*Adv. Appl. Prob.* **15,** 892–893 (1983)
*Printed in N. Ireland*


# THE HAZARDS OF OPTIMAL PROOFREADING

LUKE TIERNEY,* *Carnegie–Mellon University*

**Abstract**

In a recent paper (Yang et al. (1982)) a model for proofreading was proposed in which a reader has a fixed probability $p$ of detecting a misprint in a document containing a Poisson number of errors. This note points out that the conclusions derived from such a model can be extremely misleading if the probability of detecting a misprint varies from one misprint to another.

In a recent paper Yang, Wackerly and Rosalsky (1982) present a model to assist in determining an optimal strategy for proofreading. The model assumes that the number of errors in a text has a Poisson distribution with rate $\lambda$, that error detections are independent and that on each reading each error has the same probability $p$ of being detected. A strategy is then selected to minimize the expectation of a cost function that is linear in the number of proof readings and in the number of undetected errors. The authors treat the case where $\lambda$ and $p$ are known exactly, as well as the case where one or both are unknown but reasonable prior distributions can be specified.

The assumption that detection probabilities remain constant from one reading to the next is probably reasonable for documents of, say, fifty or more pages. However, it does not seem reasonable to assume that all errors are equally easy or hard to detect. For most readers the two typographical errors 'adn' and 'hte' are far easier to recognize than the two errors in '*Zietschrift für Wahrschleinichkeitstheorie*'. Furthermore, most typographical errors of this nature are easier to detect than, say, errors in the cross-referencing of equations.

In view of these observations it is reasonable to ask how sensitive a model assuming a constant $p$ is to departures from this assumption. To shed some light on this question, consider the expected number of undetected errors remaining in the manuscript after $k$ readings. If $p$ is fixed then this number is $\lambda(1-p)^k$. On the other hand, if the $p$'s for different errors are assumed to be independent and identically distributed random variables then the expected number of errors remaining in the text is $\lambda E[(1-p)^k]$. Thus if $\lambda$ and $E[p]$ are assumed known and we act as if $p$ were a constant, then by Jensen's inequality the expression $\lambda(1-E[p])^k$ would strictly underestimate the expected number of remaining errors unless the distribution of $p$ was in fact degenerate.

To assess the magnitude of this underestimate, suppose that $p$ has a beta distribution with parameters $\alpha$ and $\beta$. In Table 1 we consider beta distributions with means of 0·75 and 0·9 and standard deviations of 0·05 and 0·025. The table lists the value of

$$E[(1-p)^k]/(1-E[p])^k = \prod_{i=0}^{k-1} \frac{1+i/\beta}{1+i/(\alpha+\beta)}.$$

Received 6 April 1983.

* Postal address: Department of Statistics, Carnegie–Mellon University, Schenley Park, Pittsburgh, PA 15213, USA.

TABLE 1
Ratio of true to estimated expected remaining errors

| mean | standard deviation | beta parameters $\alpha$ | $\beta$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|---|---|
| 0·75 | 0·050 | 5·4921 | 1·8307 | 1·3605 | 2·2360 | 4·1855 |
| | 0·025 | 8·7507 | 2·9169 | 1·2368 | 1·7798 | 2·8718 |
| 0·90 | 0·050 | 6·5905 | 0·7323 | 2·0814 | 6·0999 | 22·0547 |
| | 0·025 | 10·5008 | 1·1667 | 1·7105 | 3·9631 | 11·2585 |

the ratio of the true expected number of remaining errors to the estimate of that number obtained by assuming a constant $p$, for values of $k$ ranging from 2 to 4. In all cases the estimate for $k=4$ is off by at least a factor of 2·5, and for $E[p]=0·9$ and $\sigma=0·05$, values that seem quite plausible for applications, the estimate is off by a factor of 22!

The beta model for the distribution of $p$ may not be appropriate, in particular for values of $E[p]$ close to 1. However, the magnitudes of the errors computed under this distribution do suggest that a model with constant $p$ should be used with extreme caution. This warning also applies to cases where $\lambda$ or $E[p]$ are not known exactly, since assuming a constant $p$ could lead to serious underestimates of $\lambda$ and overestimates of $E[p]$.

The results of Yang, Wackerly and Rosalsky can be used to obtain an optimal proofreading strategy for a random $p$ if we assume that $\lambda$ and the distribution of $p$ are known. In this case, by their Lemma 1, the fact that $E[p(1-p)^k]$ decreases in $k$ implies that the optimal number of proof readings is the smallest $k$ for which

$$\lambda E[p(1-p)^k] \leqq C_1/C_2,$$

where $C_1$ is the cost of a reading and $C_2$ is the cost of an undetected error. Results for the case where the distribution of $p$ is not completely known are more difficult to obtain since conjugate priors are no longer available.

In practice, if the distribution of $p$ is in fact not known exactly, then it may pay to try to group errors into different categories within which the value of $p$ is nearly constant and to keep track of the number of errors of each type that have been found. The posterior distributions of Yang, Wackerly and Rosalsky could then be computed for each category and could be used to select an optimal proofreading strategy. It might be useful to investigate the tradeoff between the cost of the additional bookkeeping that is required and the reduction in the cost of the optimal strategy.

Finally, it is worth emphasising that the cautionary remarks of this note apply, perhaps *a fortiori*, to other contexts, such as program debugging, where similar models might be considered.

## Reference

YANG, M. C. K., WACKERLY, D. D. AND ROSALSKY, A. (1982) Optimal stopping rules in proofreading. *J. Appl. Prob.* **19,** 723–729.