

AD-776 538

STATISTICAL APPROACHES TO MANAGING
MANPOWER DATA BASES

D. P. Gaver, et al

Naval Postgraduate School
Monterey, California

February 1974

DISTRIBUTED BY:

NTIS

**National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151**

NPS-55Gv55Lw74021

AD 776538

NAVAL POSTGRADUATE SCHOOL

Monterey, California



1974
RECEIVED

STATISTICAL APPROACHES
TO MANAGING MANPOWER DATA BASES

by

D. P. Gaver

and

P. A. W. Lewis

February 1974

Approved for public release; distribution unlimited.

Prepared for:
Office of Naval Research, Arlington, Virginia 22217

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral Mason Freeman
Superintendent

Jack R. Borsting
Provost

The work reported herein was supported in part by the Office of
Naval Research. (All Volunteer Armed Forces Committee.)

Reproduction of all or part of this report is authorized.

This report was prepared by:

Donald P. Gaver
Donald P. Gaver
Professor of Operations Research

Peter A. W. Lewis
Peter A. W. Lewis
Professor of Operations Research

Reviewed by:

Carl R. Jones
Carl R. Jones
Department of Operations Research
and Administrative Sciences

Released by:

John M. Wozencraft
John M. Wozencraft
Dean of Research

PROVISION for

PTIG	Basic Section	<input checked="" type="checkbox"/>
	Basic Section	<input type="checkbox"/>
		<input type="checkbox"/>

INT. ORDER

W. L. EDGAR

A

STATISTICAL APPROACHES TO MANAGING MANPOWER DATA BASES

D. P. Gaver

P. A. W. Lewis

Naval Postgraduate School
Monterey, California

1. Introduction.

In a recent report MARSHALL [6] has described approximately fourteen manpower and personnel data bases that are currently maintained by the Navy and Marine Corps. Such data bases are of use for operational or managerial purposes; for example, the enlisted master tape at Pers N purports to show the current status of each enlisted man in the Navy. They are also potentially useful for staff studies which usually address some specific issue or question on a once-only basis, and for research on manpower problems. Such research is of value in developing a better understanding of manpower behavior, and thus for improving policy making in the manpower area.

However, an operational data base, such as the enlisted master file, is unlikely to be satisfactory to manpower researchers. Firstly, it represents a snapshot of close to the entire enlisted Navy at a single point in time. Previous snapshots are obtained only with some difficulty, so it is hard to calculate rates of transition, averages or trends over time, etc.--all of the measures that are useful to manpower researchers. Secondly, a real understanding of the data base is important in order not to make systematic misinterpretations and errors. MARSHALL, [6, page 32], discusses some of the possible misinterpretations (see the discussion of "loss holds"). Also, one has the lurking feeling that it is exceedingly easy for error to creep into and remain in the data base, and that these

errors may lead to incorrect impressions of what is actually occurring, and possibly to unjustified policies.

The purpose of this report is to discuss statistical procedures and viewpoints that may be of use in data base or data bank management. We will try to show how recent statistical research is aimed at the development of methods useful when various amounts of data are potentially or actually available.

Although this report devotes principal attention to manpower data bases, we recognize that other data base problems may be approached in a similar manner. In particular, problems associated with the MMM data base have suggested some of the lines of thought developed here.

The reader may notice that we do not discuss problems of accessibility of data, either in the sense of physical (computer) accessibility or of inadequate documentation. This is primarily a computer science problem, but it does relate ultimately to whether or not data in a data bank will be used by researchers.

2. Comments on Data Base Types.

A manpower data base is typically organized by identifying an individual, e.g. by social security number. Under this identification or key, then, there is a record of the individual's status as of a particular point in time. The status may be represented as a vector with alpha-numeric components. Periodically, the status vector for each individual is updated. It is at the moments of status vector introductions and update that errors may be introduced into the system.

2.1 Operational vs. Research Data Bases.

Existing data bases, e.g. those at BuPers, are constructed for operational purposes. They must be nearly up-to-date, and include all individuals actually in the service (or in lost-hold status). Elements of the status vector must be sufficient for operational purposes, e.g. must show broad capabilities; pay rate, rating, specialty, perhaps educational level. One primary purpose of such a base is to exhibit present job assignment, and to keep a record of each individual's plausible future assignments. For such a base an error may mean miscasting an individual in a new assignment, at least temporarily. Such errors are likely to be of less importance in a research data base, from which it may be hoped to draw general conclusions about characteristics of groups of individuals.

