EFFECTIVE SAMPLE SIZE IN A
DICHOTOMOUS PROCESS WITH NOISE

James Cussens

Department of Mathematics, Glasgow Caledonian University,
Cowcaddens Road, Glasgow, G4 0BA, Scotland, UK

Key Words and Phrases: Bayesian inference; Beta distribution; Information loss; Noise.

ABSTRACT

The effect of noise in a dichotomous process is studied from the Bayesian viewpoint. Winkler’s approximation to the posterior distribution in the presence of noise is shown to break down badly near the limits of its application. Information loss is measured using effective sample size. An account of the relationship between effective sample size/information loss and sampling data is given which differs sharply from that of previous work in this area.

1 INTRODUCTION

Here we are concerned with estimating a proportion \( \pi \) from data generated by a dichotomous process. For example, we might wish to estimate the proportion of defective items in a population by taking a sample from the population and using the number of defective and non-defective items to form an estimate of the proportion of defective items in the population as a whole. Such estimates clearly have an important rôle in industrial quality control.

In real-life applications, we have to contend with the possibility of noise in the sample. Suppose we have a sample in which we observe a certain number of ‘successes’ and a certain number of ‘failures’. If there is noise in the data, then there is a probability \( \tau \) that a ‘success’ is misclassified as a ‘failure’ and a probability \( \delta \) that a ‘failure’ is misclassified as a ‘success’. In brief, 
\[
\text{pr(failure observed|actual success)} = \tau, \text{pr(success observed|actual failure)} = \delta
\]
\[ \delta. \] If \( \pi \) is the proportion of successes in the population, then the probability of observing a success becomes \( \Theta = \pi(1-\tau) + (1-\pi)\delta \), where \( \Theta = \pi \) if \( \tau = \delta = 0 \).

Using a Bayesian approach in the absence of noise, the prior distribution is generally constrained to be a beta distribution

\[
f(\pi|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1}
\]

since beta distributions are conjugate with respect to Bernoulli processes. However, unless \( \tau = \delta = 0 \), i.e. there is zero noise, the data-generation process is no longer Bernoulli in \( \pi \), once noise is considered.

With noise, the posterior distribution, rather than being simply a beta distribution is a mixture of beta distributions, as shown by Winkler & Franklin [3]. In that paper an approximate approach to updating with noisy data is given, which, as well as being simple to use, claims to shed light on the relationship between sample frequency and information loss due to noise.

## 2 WINKLER’S APPROXIMATION

To overcome this problem approximate approaches have been suggested by Winkler & Franklin [3] and Winkler [2]. Two approximations are considered by Winkler & Franklin [3], the second of which focuses on approximating the likelihood function.

The likelihood function associated with \( r \) successes out of \( n \) observations is

\[
\ell(r, n|\pi) = [(1 - \tau - \delta)\pi + \delta]^r [(\tau + \delta - 1)\pi + 1 - \delta]^{n-r}
\]

Now, the likelihood function

\[
\ell^\ast(r^\ast, n^\ast|\pi) = \pi'^r (1 - \pi')^{n^* - r^*}
\]

is Bernoulli in \( \pi \) and so a beta prior distribution is conjugate for it. By equating the modes and curvature at the modes of log \( \ell(r, n|\pi) \) and log \( \ell^\ast(r^\ast, n^\ast|\pi) \), when \( \delta < r/n < 1 - \tau \), we have

\[
n^\ast = \frac{n\pi(1 - \pi)}{(\pi + c_1)(1 - \pi + c_2)}
\]  \hspace{1cm} (1)

and

\[
r^\ast = n^\ast \hat{\pi}
\]

where

\[
\hat{\pi} = (r/n - \delta)/(1 - \tau - \delta)
\]

is the maximum likelihood estimator of \( \pi \) and

\[
c_1 = \frac{\delta}{1 - \tau - \delta}, \quad c_2 = \frac{\tau}{1 - \tau - \delta}.
\]
(There is a misprint in Pham-Gia & Turkkan [1], where \( \hat{\pi} \) is given as \( ((r/n) - \tau)/(1 - \tau - \delta) \).

