Acknowledgments. This material is based upon work prepared for the National Science Foundation under Grant No. ENG-79-08139 and ECS-80-21504. This note has benefitted from extensive rewriting by Jon Bentley, one of the referees.

Postscript. The algorithm presented here was discovered independent of, and simultaneous with [2]. The Hirschberg and Sinclair algorithm requires $8N + 8$ CEILING (N lg N) message passes in the worst case. The algorithm presented here requires $3N + 2N$ FLOOR (lg N) message passes in the worst case.

Received 5/79; accepted 9/81

References

1. Chang, E. and Roberts, R. An improved algorithm for decentralized extrema-finding in circular configurations of processes. *Comm. ACM*, *22*, 5 (May 1979), 281–283.
2. Hirschberg, D.S. and Sinclair, J.B. Decentralized extrema-finding in circular configurations of processors. *Comm. ACM*, *23*, 11 (Nov, 1980), 627–628.

Applications:
Management Science and Operations Research
Harvey Greenberg
Editor

Estimating and Improving the Quality of Information in a MIS

Richard C. Morey
Duke University

Most discussions of MIS's assume that the information in the records is error-free although it is recognized that errors exist. These errors occur because of delays in processing times, lengthy correction times, and, overly or insufficiently stringent data edits. In order to enable the user to implement data edits and correction procedures tailored to the degree of accuracy needed, this paper presents functional relationships between three common measures of data quality. The MIS addressed is one where records in a MIS are updated as changes occur to the record, e.g., a manpower planning MIS where the changes may relate to a serviceman's rank or skills. Since each of the updating transactions may contain an error, the transactions are subjected to various screens before the stored records are changed. Some of the transactions including some that are correct, are rejected; these are reviewed manually and corrected as necessary. In the meantime, the record is out of date and in error. Some of the transactions that were not rejected also lead to errors. The result is that at any given time the MIS record may contain errors.

For each of several error control mechanisms, we show how to forecast the level of improvement in the accuracy of the MIS record if these options are implemented.

CR Categories and Subject Descriptors: K.6.4 [**System Management**]—*management audit.*

General Terms: Measurement, Performance, Reliability

Additional Key Words and Phrases: errors, estimation, cost-effectiveness, edits, turnaround times, processing, transactions

The effort was supported by the U.S. Marine Corps under Contract N00014-70-C-0119 with the Office of Naval Research; the author also wishes to acknowledge Thomas C. Varley for drawing the author's attention to this problem, as well as editor's and referees' comments which have substantially improved the exposition.

Author's Present Address: Richard C. Morey, Center for Applied Business Research, Fuqua School of Business Administration, Duke University, Durham, NC 27706

1. Introduction

The importance of examining the impact of information on decision effectiveness in designing and evaluating MIS's has been stressed by several researchers (for example, [1]–[4] and [8].) Implicit in many of these discussions is that the information in the file is error-free and should be used as presented; in reality many potential users of such systems feel the data in the file has not been validated and lacks credibility. They often end up bypassing the MIS in their own decision-making.

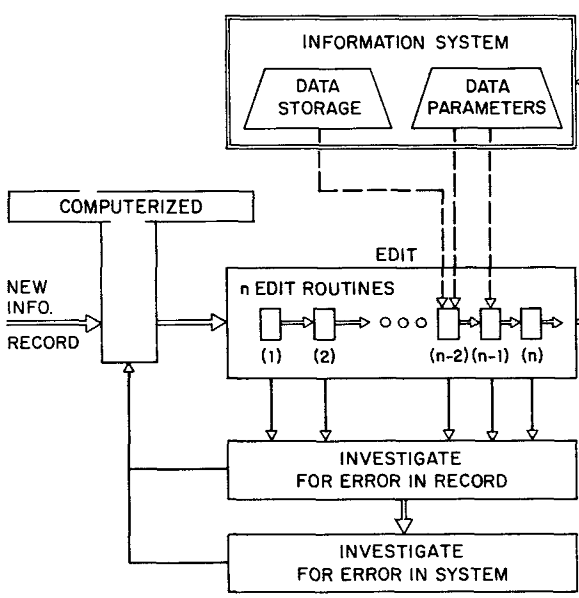
General procedures are missing which would enable one to objectively estimate and control the error rate present in the database or MIS. Since the need for accuracy differs drastically with different types of records (e.g., rank, birth date, Mother's maiden name, etc.), one desires to accomplish the above as a function of the type of record. This information, together with the impacts of using erroneous information and the costs of error reduction, enables one to determine objectively the amount and mix of efforts warranted for each type of record. The same approaches can also be used to help assess the benefits accruing from switching from a batch system to an on-line operation or by reducing various transaction turnaround times.