In order to construct a data base for research purposes one must have considerably more detail than is usual in an operational data base. Some such detail may be provided by accumulating over time the snapshots alluded to previously: one can then, for example, compute estimates of the continuation rates of various classes of individuals, e.g. electronic technicians, and thereby devise policies directed at building retention in needed areas. This sort of problem is being investigated by K. T. MARSHALL. Other investigations, e.g. those of drug abuse, crime, suitability for certain educational programs and jobs, require more detailed and personal information.

2.2 Samples.

For research purposes there should be no need to have in-depth information on all members of a population of interest. A properly constructed and well maintained random or stratified statistical sample is adequate for forming hypotheses concerning associations, and even the effect of policies, such as lump sum payments to induce recruitment or aid retention. Such a sample might be analogous to the test panels used in consumer goods marketing research, although the members of the sample need not be aware of their special status. If a sample is constructed, a suitable compromise can be made between original selection cost and the cost of maintenance, as compared to the cost of sampling error. It will be possible to maintain in essentially error-free condition an in-depth sample of a fraction of the individuals in a target population much more cheaply than is possible with a complete roster of all individuals in the population. The latter cost is difficult to establish for a research data base, for potential usage is unlikely to be well-known or specifiable in advance. One can imagine

collecting a multi-dimensional profile of selected individuals, only to find that certain aspects of this profile are of no interest to researchers or problem solvers. Provision for gradually clearing the inventory of such detail should be made whenever data bases are constructed. The problem of deciding when and how much to reduce the magnitude of stored data may be approached by way of statistical decision theory, combined with informed technical judgment. It bears some relation to problems of control of physical items that are kept in inventory. The control of the errors in the data that is inventoried is analogous to problems encountered in the quality control of manufactured items. In later sections of this report we address problems of this sort.

There are also problems of time sampling which are comparable to so-called problems of "work sampling" which have been investigated by industrial engineers and statisticians. I.e., how often should one sample a data base to obtain an adequate idea of the dynamics of the system. In particular it is important to avoid aliasing effects, and bias effects from time sampling such as the well known effect of length-biased sampling [3,4].

3. Statistical Methods and Data Bases.

In this section we will discuss some specific statistical methods likely to be useful in (a) summarizing data base information, and (b) checking data inputs for errors. In fact, (a) and (b) are related activities.

3.1 Data Summarization by Quantiles.

Data bases typically contain entries pertaining to a great many individuals. For example, studies of retention might benefit from data on the distribution of time in active service for individuals of different backgrounds, and studies on promotion and advancement would require the distribution of time spent at each rate or rank level by individuals in different categories. Since the data are overwhelmingly numerous, summaries which adequately characterize the data are certainly a necessity.

Obvious, and traditional, summaries are the following:

(a) The sample mean of the data:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (3.1)$$

where x_i represents the i^{th} observation on the variate X , and n is the total number of these observations. Besides being a numerical summary. This estimates the mean value $\mu = \int_{-\infty}^{\infty} xF(x)dx$, where $F(x)$ is the p.d.f. associated with x .

(b) The sample standard deviation

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (3.2)$$

and other moments. The mean measures general "location" or "magnitude" of the data, while standard deviation indicates the dispersion. Robust summaries of location, cf. ANDREWS, et al. [1] are likely to be useful here.

One is frequently interested in looking at data with reference to a mathematical model. That is, a stochastic formulation often yields or requires data that may be regarded as a collection of observations on a random variable X . While the mean and variance of observations are useful, details concerning probabilities or distributions are even more so for focussing attention on the departures that may occur between model input or output, and actual data. Lately, work has been done by LEWIS, GOODMAN, and ROBBINS [5], and also in the thesis by YAGUCHI [11] written under the direction of P. A. W. LEWIS on the estimation of distributional quantiles from large samples. We define the α -quantile of the distribution $F(x)_\alpha$ of the random variable X to the number, X_α , that satisfies

$$F(X_\alpha) = \alpha \quad (3.3)$$

for α some number between zero and unity. LEWIS, et al, consider efficient computational methods for estimating such quantiles from vast amounts of data typical of certain data base problems. This work is specifically aimed at computational efficiency, and attempts to avoid sorting of extensive data, which can be time consuming and costly in terms of computer memory. It is based on a data transformation and the use of stochastic approximation methods and, in summary, is as follows.

