shape parameters of the marginal beta distribution and C is the constant for which this expression is a density function; its value is immaterial for maximising $L(\alpha,\beta)$. The kth integrand is proportional to the density of the beta distribution with parameters $m_k+\alpha-1$ and $n_k-m_k+\beta-1$, and so the integral has a closed form. The loglikelihood is

$$\begin{aligned} l(\alpha,\beta) \;&=\; \log(C) + K\left\{\Gamma^{(1)}(\alpha+\beta)-\Gamma^{(1)}(\alpha)-\Gamma^{(1)}(\beta)\right\} \\ &\quad + \sum_{k=1}^K \left\{\Gamma^{(1)}(\alpha+m_k)+\Gamma^{(1)}(\beta+n_k-m_k)-\Gamma^{(1)}(\alpha+\beta+n_k)\right\}\,, \end{aligned}$$

where $\Gamma^{(1)}(x)=\log\{\Gamma(x)\}$. Denote the digamma function by $\Gamma^{(2)}$ and the trigamma function by $\Gamma^{(3)}$. They are the respective first- and second-order derivatives of $\Gamma^{(1)}$.

The Newton-Raphson algorithm comprises iterations; the new (provisional) solution in iteration $t+1$ is

$$\hat{\boldsymbol{\gamma}}_{t+1} \;=\; \hat{\boldsymbol{\gamma}}_t + \mathbf{H}^{-1}(\hat{\boldsymbol{\gamma}}_t)\,\mathbf{s}\,(\hat{\boldsymbol{\gamma}}_t)\;,$$

where $\hat{\boldsymbol{\gamma}}_t=(\hat{\alpha}_t\,,\hat{\beta}_t)$, $\mathbf{s}$ is the score vector and $\mathbf{H}$ the observed information matrix. In iteration t, $\mathbf{H}$ and $\mathbf{s}$ are evaluated at the current solution $\hat{\boldsymbol{\gamma}}_t$. The expressions for them

are easy to derive thanks to the definitions of $\Gamma^{(2)}$ and $\Gamma^{(3)}$:

$$\begin{aligned}
\frac{\partial L}{\partial \alpha} &= K\Gamma^{(2)}(\alpha+\beta) - K\Gamma^{(2)}(\alpha) + \sum_{k=1}^{K} \Gamma^{(2)}(\alpha+m_k) - \Gamma^{(2)}(\alpha+\beta+n_k) \\
\frac{\partial L}{\partial \beta} &= K\Gamma^{(2)}(\alpha+\beta) - K\Gamma^{(2)}(\beta) + \sum_{k=1}^{K} \Gamma^{(2)}(\beta+n_k-m_k) - \Gamma^{(2)}(\alpha+\beta+n_k) \\
\frac{\partial^2 L}{\partial \alpha^2} &= K\Gamma^{(3)}(\alpha+\beta) - K\Gamma^{(3)}(\alpha) + \sum_{k=1}^{K} \Gamma^{(3)}(\alpha+m_k) - \sum_{k=1}^{K} \Gamma^{(3)}(\alpha+\beta+n_k) \\
\frac{\partial^2 L}{\partial \beta^2} &= K\Gamma^{(3)}(\alpha+\beta) - K\Gamma^{(3)}(\beta) + \sum_{k=1}^{K} \Gamma^{(3)}(\beta+n_k-m_k) - \sum_{k=1}^{K} \Gamma^{(3)}(\alpha+\beta+n_k) \\
\frac{\partial^2 L}{\partial \alpha\, \partial \beta} &= K\Gamma^{(3)}(\alpha+\beta) - \sum_{k=1}^{K} \Gamma^{(3)}(\alpha+\beta+n_k)\ .
\end{aligned}$$

The algorithm converges rapidly, requiring 6 – 15 iterations to achieve precision to eight decimal places. That is, iterations are stopped when

$$\left\|\hat{\boldsymbol{\gamma}}_{t-1} - \hat{\boldsymbol{\gamma}}_t\right\|^2 + (l_{t+1} - l_t)^2 \ < \ 10^{-16}\,.$$

An initial solution $\hat{\boldsymbol{\gamma}}_0$ is required for the first iteration. It has to be such that $\mathbf{H}(\hat{\boldsymbol{\gamma}}_0)$ is negative definite. An estimate that implies an expectation similar to the overall rate $\hat{p}$ is satisfactory. For example, we set $\hat{\alpha}_0 = 3$ and $\hat{\beta}_0 = 100$, so that $\hat{p}_0 = \hat{\alpha}/(\hat{\alpha}+\hat{\beta}) = 0.029$, and after 11 iterations obtained $\hat{\alpha} = 31.0$ and $\hat{\beta} = 959.5$, implying $\hat{p} = 0.0313$.