As an example of the types of mechanisms to be explored, Fig. 1 depicts a series of computerized edits to which all incoming changes to the record are subjected. Some of these edits are completely limited to the entry or transaction itself. This is the case if simple admissibility edits, employing alphanumeric or range checks, are used. Other edits employ so-called logical relationship tests which compare the data contained in the submitted entry with that in the file to discern its "reasonableness." An example of the latter is used in the U.S. Marine Corps' Manpower Management System (MMS) and the Navy's Pay/Personnel Administrative Support System (PASS). These systems provide data on a continuing basis for payroll purposes and to assist in training, assignment, promotion, classification, and separation. Consider a transaction which is to change the rank of a serviceman. Before the stored record can be modified, the rank currently in the record is compared to the one on the transaction; if it is more than a single step, the transaction is automatically rejected by the edits for manual review to discern if the information is correct. (See [8] for other examples of logical edits.)

Beyond the inherent errors in transactions, the time it takes to batch and screen transactions and to manually review rejected transactions strongly impacts the error rates of the information in the system. In addition the relative likelihood of not catching erroneous transactions (i.e., a Type I error, or rejecting transactions that should not be rejected (i.e., a Type II error) also affects the final error rate present in the stored record.

These types of dynamic interactions have been recognized by [2], [3] in their research, but not in the context of an MIS. Such considerations become even more complex since different sets of users within a given organization may have different needs and uses for the same type of information. As an example, personnel managers need timely and responsive answers to their queries, e.g., the number of personnel with certain qualifications. In contrast, the financial managers demand a detailed audit trail and consistency of the historical record. To be able to meet these different needs, one needs the ability to custom tailor the error rate present in the stored record by type of record.

Fig. 1. Diagrammatic Representation of Typical Flow of Data from Updating Transaction.



2. Measures of Error Rates in a MIS

The following discussion applies for a MIS whose records are updated periodically or whenever changes to the record are reported. In the illustrative manpower management systems, the records could be the rank of a given individual or his "leave" status. In other MIS's, the inventory of stock or the usage rates by type of item, or the failure rates or repair costs of certain types of equipment, are examples of types of records.

Within this setting three key measures associated with the accuracy of the information to a given type of record in a MIS are relevant. They are presented below and then discussed in turn. The three measures are:

(*i*) *The transaction reject rate*, i.e., the proportion of incoming transactions which fail, either correctly or incorrectly, the various edits and logical tests used;

(*ii*) *The intrinsic transaction error rate*, i.e., the proportion of the incoming transactions that are truly in error;

(*iii*) *The stored MIS record error rate*, i.e., the probability that the stored record is in error for any reason. It is defined as the likelihood that a randomly chosen record (for the particular record type of interest) examined at a random point in time, is in error. It includes the situation where a change in the record has occurred but has not yet been updated in the record.

The reject rate is a widely used measure and is often used as a proxy for the quality of the incoming transactions. However, the reject rate may exceed the intrinsic transaction error rate if the edits are overly stringent and reject correct transactions. On the other hand the reject rate may significantly underestimate the intrinsic transaction error rate if the edits allow errors to pass through undetected. This might be the case if the MIS relies only on simple admissibility checks but cannot catch errors of substance.