It is assumed that the data is a homogeneous sequence of random variables X_1, X_2, \dots with continuous distribution $F(x)$, the x_i 's being, for example, length of service of Naval Personnel at a given instant of time.

- (1) The procedure for estimating quantiles must be fast, on-line and economical of storage. This precludes the standard estimates based on order statistics, except for small data sets.

- (2) The alternative is to estimate the quantiles by using stochastic approximation (Robbins-Monro) techniques. These require very little computer storage. However, as one gets further away from the median ($\alpha = 0.5$) the Robbins-Monro technique has been shown to be a very poor estimator in the sense that convergence is extremely slow.
- (3) Extensive simulations (YAGUCHI, [11] have enabled us to understand the problem with the Robbins-Monro technique; this understanding will allow us to use jackknifing techniques (GRAY and SCHUCANY, 1972) to improve the technique if necessary.
- (4) An even greater improvement in the performance of the R.M. technique for extreme quantiles is obtained by using the maximum transformation (GOODMAN, LEWIS and ROBBINS [5]) for $\alpha > 1/2$, and the minimum transformation for $\alpha < 1/2$. The maximum X_1 in each successive group of v X_1 's is obtained and its median value (equal to the α quantile of X_2) is estimated using the Robbins-Monro technique. The group size v is chosen so that the median of the distribution of the maximum equals the desired α -quantile. For example for $\alpha = 0.975$, $v = 28$.
- (5) The procedure using the maximum transformation is unfortunately very sensitive to outliers (errors in the data; we only discuss those in the positive direction). A possible solution is to use the next-to-maximum or next-to-next-to-maximum transformation. Other robust procedures are worthy of investigation.
- (6) For a given v which transforms the α -quantile of $F(x)$ to the $\alpha' = 0.50$ quantile of the maximum of v X_1 's, the level α'' of this quantile, is generally higher for the next-to-maximum, e.g. $\alpha'' = 0.85$. However, we can now use the Robbins-Monro technique to estimate the

α -quantile (equivalently the α'' -quantile of the next to maximum distribution), even though α'' is so high.

3.2 Other Data Summaries.

Suppose that either the basic data is not very numerous, as will often be true when a new base is being established, or that the data is segmented because of a suspected change in the underlying process. Then one can estimate quantiles more straightforwardly. An obvious approach is by ordering the data and using order statistics, e.g. in a sample of 100, let $x_{(90)}$ be the 90th observation in order of increasing size. One can let $x_{(90)}$ estimate quantile $\frac{90}{101} \approx 0.9$, which it does nearly unbiasedly.

Perhaps a more informative approach is the following: Compute the mean, \bar{x} , and standard deviation, s , of either the complete body of data, or a sample thereof. Then simply tabulate the data that lies in intervals of the form

$$(\bar{x}, \bar{x}+s), (\bar{x}+s, \bar{x}+2s), (\bar{x}+2s, \bar{x}+3s), \dots, \\ (\bar{x}, \bar{x}-s), (\bar{x}-s, \bar{x}-2s) \quad (3.4)$$

This kind of sorting procedure can be replaced by intervals of the form

$$(\bar{x}+khs, \bar{x}+(k+1)hs) \quad (3.5)$$

where h is a multiplier of the scale factor s , and $k = 0, \pm 1, \pm 2, \dots$ to obtain a finer-scaled determination. The tabulation referred to can be (i) a summary count (ii) a count, plus a summary of the data point locations in the interval (\bar{x}_k is the mean of the d.p. in the k^{th}

interval, s_k the sample variance, $\min x_k$ the smallest, etc.), (iii) a count, plus a display--perhaps on cathode-ray scope face--of the individual observation in interval k --all for all, or selected, k . Among other possibilities are the stem-leaf representation of TUKEY [8].

One can consider extending the above procedure to several dimensions, developing cells of dimension s_x by s_y , for example. Graphical display is still possible, but becomes more difficult if three or more dimensions are desired.

The above procedures are difficult to implement for every skewed data and would perhaps have to be used in conjunction with the quantile estimation methods. Alternatively, the data should be transformed (e.g. logged, cube-rooted) to reduce skewness and bring nearer to normal or Gaussian form. Thus a histogram with quantiles indicated can be useful in examining very skewed data. A package like this has been developed at the Naval Postgraduate School by D. W. ROBINSON.