This approximation (which applies only when \( \delta < r/n < 1 - \tau \)) is of particular interest, since \( r \) successes out of a sample of size \( n \) are approximately equivalent to \( r^* \) successes from a noise-free sample of size \( n^* \), \( n^* \) is intended to represent effective sample size. We have \( n^* \leq n \) which corresponds to a loss of information or decrease of sample size due to the presence of noise.

3 PROBLEMS WITH WINKLER’S APPROXIMATION

3.1 WHERE THE APPROXIMATION BREAKS DOWN

One of the main purposes of this paper is to demonstrate deficiencies in Winkler’s approximation. This will be done by comparing various results obtained using the approximation with those obtained using the exact approach given by Pham-Gia & Turkkan [1].

As noted by Winkler [2], \( n^* \) and consequently \( n^*/n \) tend to 0 as \( r/n \rightarrow \delta \) or \( r/n \rightarrow 1 - \tau \). One consequence is that even with very small positive values for \( \tau \) and arbitrarily large \( n \), \( n^* \) can become vanishingly small if \( r/n \) is close to \( 1 - \tau \). So for high values of \( r/n \) there can apparently be a massive decrease in the value of the effective sample size, as soon as we introduce even the smallest amount of noise.

Against this, it will be shown that as \( r/n \rightarrow \delta \) or \( r/n \rightarrow 1 - \tau \), the posterior distribution, as calculated by Winkler’s approximation, diverges sharply from the exact posterior, and that, contrary as to what is claimed by Winkler & Franklin [3], the approximation can become arbitrarily bad as \( n \) increases. We find that, as with many approximations, Winkler’s approximation behaves badly near the limits of its application. The divergence also demonstrates that the claim that effective sample size tends to 0 as \( r/n \rightarrow \delta \) or \( r/n \rightarrow 1 - \tau \) is incorrect.

3.2 COMPARING LIKELIHOOD FUNCTIONS

Let us consider the shapes of \( \ell \) and \( \ell^* \) as \( r/n \rightarrow 1 - \tau \). Analogous results follow as \( r/n \rightarrow \delta \). Since we are considering what occurs as \( r/n \rightarrow 1 - \tau \), we can assume that \( r/n > 1/2 \) in what follows.

Since \( r/n > 1/2 > \delta \), any value of \( \pi \) lying in the interval \([0, \hat{\pi}]\) can be expressed as

\[
\pi = \frac{\theta (\hat{\pi}) - \delta}{1 - \tau - \delta}
\]

for \( \delta/(r/n) \leq \theta \leq 1 \). Now we will compare general values of \( \ell(\pi) \) to the

3
maximal value $\ell(\hat{\pi})$, for $\pi \in [0, \hat{\pi}]$,

\[
\frac{\ell(\pi)}{\ell(\hat{\pi})} = \frac{(\theta (\hat{\pi}))^r (1 - \theta (\hat{\pi}))^{n-r}}{(\theta (\hat{\pi}))^r (1 - \theta (\hat{\pi}))^{n-r}} = \left( \frac{1 - \theta (\hat{\pi})}{1 - \theta (\hat{\pi})} \right)^{n-r} = O(\theta)^n
\]

(2)

Since $\theta < 1$ if $\pi \in [0, \hat{\pi})$, it follows that the value of $\ell$ at non-modal values of $\pi$ can become arbitrarily small as a fraction of the modal value, as $n$ increases. It follows that the modal value of the normalised version of $\ell$: $\ell(\hat{\pi}) / \int_0^1 \ell(\pi)$ can become arbitrarily large.