The stored MIS record error rate is, of course, the "bottom line." It is the result of erroneous transactions not being caught by the edits, correct transactions being improperly rejected by the edits, lengthy times between batchings of transactions for initial entry into the computer, and excessive correction and turnaround times. The intrinsic transaction error rate can actually be less than the stored MIS record error rate if the processing and correction times are unduly long or if there is a significant likelihood of Type II errors occurring. If there were no data edits and transactions were allowed to update the stored records instantaneously without any screening, the MIS record would eventually be the same as the intrinsic transaction error rate.

When the above measures were estimated for the leave status record of the Marine Corps' Manpower Management System shortly after its startup, a reject rate of 48 percent was observed for all leave transactions. However, the intrinsic transaction error rate and the MIS error rate for the leave records were later estimated to be 25 percent and 9 percent, respectively. This concretely points out the differences and interactions between these three measures and the need for MIS managers to be aware of their differences when setting accuracy goals and priorities for the MIS's.

3. Statistical Estimates of Error Rates

We address in this section the functional relationships between the three rates and their estimation; a numerical example is also presented to help fix the ideas. The mathematical proofs are in the Appendix.

3.1. Notation

The following notation will facilitate the discussion:

(1) r denotes the earlier defined [(*i*) of Sec. 2] parameter termed the transaction reject rate; it is clearly a function of the edits being used.

(2) P denotes the conditional probability that an erroneous transaction is properly rejected by one of the edits. Hence $1\text{-}P$ is the probability of the Type I error occurring.

(3) P' denotes the conditional probability that a correct transaction is improperly rejected by one of the edits, thereby delaying proper updating of the record. P' is the probability of the Type II error occurring.

(4) e_T denotes the earlier defined [(*ii*) of Sec. 2] intrinsic transaction error rate parameter for the type of record in question.

(5) T denotes the nonnegative random variable representing the time interval or spacing between transactions for a given record of the type being analyzed; these are further assumed to be independent and identically distributed. Candidates for T might be the exponential, uniform, lognormal random variables or even a constant. Also let μ_T denote the mean of the intertransaction times.

(6) C_1, a constant, denotes the minimum processing time, measuring the elapsed time from when a transaction is submitted to the system until it updates the record. This occurs for a transaction not rejected by any of the edits.

(7) C_2, a constant, denotes the additional processing time delay, over and above C_1, to manually review and correct transactions which (*i*) were in error, and (*ii*) were properly rejected by the edits. It will be assumed, to facilitate the exposition, that the manual review perfectly resolves all discrepancies and correctly updates the stored record.

(8) C_3, a constant, denotes the additional processing time over and above C_1, to manually review and allow to enter into the system any intrinsically correct transactions which were rejected by the edits. It is assumed the reviewer is able to ascertain the correct situation so that the stored record is updated accurately.

(9) e_M denotes the earlier defined [(*iii*) of Sec. 2] stored MIS record error rate parameter.

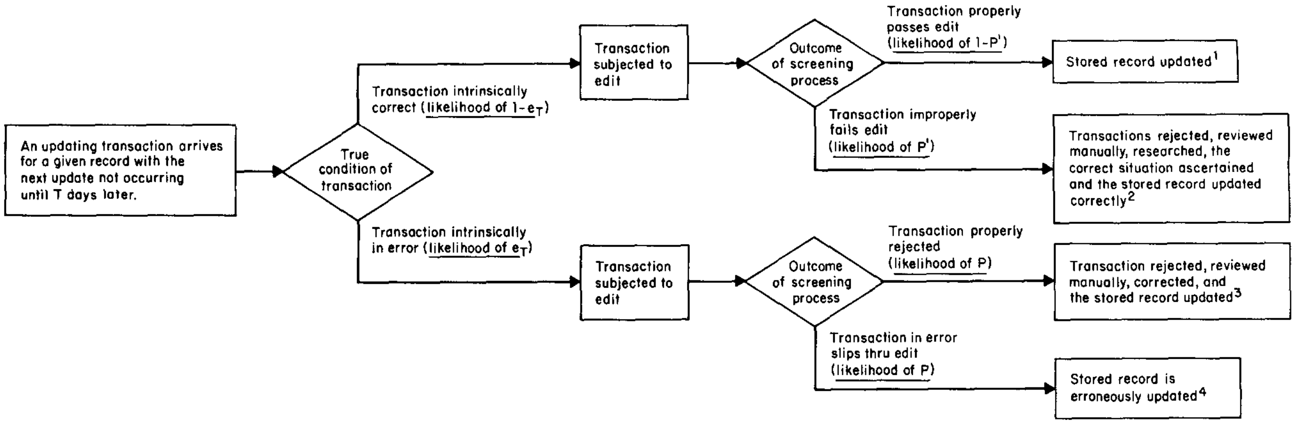
Figure 2 is presented to help the exposition of the various contingencies.