Table 1: Infants born at gestational ages of less than 32 weeks and cases of NEC in the English neonatal networks in 2013 – 14.

k	n_k	m_k	%	k	n_k	m_k	%	k	n_k	m_k	%
1_4	391	8	2.05	9_7	755	31	4.11	17_8	344	7	2.03
2_6	635	24	3.78	10_8	600	23	3.83	18_9	589	18	3.06
3_5	800	37	4.62	11_{12}	418	13	3.11	19_5	639	14	2.19
4_8	871	35	4.02	12_{11}	780	37	4.74	20_{10}	591	26	4.40
5_9	470	15	3.19	13_7	797	17	2.13	21_6	520	15	2.88
6_6	409	11	2.69	14_5	649	18	2.77	22_7	658	13	1.98
7_4	421	9	2.14	15_6	623	20	3.21	23_9	931	22	2.36
8_4	1014	30	2.96	16_8	773	19	2.46	All_{164}	14 678	462	3.15

Notes: k — network (the number of hospitals that contribute to the data with 10 or more infants is given in the subscript); n_k — number of qualifying infants cared for in the network's neonatal units (including small units); m_k — number of cases of NEC; % — observed rate of NEC, $100m_k/n_k$.

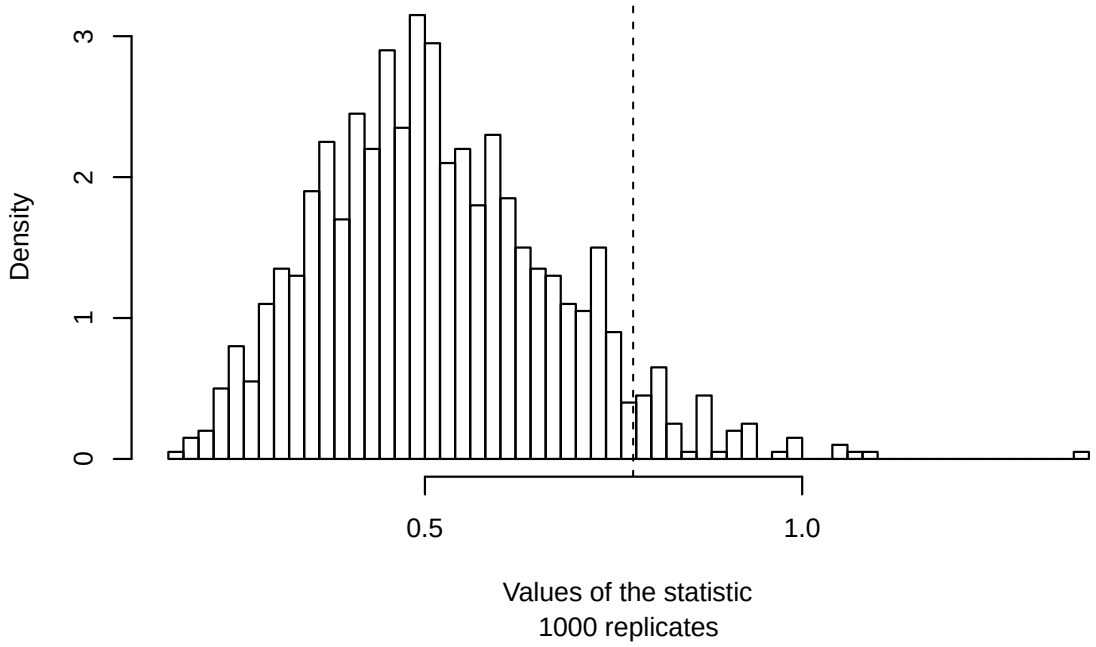

Figure 1: The observed and replicate (simulated) values of the statistic S.