This is in contrast to the behaviour of $\ell^*$ for values of $r/n$ near $1 - \tau$. As $r/n \to 1 - \tau$, we have $n^* \to 0$ and hence $\ell^*(\hat{\pi}) \to 1$ for all values of $\pi$. $\ell^*$ comes close to the constant function $c(\pi) = 1$ for such values of $r/n$, and hence the normalised version of $\ell^*$ is similarly flat.

We have just seen that $\ell$ and $\ell^*$ can have very different shapes for $r/n$ close to $1 - \tau$. This will shortly be illustrated graphically. However, since equating the modal curvatures of $\log \ell$ and $\log \ell^*$ was intended to ensure that $\ell$ and $\ell^*$ had similar shapes it is important first to investigate what this equation actually implies for $\ell$ and $\ell^*$.

Consider $\ell$ and $\ell^*$ as functions in $\pi$. Since $(\log \ell)'(\hat{\pi}) = (\log \ell^*)'(\hat{\pi}) = 0$, the curvatures at the mode $(\hat{\pi})$ are given simply by $(\log \ell)''(\hat{\pi})$ and $(\log \ell^*)''(\hat{\pi})$. So equating curvatures, we have that

\[
(\log \ell)''(\hat{\pi}) = (\log \ell^*)''(\hat{\pi})
\]

Using the following simple lemma,

**Lemma 1**

\[
f(a) \neq 0, f'(a) = 0 \Rightarrow (\log f)''(a) = \frac{f''(a)}{f'(a)}
\]

**Proof:**

\[
(\log f)''(a) = \left( \frac{f'(a)}{f(a)} \right)' = \frac{f(a)f''(a) + f'(a)^2}{f(a)^2} = \frac{f''(a)}{f(a)}
\]

it follows that

\[
\frac{\ell''(\hat{\pi})}{\ell(\hat{\pi})} = \frac{(\ell^*)''(\hat{\pi})}{\ell^*(\hat{\pi})}
\]

(3)

Now since $\ell(\hat{\pi}) = (\ell^*)'(\hat{\pi}) = 0$, the curvatures at $\hat{\pi}$ are given by $\ell''(\hat{\pi})$ and $(\ell^*)''(\hat{\pi})$. However, Equation 3 falls short of ensuring that these curvatures are equal, which would only occur if $\ell(\hat{\pi}) = \ell^*(\hat{\pi})$. We can have $\ell''(\hat{\pi})$ ‘big’ and $(\ell^*)''(\hat{\pi})$ ‘small’ as long as $\ell(\hat{\pi})$ and $\ell^*(\hat{\pi})$ take appropriate values. Consider
the ratio of the curvatures of the normalised likelihood functions. We have, using Equation 3:

\[
\frac{\ell'(\hat{\pi}) \int_0^1 \ell'(\pi) d\pi}{\int_0^1 \ell(\pi) d\pi} = \frac{\ell(\hat{\pi}) \int_0^1 \ell'(\pi) d\pi}{\int_0^1 \ell(\pi) d\pi} \tag{4}
\]

From the previous section, we know that \(\int_0^1 \ell'(\pi) d\pi / (\ell'(\hat{\pi})) \rightarrow 1 \) as \(r/n \rightarrow 1 - \tau\) and \(\ell(\hat{\pi}) / \int_0^1 \ell(\pi) d\pi \rightarrow \infty\) as \(n \rightarrow \infty\), so it follows that the RHS of Equation 4 increases arbitrarily with \(n\) when \(r/n\) is near \(1 - \tau\). It follows then that, in this case, the modal curvature of \(\ell\) (normalised) also becomes arbitrarily large as a fraction of the modal curvature of \(\ell^*\) (normalised).

### 3.3 Empirical Comparison of Approximate and Exact Posterior Distributions

We will compare approximate and exact posteriors using a uniform prior. If the prior is uniform, then the posterior distribution is simply the normalised likelihood function, viewed as a function of \(\pi\). From above then, we expect the modal value of the posteriors, and the curvature there, to differ greatly if \(r/n\) is near \(1 - \tau\) and \(n\) is fairly large. Figures 1 and 2 show that this is indeed the case.