3.2 Estimation of Various Error Rates

This section presents the formulas for calculating the three error rates discussed in Sec. 2 and their interdependencies. They are presented as functions of the various parameters in the previous subsection. The derivations of the formulas have been relegated to the Appendix.

First consider the maximum likelihood estimate of e_T, the intrinsic transaction error rate, given an empirical estimate of the transaction reject rate $\hat{r}$ and the parameters P and P'. (Note that $\hat{r}$ is simply the fraction of transactions in any sample that were rejected by the edit.) The quantity denoted $\hat{e}_T$ is helpful in monitoring the general quality of the incoming transactions. It can also help determine if reporting procedures need to be

Fig. 2. The Various Possible Dispositions for New Transactions.



[1] In this contingency, the fraction of the inter transaction interval $(0, T)$ over which the record is incorrect, is the smaller of C_1/T and 1.

[2] In this case, the fraction of $(0, T)$ over which the record is in error is the smaller of $(C_1 + C_3)/T$ and 1.

[3] In this case the fraction of $(0, T)$ over which the record is in error is the smaller of $(C_1 + C_2)/T$ and 1.

[4] In this case the record is in error for the entire interval $(0, T)$.

tightened or if the edits should be relaxed or tightened.

Assuming $P > P'$ (that is, the edit is more likely to reject erroneous transactions than correct ones), one can show:

$$\hat{e}_T = \begin{cases} 0 & \text{if } \hat{r} < P' \\ \dfrac{\hat{r} - P'}{P - P'} & \text{if } P' \le \hat{r} \le P \\ 1 & \text{if } \hat{r} > P \end{cases} \tag{1}$$

If $P \le P'$, then the edit is not providing a useful function and should be dropped.

Next consider the estimate of e_M, the stored MIS record error rate, as a function of $\hat{e}_T$, the various processing times, i.e., C_1, C_2, and C_3, the likelihoods of Type I and Type II errors, and the intertransaction statistics. If the user has expended the effort to develop a complete probability distribution for the time between transactions (of the type of interest), i.e., a distribution for T, then an exact, albeit somewhat intractable, expression for e_M is available; this is presented in the Appendix [See Eq. (5)]. In those situations where only the mean of the intertransaction times is available, and where a tight lower bound on $\hat{e}_M$ would suffice, the following relationship is useful:

$$\hat{e}_M \ge \hat{e}_T(1 - P) + [C_1(1 - \hat{e}_T)(1 - P') + (C_1 + C_2)\hat{e}_T P + (C_1 + C_3)(1 - \hat{e}_T)P']/\mu_T. \tag{2}$$

In order to apply this result, one should be able to assume that essentially all of the intertransaction times are larger than $C_1 + C_2$ and $C_1 + C_3$, a very reasonable assumption for most practical cases. It is reassuring to note from Eq. (5) that, in the special case where $P = P' = C_1 = C_2 = C_3 = 0$, (that is, no edit is applied and the record is updated instantaneously), $\hat{e}_M$ properly reduces to $\hat{e}_T$. In addition, if T is a constant and this constant is allowed to grow without bound, one obtains $\hat{e}_M$ approaching $\hat{e}_T(1 - P)$, namely, the joint likelihood that the transaction is in error and is not caught by the edit.