4. Data Bases as Storage Systems.

Any data base may be regarded as an inventory, into which items of information are placed, and from which demands are satisfied. Consequently, one is tempted to apply ideas of inventory and storage theory to the data base management problem. We mention some possibly fruitful viewpoints for data base systems design and management.

4.1 The Data Base as an Inventory of Errors: Estimation of Error Content.

Each element of a data base--entry in an individual's status vector--is subject to occasional change, and such changes often introduce errors. In fact, errors are introduced along with changes, e.g. when incorrect entries are made, either by the originator or by the data base custodian. Of course, once introduced, errors remain present in the data base if actual changes are not made. Also, unless checks are applied, some actual changes will never be recorded. The data base will thus not be topical.

How can a data base be checked for error content? A model for error introductions and residence times in an operating data base was given by CARTER [2] in his Naval Postgraduate School M.S. thesis, written under the direction of K. MARSHALL. In this thesis, it was presumed that errors enter (along with changes) at Poisson rate λ , and survive for a random time which is essentially the time until the next change. These assumptions lead to an $M/G/\infty$ queueing or service system model, from which one can predict the distribution of the number of errors resident in the system. According to this model, $N(t)$, the number of errors present at t is approximately Poisson with parameter close to $\frac{\lambda}{\mu}$, μ^{-1} being the expected residence time of an error. The model assumes that errors are removed individually, by attrition, when changes are made in elements.

Now one can estimate the incidence of errors by sampling. Suppose that periodically a random sample of n elements is selected from the data base and these elements are checked for accuracy by referring to the data source. If there are D elements in the base at the time of sampling then the conditional probability of sampling an erroneous element--an error, for short--is, given $N(t)$, and assuming that all elements are equally likely,

$$P\{\text{error is sampled element} | N(t)\} = \frac{N(t)}{D}. \quad (4.1)$$

The unconditional probability is simply

$$P\{\text{error in sample}\} = \frac{E[N(t)]}{D} = \frac{\lambda}{\mu D} \quad (4.2)$$

Under our assumptions, we can now estimate λ , for D is known and μ^{-1} is the expected residence time of data base elements between changes, a quantity that is possible to estimate from the rate at which change orders appear at the base. If samplings are taken infrequently enough so that the error content stochastic process, $N(t)$, is essentially in stochastic equilibrium, then if k errors are found in the sample of n we can estimate λ by

$$\hat{\lambda} = \left(\frac{k}{n}\right)\mu D. \quad (4.2)$$

If elements sampled are independent then

$$\text{Var}[\hat{\lambda}] = \text{Var}\left[\frac{k}{n}\right]\mu^2 D^2 = \frac{\lambda}{\mu D} \left(1 - \frac{\lambda}{\mu D}\right) \frac{1}{n} \mu^2 D^2 = \frac{\lambda \mu D}{n}. \quad (4.3)$$

This does not take account of errors in estimating μ .

Numerical Example.

Suppose $D = 3 \times 10^7$ separate data elements (see CARTER, p. 9).

Let the probability of a change on a day be $\mu = 3 \times 10^{-5}$. Suppose that

$n = 1000$, and the fraction of changes that are in error is $p = 1 \times 10^{-3}$. The rate of occurrence of changes is 1000, so $\lambda = 1000 \times 1 \times 10^{-3} = 1$. Therefore,

$$\text{Var}[\hat{\lambda}] \approx \frac{1 \times 3 \times 10^{-5} \times 3 \times 10^7}{1000} = \frac{9}{10}. \quad (4.4)$$

Clearly a sample of $n = 1000$ is not really sufficient to estimate this relatively small error introduction rate. Yet the expected number of errors in the base is, on the average, equal to $\frac{\lambda}{\mu} = \frac{1}{3 \times 10^{-5}} \approx 3 \times 10^4$.