Since Winkler’s approximation is derived by equating modes and curvature at the modes of the log of the likelihood functions, the posterior distributions have equal modes at \(\pi = \hat{\pi}\). However, as expected, the values at the mode and the modal curvatures are quite different.

### 3.4 Comparing Modal Values

We now show how the modal value of the posterior varies with \(r/n\), using again a uniform prior and for fixed \(n, \tau\) and \(\delta\). We do this for both exact and approximate methods, putting both graphs on the same axis. The results are presented in Figures 3 and 4.

As expected, the posterior modal values are very close for most value of \(r/n\), only diverging significantly as \(r/n\) approaches \(1 - \tau\) or \(\delta\). Also, for the higher value of \(n\) (\(n = 30\)), the approximation is better for ‘moderate’ values of \(r/n\). On the other hand, if \(n\) is big, the divergence is greater for values of \(r/n\) which approach the limits of the approximation’s application (\(\delta\) and \(1 - \tau\)).

### 4 Effective Sample Size

One aim of Winkler’s approximation, discussed above, was to provide an easily calculated approximation to Bayesian updating of a beta distribution in the presence of noise. Given advances in statistical computing, it is now reasonably
easy to compute an exact answer. However, the quantity \( n^* \) as defined in Equation 1 is of general interest nonetheless, since it is intended to give the size of a noise-free sample which is equivalent to the actual noisy sample in the sense of leading to the same posterior distribution.

As the previous section has demonstrated, Winkler’s approximation breaks down badly with \( r/n \) near \( \delta \) or \( 1 - \tau \), with the ‘approximate’ posterior being considerably flatter than the exact one. This means that, for such values of \( r/n \), the value \( n^* \) is much too small. Winkler & Franklin [2] and also Pham-Gia & Turkkan [1] show that \( n^*/n \rightarrow 0 \) as \( r/n \) approaches \( \delta \) or \( 1 - \tau \), and graphs are presented illustrating this and related facts. These are meant to show that information loss due to noise tends to be total for these ‘extreme’ values of \( r/n \). However, since the value \( n^* \) is far too small for such values, this inference does not follow.

We will tackle the question of information loss by finding values \( r_\eta \) and \( n_\eta \) such that the mean and variance of a posterior which is generated by updating with a noise-free sample parameterised by \( r_\eta \) and \( n_\eta \), are equal to the mean and variance of the exact posterior generated by updating with the noisy data.

Let \( \mu \) denote the exact posterior mean and \( \sigma^2 \) denote the exact posterior variance. Let \( \alpha \) and \( \beta \) be the parameters of the prior beta distribution. We have

\[
\mu = \frac{\alpha + r_\eta}{\alpha + \beta + n_\eta}
\]

\[
\sigma^2 = \frac{(\alpha + r_\eta)(\beta + n_\eta - r_\eta)}{(\alpha + \beta + n_\eta)^2(\alpha + \beta + n_\eta + 1)}
\]

From which it follows that

\[
n_\eta = \frac{\mu(1 - \mu)}{\sigma^2} - (\alpha + \beta + 1)
\]

\[
r_\eta = \mu \left( \frac{\mu(1 - \mu)}{\sigma^2} - 1 \right) - \alpha
\]

Using the technique given in Pham-Gia and Turkkan [1], we can calculate \( \mu \) and \( \sigma^2 \), and hence \( n_\eta \) and \( n_\eta/n \) for any values of \( \alpha, \beta, \tau, \delta, r, n \). Figures 5–8 give several graphs mapping \( n_\eta/n \) against \( r/n \) for various values of \( \alpha, \beta, \tau, \delta \) and \( n \). We find that \( n_\eta/n < 1 \) for all values given in the graphs, and it seems likely that this is always the case.