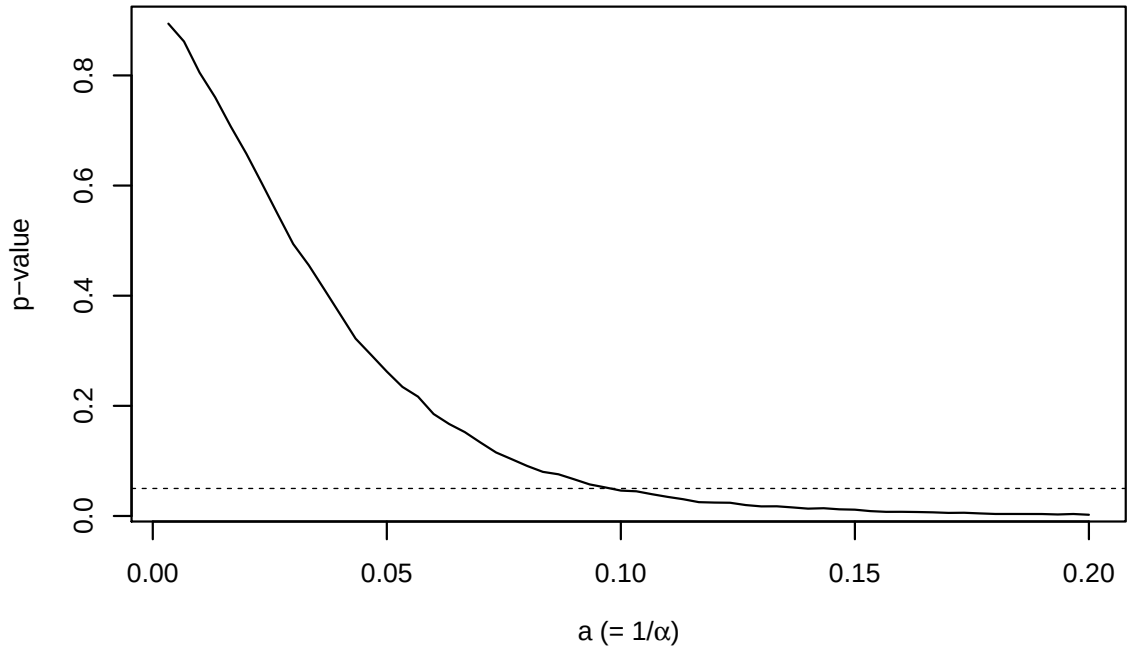

Figure 2: Empirical quantiles of the realised value of S in sets of 1000 replicates; beta-distributed rates p_k .