One of the real uses for the lower bound of Eq. (2) is in quantitatively assessing the relative improvement in MIS accuracy to be gained from batching transactions more frequently, or from converting to an on-line operation. By varying C_1 in Eqs. (2) or (5), one can compare the changes in $\hat{e}_M$ (or its lower bound) with the cost consequences to determine, in an objective and defensive manner, the relative cost-effectiveness of various proposals. One can even obtain a confidence interval which takes into account the uncertainty associated with the transaction reject rate estimate. This follows since the lower and upper γth level confidence limits for the reject rate parameter r are given by:

$$\hat{r} \pm k_\gamma \sqrt{\frac{\hat{r}(1 - \hat{r})}{N}} \tag{3}$$

In Eq. (3), N is the size of the sample on which the reject rate is being estimated, and k_γ satisfies $1 - \gamma = 2(1 - \Phi(k_\gamma))$, $\Phi(x)$ being the cumulative distribution function for the standard normal. By substituting the upper and lower confidence limits from Eq. (3) for $\hat{r}$, one has confidence limits on e_T, and hence, from Eq. (5) or (2), on e_M or its lower bound.

3.3 Application

The following numerical illustration is given using actual data associated with the aforementioned Marine Corps Manpower Management System shortly after its startup. The transaction type of interest was the leave status for a typical enlisted Marine. The edits were actually a series of simple admissibility and logical edits which together yielded estimates of P and P' of 0.95 and 0.03, respectively. These estimates of P and P' were based on a detailed analysis of a sample of transactions which had been carefully researched to reveal the true situation for each transaction in the sample. The various processing times were estimated to be as follows: $C_1 = 3$

days, $C_2 = 2$ days, and $C_3 = 4$ days. The mean time between leave transactions for a given Marine was 43 days; this was based on the total number of leave transactions for all enlisted Marines in the past year and the average number of active duty Marines during that period.

Based upon the results of all leave transactions contained in a randomly drawn processing cycle, the empirical estimate of the transaction reject rate parameter was 0.26. The 90 percent confidence limits on r were 0.26 ± 0.0272. From Eq. (1), we obtain the estimate of e_T to be:

$$\hat{e}_T = \frac{\hat{r} - P}{P - P'} = \frac{0.26 - 0.03}{0.95 - 0.03} = 0.25.$$

By substituting in the upper and lower confidence limits, at the 90 percent level, for r into Eq. (1), we also obtain (0.22, 0.28) as the 90 percent confidence interval on e_T. The next step was to lower bound the stored MIS error rate for this type of transaction. Before Eq. (2) could be used for that purpose, a review of intertransaction times for randomly selected Marines was conducted. This was to ascertain that, for all practical purposes, the time between transactions exceeded a week, i.e., the larger of $C_1 + C_2$ and $C_1 + C_3$. This having been accomplished, Eq. (2) was applied yielding a lower bound on $\hat{e}_M$ of 9.45 percent. In order to determine how close the lower bound of $\hat{e}_M$ was to its exact estimate, a detailed review of intertransaction times yielded the probability distribution for T to be well approximated by a uniform distribution between 14 and 72 days and a mean of 43 days. Upon substitution of this probability distribution into Eq. (5), one obtains $\hat{e}_M = 11.21$ percent, so that the lower bound of 9.45 percent was reasonably tight. The analysis enabled the MIS manager to estimate objectively that his edits, processing times, and Type I and II errors all combined to take an intrinsic transaction error rate of 25 percent and yield a stored MIS record error rate of 11.21 percent; this represented an improvement of about 56 percent in the quality of information.