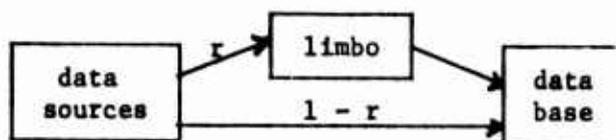
Actually, it will probably be most convenient to extract a sample of individuals, not elements. Using Carter's example, suppose that a sample consists of 600 men out of 600,000; this corresponds to $n = 600 \times 50 = 30,000$. Then

$$\text{Var}[\hat{\lambda}] \approx \frac{1 \times 3 \times 10^{-5}}{3 \times 10^4} = 6 \times 10^{-4}.$$

It is probably unnecessary to determine λ to the above accuracy, so a smaller sample should suffice. Sampling data sources is also sensible.

4.2 Preventive Sampling

The above approach is devoted to estimation of data base error content, and not directly to error prevention or reduction. A sampling procedure can also be used to screen data input. For example, one can select certain incoming change notices and carefully verify them, correcting those found to be in error. We can put those items in the process of being checked into a "limbo" status, later admitting them to complete status when checking and correction is finished. The path of data may be depicted as follows:



Let δ be the rate at which data base changes occur. With probability r --note that r is a decision variable--a change is first subjected to check and verification before being entered into the data base, while with probability $1 - r$ the data is entered directly into the base. Any element that is in process of check is in the limbo state, and is delayed for an average time v^{-1} until it is verified and sent to the data base. Since change arrivals occur at Poisson rate δ , then those into limbo occur at rate $r\delta$, of which a fraction p are in error upon entry. Assume for simplicity that none are in error upon final introduction to the data base, of which a fraction p are in error. Therefore, the rate of input of errors to the data base is $(1-r)p\delta = (1-r)\lambda$, and consequently the expected number of errors resident in the data base is reduced to $(1-r)\frac{\lambda}{\mu}$, at the expense of maintaining a limbo base of average size $\frac{r\delta}{v}$, of which a fraction p are in error. We remark that $L(t)$, the total number of elements in the limbo file is, under our assumptions, Poisson distributed, and independent of the number of errors present in the data base.

The above notion of limbo file can be generalized in various sensible directions. One is to adjust the limbo input to the error proneness of each individual data source, or class of sources. If $r_i(t)$ is the fraction of items selected from source i at time t (i.e. during month t) then if sampling indicates that $p_i(t)$, the probability of an error from source i , is on the increase, $r_i(t+1)$ can be increased,

perhaps to unity. Another, related, approach is to subject each change to a consistency check before admitting it to the data base; this approach has been discussed by NAUS [8].

None of the approaches for error control mentioned above directly concern themselves with usage or frequency of attempted access to the data base. If the data base is seldom utilized, then its continued existence will be questioned. If, upon attempted access, many data base elements are missing or in error the data base will certainly lose credibility and fall into disuse. Therefore, a sampling scheme based upon usage may well be considered. A simple version may run as follows: If an element is accessed, subject it to check before transmission with probability $c(0 < c < 1)$. Note that such a check requires a non-zero time that may be unacceptable to the user. In that case, the required data may be transferred, but without a "Grade A" rating. The data base then retrospectively samples the elements transmitted and corrects those found to be in error. If the latter fraction is high, further cleanup is carried out. Such a procedure tends to keep those elements frequently accessed by users in relatively low error condition, and tends to neglect elements that are infrequently accessed. An extension is to attempt to classify the attempted accesses by users, and by sensitivity of the user to errors. Those elements most effected can then be maintained most vigorously. One can even selectively cease maintaining certain elements in the base--or add others--depending upon experienced or projected demand. The existence of a satisfactory base depends upon effective interaction between the prospective users and those responsible for managing the base itself; such an interaction should not be time consuming or tedious if it is aided by

use of some simple decision rules for data management. Those suggested will perhaps stimulate other ideas.

4.3 Automatic Error Selection and Correction.

Schemes for automatically selecting data which might be in error for the limbo file are being investigated. In effect, this uses the quantile estimation scheme detailed above, but now the next-to-maximum in a set of v of the x_i 's is used to estimate an α -quantile so that a judgment can be made as to whether to put the maximum of the x_i 's in the limbo file.

This work is still very tentative. One implementation successively takes k groups of size v , and estimates the α -quantile S_α of the x_i 's with the next-to-maxima. Then S_α is the median point of the distribution of the maxima of the groups of size k . If all k maxima are greater than the estimated S_α , these maxima are put in the limbo file for examination, possibly being first ordered.

In the above, v and k are chosen to control various error probabilities. There are many possible variants on the basic procedure which are sensible and will be explored by means of simulation experiments.