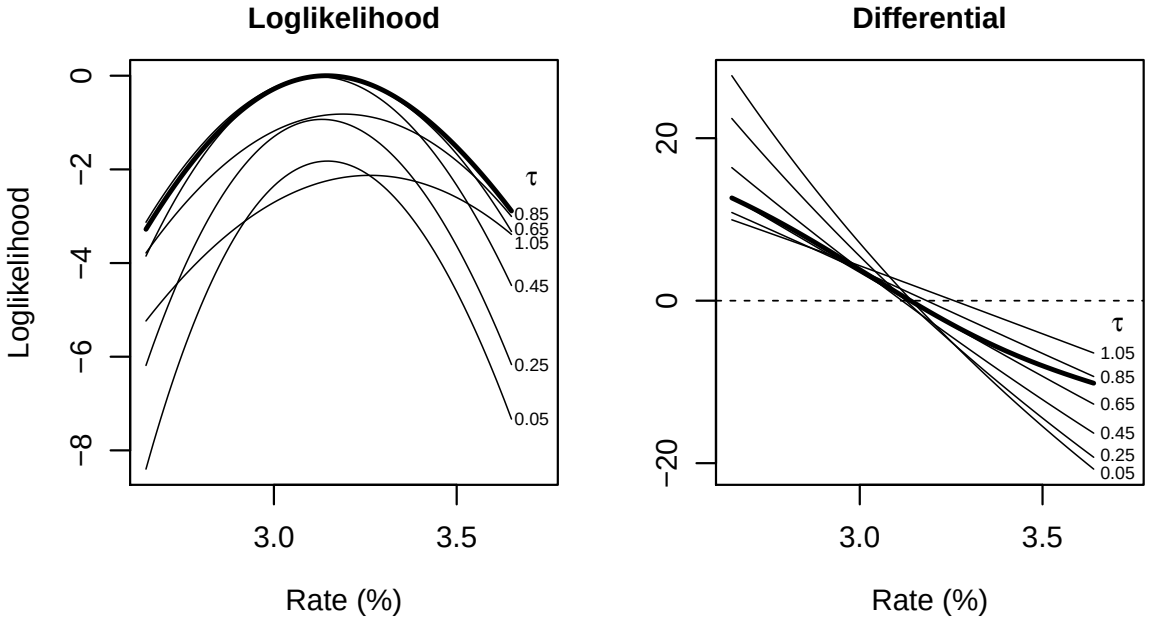

Figure 3: Loglikelihoods for $100p$ and their differentials, in profile with conditioning on the standard deviation τ (thin lines) and marginal after integrating over τ^2 (full solid line).

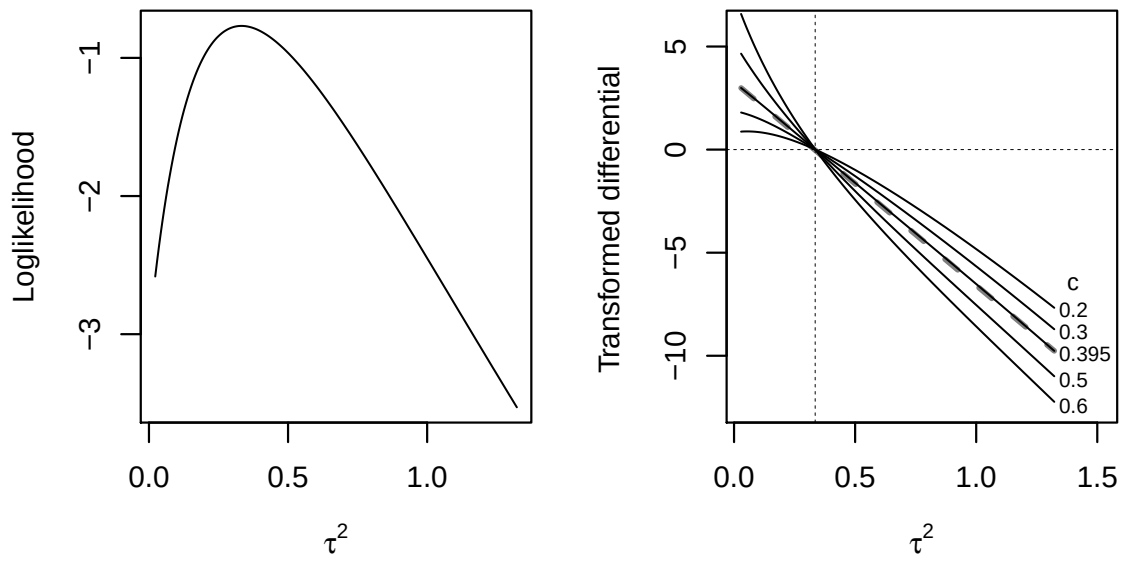

Figure 4: Loglikelihood for τ^2 ($\%^2$) and its transformed differential, for a set of values of c indicated at the left-hand margin.