4.1. Analysis of Marginal Effectiveness of Various MIS Error Rate Reduction Mechanisms

Consider the use of the functional form, available in Eq. (5), for gauging the reductions possible in the stored MIS error rate. The various control mechanisms available to the MIS manager include:

(*i*) Improving the quality of the incoming transactions, i.e., reducing e_T; presumably this could be accomplished by more training of the preparers of the transactions, use of optical character recognition (OCR) equipment, more emphasis on the care exercised in preparing transactions, etc.

(*ii*) Reducing the cycle time for processing of transactions, i.e., reducing C_1; this could be accomplished by batching more frequently or use of an on-line operation.

(*iii*) Reduction of the time required for manual review, researching, and correction of rejected transactions, i.e., reduction of C_2 and C_3. This depends upon more and/or better trained clerks, as well as improving their access to the historical records or individuals involved. This is exactly the thrust of the Navy's new PASS system mentioned earlier where there is to be one single location for each Navy person, handling all payroll, re-enlistment, separation, vacations, etc. issues. This improved interface will facilitate the researching and correction of rejected transactions.

(*iv*) A tightening of the edits to reduce the frequency of Type I and Type II errors, i.e., to increase P and and reduce P'. This requires more precision in the screens used and requires a careful analysis of the relative advantages and disadvantages of deleting or adding edits.

To investigate which of the above mechanisms has the most impact, the functional form of Eq. (5) in the Appendix can straightforwardly be differentiated to yield the rate of change of $\hat{e}_M$ with respect to P, P', $\hat{e}_T$, C_1, C_2 and C_3. As an illustration, consider the scenario discussed in the previous section. Upon differentiating Eq. (5) and evaluating it at the stated levels, one finds:

$$\left.\frac{\partial \hat{e}_M}{\partial P}\right|_{\substack{\hat{e}_T = 0.25,\ C_1 = 3,\ C_2 = 2,\\ C_3 = 4,\ P = 0.95,\ P' = 0.03}} = -0.215$$

Hence, in a small neighborhood about $P = 0.95$, an increase in P of one percentage point (by tightening the edit) is likely to lower the stored MIS error rate by 0.215 of one percentage point. In a similar manner one can compute, in a small neighborhood about the current values, a marginal percentage reduction in $\hat{e}_M$ per percentage reduction in P' of 0.084, and a marginal percentage reduction in $\hat{e}_M$ per percentage reduction in e_T of 0.097. Upon carrying out a similar exercise for C_1, C_2, or C_3, one can estimate that a one day reduction in C_1, C_2, or C_3 reduces $\hat{e}_M$ by 2.79, 0.67, and 0.06 percent, respectively. The MIS manager, armed with these respective reductions and with respective costs to implement each of the error control mechanisms, can objectively and defensibly determine the best method of attack to reach some given accuracy goals.

5. Summary

An assessment of the quality of information stored in a MIS database where records are updated as changes occur to them is considered. It is assumed the updating transaction is subjected to edits before it is allowed to modify the record. The key result is a mathematical relationship for the stored MIS record error rate as a function of the quality of the incoming data, the various processing times, the likelihoods of Type I and II errors, and the probability distribution for the intertransaction time. This yields an easy-to-compute lower bound on the stored MIS error rate and enables one to compute the impacts on the MIS error rate of marginal changes in the above error control mechanisms. Also important are the interrelationships developed between other standard measures of data quality. Finally, the method is useful for estimating in a quantitative fashion the relative cost-effectiveness of various error reduction strategies including converting from a batch processing operation to an on-line operation.

Appendix. Derivation of Estimation Formulas

The proof of Eq. (1) follows from the realization that

$$r = e_T P + (1 - e_T) P', \tag{4}$$

r being a convex combination of P and P'. For realizations of r outside of (P', P), the best estimate of e_T is 0 or 1. Consider the derivation of Eq. (2). We assume that the successive transactions are independent in terms of their likelihood of being in error. It is recognized that this may be somewhat of a departure from reality but is plausible if the sources of the error are more from keypunching or human errors.