From Figures 5–8, we see that, for most values of \( r/n \), the greater the noise, the lower the graph is. However, with the exception of Figure 7, we find that the graph for \( \delta = 0.3, \tau = 0.1 \) is higher than that for \( \delta = \tau = 0.1 \) for a certain range of low values of \( r/n \). Also, in Figure 6, we find that the graph for \( \delta = \tau = 0.3 \) is higher than that for \( \delta = 0.3, \tau = 0.1 \) for a small range of fairly high values of \( r/n \).

This might appear paradoxical, in that one expects a decrease in \( n_\eta \), the effective sample size, as noise increases. However, the results given should not
be interpreted as indicating that one can increase the effective sample size by adding extra noise to a sample, which would indeed be paradoxical. The key point is that those values of \( r/n \) which render a higher value of \( n_\eta \) for \( \delta = 0.3, \tau = 0.1 \) than for \( \delta = \tau = 0.1 \) are less likely to occur in a sample with noise of \( \delta = 0.3, \tau = 0.1 \) than in a sample with noise at \( \delta = \tau = 0.1 \), whatever the true value of \( \pi \).

Figure 8 is remarkable in that some values of \( n_\eta \) are negative. In these cases, it is not possible to interpret \( n_\eta \) as effective sample size. Informally, these negative values have occurred because the very skewed prior distribution (\( \alpha = 1, \beta = 10 \)) is more ‘informative’ than the flatter posterior, so it is as if data has to be taken away to move from prior to posterior.

Another important feature of Figures 5–8 is that the value of \( n_\eta/n \), unlike that of \( n^*/n \), increases considerably when \( r/n \) is near 0 or 1. Indeed, \( n_\eta \) increases as \( r/n \) approaches \( \delta \) or \( 1 - \tau \), which again differs from \( n^* \), which tends to zero as these values are approached. In short: using \( n_\eta \) demonstrates that information loss due to noise decreases if \( r/n \) is low or high, and this effect is especially apparent for high noise values, and is accentuated as \( n \) increases.

In contrast, previous workers in this area have accepted that information loss is total for high or low values of \( r/n \). Pham-Gia & Turkkan [1] state that

\[
\ldots \text{if } L = (1 - n^*/n) \text{ reflects the information loss due to noise, this loss increases with } \tau \text{ and } \delta \text{ and is almost total as } r/n \text{ approaches } \delta \text{ or } 1 - \tau. \text{ This fact can be easily understood, since then almost all observations classified as successes or failures, can be attributed to misclassifications.}
\]

However it is not the case that, in general, almost all observations will be attributed to misclassifications as \( r/n \) approaches \( \delta \) or \( 1 - \tau \). Suppose \( \pi = 1 \) and \( \tau = \delta = 0.1 \) and that \( r/n \) takes it expected value 0.9. In this case only 10% of the sample has been misclassified, as expected. Suppose that \( \pi \) is unknown, and we wish to use Bayesian inference to find probable values for it, then a result of \( r/n = 0.9 \) with \( \tau = \delta = 0.1 \), rather than containing zero information, is strong evidence that \( \pi \) has a high value. This fact is attested to by the relatively large value of \( n_\eta \) in this case.

## 5 CONCLUSIONS

The central conclusion of this paper is, from a practical perspective, a positive one: information loss due to noise is nowhere near as severe for high or low values of \( r/n \) as has previously been thought. In fact, information loss decreases markedly if \( r/n \) is close to 0 or 1.

To agree with Pham-Gia and Turkkan, accounting for noise is a significant problem in Bayesian statistics, and will undoubtedly lead to further work that is both theoretically interesting and of great practical use. Amongst
other things, future work will involve generalising to the multinomial case and employing other approaches to measuring information gain and loss due to updating.

ACKNOWLEDGEMENTS

Helpful comments were received from Philip Dawid, Robert Winkler and Brian Davies.

BIBLIOGRAPHY