5. Transient Behavior of a Data Base: An Approximate Analysis.

In this section we make an explicit mathematical model for a data base, viewed as a storage system of records or data with time-varying content. Such a model allows one to estimate possible costs of maintaining an actual base, where the cost components are computer storage requirements, change processing, and user response. We assume that each individual in the system is represented by a status vector in the data base. The implication is that we are dealing with an operating data base. However, if a sample is selected our model may as well represent the behavior of a smaller, research, data base.

5.1 Formulation.

Let $I(t)$ denote the number of individuals represented at time t in the data base. If $I(t)$ represents even a subpopulation of individuals, e.g. all medical doctors in the U.S. Navy, this population will change in time, with changes that tend to be related to population size. Let $a(t, I(t))dt$ represent the probability that a new individual is acquired or recruited at time t , and $b(t, I(t))dt$ represents the probability that an individual leaves the system. Consequently, these represent probabilities of change in the data base itself.

5.2 Illustrative Model.

Consider the following simple model. Suppose \bar{I} is an upper limit for the size of the organization, $\alpha(t)$ is recruitment or enlistment rate, and $\beta(t)$ represents the time-dependent rate at which individuals leave the organization. Then the change in organization size, $dI(t)$, can be considered to consist of a drift or deterministic part, plus a random disturbance.

It is plausible that recruitment effort (or expense) should be proportional to $\bar{I} - I(t)$ for $0 < I(t) \leq \bar{I}$. Thus the expected number of recruits or addition is represented by $\alpha(t)\{\bar{I} - I(t)\}dt$ in time dt . Likewise, the expected number of defections is approximately $\beta(t)I(t)dt$. We write the stochastic differential equation

$$dI(t) = \alpha(t)\{\bar{I} - I(t)\}dt - \beta(t)I(t)dt + \sigma(t)dW(t), \quad (5.1)$$

where $dW(t)$ is the differential of a Wiener process; see COX and MILLER, p. 217. If individuals tend to behave independently and in such a way that the point process of changes is orderly then the variance of the change in dt is representable as

$$\text{Var}[dI(t)] = \sigma^2(t)dt = [\alpha(t)\{\bar{I} - I(t)\} + \beta(t)I(t)]dt. \quad (5.2)$$

Now suppose, following McNEILL [7] that $I(t)$ consists of a deterministic part, $x(t)$, plus a stochastic noise part, $S(t)$:

$$s(t) = \frac{I(t) - \bar{I}x(t)}{\sqrt{\bar{I}}} \quad (5.3)$$

and consider what may occur if \bar{I} becomes large ($\bar{I} \rightarrow \infty$). Substituting $dI(t) = \bar{I}dx(t) + \sqrt{\bar{I}} dS(t)$ into (5.1) we find

$$dI(t) \equiv \bar{I}dx(t) + \sqrt{\bar{I}} dS(t) = \alpha(t)\{\bar{I} - \bar{I}x(t) - \sqrt{\bar{I}} S(t)\}dt - \beta(t)\{\bar{I}x(t) + \sqrt{\bar{I}} S(t)\}dt$$

$$\sqrt{\alpha(t)\{\bar{I} - \bar{I}x(t) - \sqrt{\bar{I}} S(t)\} + \beta(t)\{\bar{I}x(t) + \sqrt{\bar{I}} S(t)\}} dW(t) \quad (5.4)$$

Next divide through by $\sqrt{\bar{I}}$, and allow $\bar{I} \rightarrow \infty$. Identification of terms of order $\sqrt{\bar{I}}$ gives a differential equation for $x(t)$, the deterministic component:

$$\frac{dx(t)}{dt} = \alpha(t) - [\alpha(t) + \beta(t)]x(t). \quad (5.5)$$

The solution is

$$x(t) = x(0)e^{-\int_0^t [\alpha(z) + \beta(z)] dz} + \int_0^t \alpha(y) e^{-\int_y^t [\alpha(z) + \beta(z)] dz} dy \quad (5.6)$$

Example: Suppose $\alpha(t) = \alpha$, $\beta(t) = \beta$, both constants. Then

$$x(t) = x(0)e^{-(\alpha+\beta)t} + \frac{\alpha}{\alpha+\beta} \{1 - e^{-(\alpha+\beta)t}\}. \quad (5.7)$$