Let X, a random variable denote the length of time the file is incorrect in the period T (itself a random variable) between individual transactions. Then the random variable X/T represents the fraction of the interval $(0, T)$ that a record, chosen at random and examined at a random point in time, is in error. Consider the random variable $X/T \mid T = t$, using the notation that $a \Lambda b$ denotes the smaller of a and b. One has:

$(X/T \mid T = t =)$

1	with probability $e_T(1 - P)$, i.e., the transaction contains errors but slips through undetected.
$\dfrac{C_1 \Lambda t}{t}$	with probability $(1 - e_T)(1 - P')$, i.e., the errorless transaction properly passes the edits and updates the file correctly.
$\dfrac{(C_1 + C_2) \Lambda t}{t}$	with probability $e_T \times P$, i.e., the erroneous transaction properly is failed by the edit and after manual review, is allowed to update the record correctly.
$\dfrac{(C_1 + C_3) \Lambda t}{t}$	with probability $(1 - e_T) P'$, i.e., an errorless transaction improperly fails the edits, but after manual review is allowed to update the record correctly.

Upon computing the expectation of $X/T \mid T = t$ and taking the expectation of this with respect to T, one obtains the result:

$$\begin{aligned} E_T\left[E\left(\frac{X}{T} \,\middle|\, T\right)\right] &= e_M \\ &= e_T(1 - P) + (1 - e_T)(1 - P') \\ &\quad \times \left[C_1 F_T(C_1) + \int_{C_1}^{\infty} C_1/t \, dF_T(t)\right] \\ &\quad + e_T P'\left[(C_1 + C_2) F_T(C_1 + C_2) \right. \\ &\quad \left. + \int_{C_1 + C_2}^{\infty} (C_1 + C_2)/t \, dF_T(t)\right] \\ &\quad + (1 - e_T) P'\left[(C_1 + C_3) F_T(C_1 + C_3) \right. \\ &\quad \left. + \int_{C_1 + C_3}^{\infty} (C_1 + C_3)/t \, dF_T(t)\right] \end{aligned} \tag{5}$$

where $F_T(t)$ is the cumulative distribution function for the random variable T. The result Eq. (2) of Sec. 3 is then a direct corollary from Eq. (5) where Jensen's Inequality (for example, see [5]) is used to obtain $E_T(1/T) \geq 1/E_T(T) = 1/\mu_T$. Finally, the numerical application is based on $F_T(t)$ being a uniform distribution over the interval (a, b). In this case the computation of Eq. (5) is facilitated by realizing that:

$$\int_k^{\infty} \frac{k}{t} \, dF_T(t) = \int_a^b \frac{k}{t(b - a)} \, dt = \frac{k}{b - a} \ln \frac{b}{a}$$

whenever $k < a$.

Received 2/81; revised 6/81; accepted 8/81

References

1. Chervany, N.L. and Dirkson, G.W. Economic evaluation of management information systems: An analytical framework. *Decision Sciences, 1,* 1970, 96–308.
2. Forrester, J. W. *Principles of Systems.* Wright-Allen, Cambridge, Mass., 1976.
3. Greenberg, Harold. Optimal test procedures under stress. *Operations Research 12,* 5, (Sept–Oct 1964), 689–692.
4. Haber, S. E., Segel, F., and Solomon, H. Statistical auditing of large scale management information systems. *Naval Res. Logistics Quarterly. 19,* 3, (Sept. 1972), 449–459.
5. Harris, B. *Theory of Probability.* Addison-Wesley, Reading, MA, 1966.
6. Morey, R. C. and Iglehart, D. L. Inventory systems with imperfect asset information. *Management Science, 18,* 8, (April 1972) 388–394.
7. Narasimhan, R. A simulation analysis of timeliness and accuracy of information on system performance. *Proceedings of the 10th Annual Conference, American Inst. for Decision Sciences,* Vol. 1 November 1978, 143–145.
8. Varley, Thomas C. Data input error detection and correction procedures. Office of Naval Research Report, June 1969, (unpublished Ph.D. dissertation, George Washington University.) 25–27.