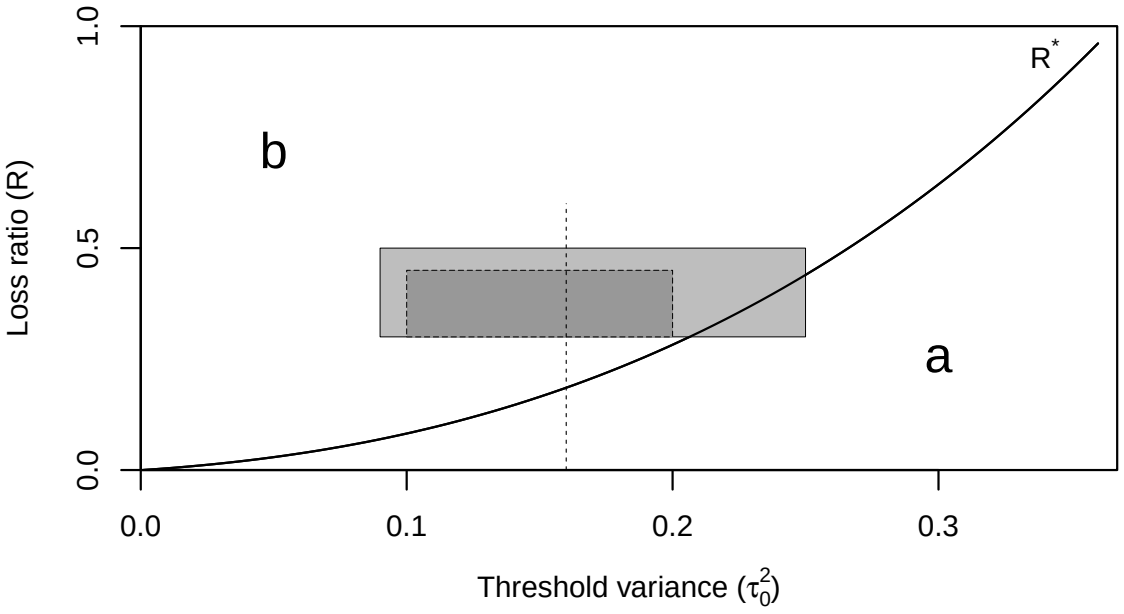

Figure 5: Borderline loss ratio R^* as a function of the threshold variance τ_0^2. The shaded rectangles delineate two plausible ranges for (τ_0^2, R).

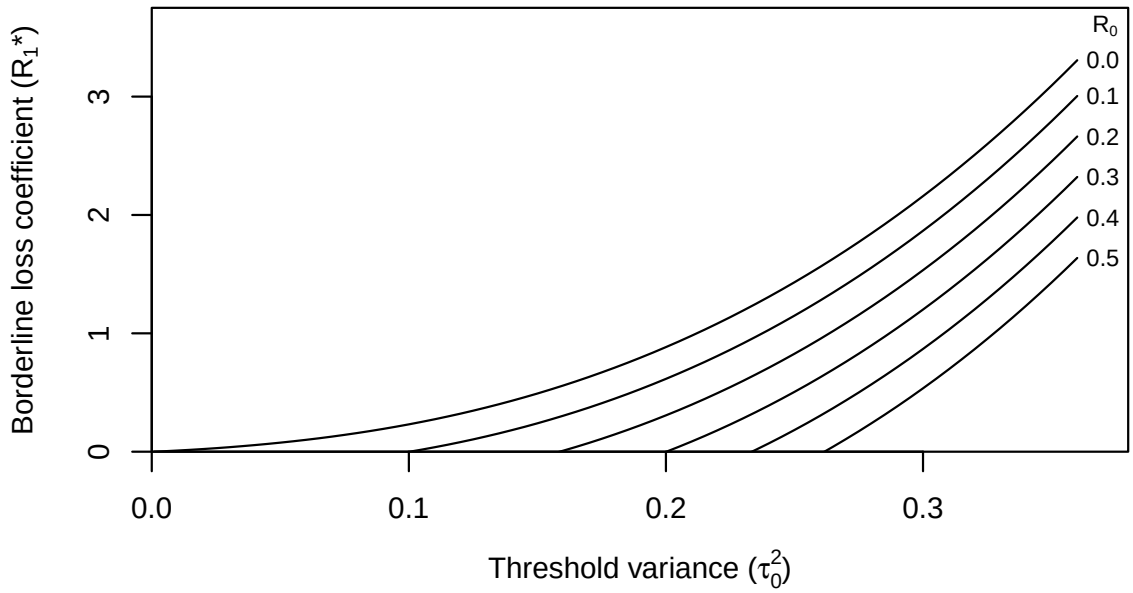

Figure 6: Borderline linear loss coefficient R_1^* as a function of the threshold variance τ_0^2 and the absolute loss coefficient R (values printed at the right-hand margin).