Consequently our approximation to the expected value of $I(t)$ is

$$E[I(t)] \approx \bar{I}x(t) = \bar{I} \left[x(0)e^{-(\alpha+\beta)t} + \frac{\alpha}{\alpha+\beta} \{1 - e^{-(\alpha+\beta)t}\} \right]. \quad (5.8)$$

Return now to (5.4). After division by \sqrt{I} the following limiting differential equation occurs for $S(t)$:

$$dS(t) = -[\alpha(t) + \beta(t)]S(t)dt + \sqrt{\alpha(t)\{1-x(t)\} + \beta(t)x(t)} dW(t) \quad (5.9)$$

This is recognized as describing a non-stationary Ornstein-Uhlenbeck process; see COX and MILLER [4]. A formal solution to (5.9) is

$$S(t) = S(0)e^{-\int_0^t [\alpha(z) + \beta(z)] dz} + \int_0^t e^{-\int_u^t [\alpha(z) + \beta(z)] dz} \sigma(u) dW(u), \quad (5.10)$$

from which one can see directly that, given $S(0)$, $S(t)$ is normally distributed or Gaussian, being the weighted sum of Gaussian elements, dW . Further calculations give

$$E[S(t)] = E[S(0)]e^{-\int_0^t [\alpha(z) + \beta(z)] dz} \quad (5.11)$$

and

$$\text{Var}[S(t)] = \text{Var}[S(0)]e^{-2\int_0^t [\alpha(z) + \beta(z)] dz} + \int_0^t e^{-2\int_u^t [\alpha(z) + \beta(z)] dz} \sigma^2(u) du. \quad (5.11)$$

Example (continued): $\alpha(t) = \alpha, \beta(t) = \beta.$

In this case $S(t)$ is Gaussian with mean

$$E[S(t)] = E[S(0)]e^{-(\alpha+\beta)t}$$

and variance

$$\begin{aligned} \text{Var}[S(t)] = & \text{Var}[S(0)]e^{-2(\alpha+\beta)t} + \frac{\alpha\beta}{(\alpha+\beta)^2} [1-e^{-2(\alpha+\beta)t}] \\ & + \left(\frac{\beta-\alpha}{\beta+\alpha}\right) [x(0) - \frac{\alpha}{\beta+\alpha}]e^{-(\alpha+\beta)t} [1-e^{-(\alpha+\beta)t}]; \end{aligned} \quad (5.12)$$

the variance components may be added because $\{dW, u \geq 0\}$ is independent of events prior to $u = 0$, which determine $S(0)$. Thus for large t the present model suggests that $I(t)$, the number of individuals in the organization (which equals the number of records in the operating data base) is Gaussian, with mean

$$E[I(t)] \approx \frac{\alpha \bar{I}}{\alpha + \beta},$$

and variance

(5.13)

$$\text{Var}[I(t)] \approx \frac{\alpha\beta \bar{I}}{(\alpha+\beta)^2}$$

Of interest also is the fact that various functionals of $I(t)$ can be computed. For instance, the amount of storage space--reels of tape, or disk units--required to store each organization incumbent's status vector is proportional to $I(t)$. Roughly, we pay $kI(t)dt$ to store the data base elements for time period $(t, t+dt)$. Thus, total storage cost for a time T is

$$\begin{aligned}
D(T) &= \int_0^T kI(t)dt \\
&\approx k \int_0^T [\bar{I}x(t) + \sqrt{I} S(t)]dt \\
&= k\bar{I} \int_0^T x(t)dt + k\sqrt{I} \int_0^T S(t)dt
\end{aligned} \tag{5.14}$$

Since $\{S(t)\}$ is Gaussian, $\int_0^T S(t)dt$ is also Gaussian, with mean zero and variance that is explicitly expressible in terms of $\alpha(t)$ and $\beta(t)$. An additional twist is to calculate the cost of storing the data base, but allowing for discounting of future costs. Our approximation is

$$\begin{aligned}
D &= \int_0^{\infty} e^{-\Delta t} kI(t)dt \\
&\approx k\bar{I} \int_0^{\infty} e^{-\Delta t} x(t)dt + k\sqrt{I} \int_0^{\infty} e^{-\Delta t} S(t)dt
\end{aligned} \tag{5.14}$$

which is again Gaussian; Δ is the discount rate.

We next attempt to account for routine changes, each of which we will assume costs an average of c' ; the constant c' will depend upon the logical organization of the files. Our model assumes that, given $I(t)$, the expected number of changes in $(t, t+dt)$ is

$$E[dC(t)|I(t)] = \alpha(t)[\bar{I}-I(t)]dt + \beta(t)I(t)dt \tag{5.15}$$

According to our approximation this means that

$$E[dC(t)|I(t)] \approx \bar{I}[\alpha(t)\{1-x(t)\} + \beta(t)x(t)]dt + \sqrt{I} [\alpha(t)+\beta(t)]S(t)dt, \tag{5.16}$$

where $I(t) = \bar{I}x(t) + \sqrt{I} S(t)$. To translate to cost per time interval dt multiply by c' .

Consequently the total expected number of changes over an operating period $(0, T)$ is

$$E[C(T)] \approx \bar{I} \int_0^T [\alpha(t) + x(t)\{\beta(t) - \alpha(t)\}] dt + \sqrt{T} \int_0^T [\alpha(t) + \beta(t)] E[S(t)] dt \quad (5.17)$$

The latter expression can be evaluated with the aid of (5.8) and (5.11). It may also be shown that the total number of routine changes is Gaussian in this approximation.

Example: Once again $\alpha(t) = \alpha$, $\beta(t) = \beta$. Then

$$\begin{aligned} E[C(T) | x(0), S(0)] &\approx \bar{I} \int_0^T [\alpha + (\beta - \alpha)x(0)e^{-(\alpha + \beta)t} \\ &\quad + \frac{\alpha}{\alpha + \beta} \{1 - e^{-(\alpha + \beta)t}\}] dt + \sqrt{T} \int_0^T (\alpha + \beta) S(0) e^{-(\alpha + \beta)t} dt \quad (5.18) \\ &= \bar{I} [\alpha T + (\beta - \alpha) \left(\frac{x(0)}{\alpha + \beta} \{1 - e^{-(\alpha + \beta)T}\} + \frac{\alpha}{\alpha + \beta} T - \frac{\alpha}{(\alpha + \beta)^2} \{1 - e^{-(\alpha + \beta)T}\} \right) \\ &\quad + \sqrt{T} S(0) \{1 - e^{-(\alpha + \beta)T}\}] \\ &\sim \bar{I} \frac{2\alpha\beta T}{\alpha + \beta} + R(T), \end{aligned}$$

where $R(T)$ is a remainder term that is easily and explicitly calculated in this model. Of course the initial term is available directly from a simple classical birth-and-death model analysis, but the transient (T -dependent) terms are, thereby, not available in any usable form.

Lastly, we turn to a model for query or usage costs. A research data base will experience occasional requests for information about a subset of its elements (e.g. "what fraction of all individuals having

characteristic A also have characteristic B?"). If the data base is accessed in search for individuals of type A, for instance, then in general the cost of accessing should depend upon the number of entries. Empirical information would be desirable, but is lacking and in any case depends upon file organization. We assume that (i) the rate of attempted accesses in time is $v(t)$, that (ii) the expected cost per individual in the data base examined (as a prospective type A) is h ; further assume (iii) that all individuals in the data base must be examined at least once. We do not assume that individuals are already classed and filed as type A. Then the expected cost of such an access service is $hI(t)$, conditional on $I(t)$, and the expected cost of accessing in $(t, t+dt)$ is $v(t)hI(t)dt$. Thus, formally,

$$E[dA(t)|I(t)] = v(t)hI(t)dt \quad (5.19)$$

and in terms of our approximation,

$$E[dA(t)|I(t)] \approx h\bar{I} v(t)x(t) dt + h\sqrt{I} v(t)S(t)dt. \quad (5.20)$$

Consequently, if $I(t) = \bar{I}x(t) + \sqrt{I} S(t)$,

$$E[A(T)] \approx h\bar{I} \int_0^T v(t)x(t)dt + h\sqrt{I} \int_0^T v(t)E[S(t)]dt. \quad (5.21)$$

Example: $\alpha(t) = \alpha$, $\beta(t) = \beta$, and $v(t) = v$ (constant access rate).

For this model we need merely refer to expression (5.14) with vh replacing k .

5.3 Extensions.

The model just described is representative of a class of such models that remain to be studied and applied. Several alterations or reinterpretations of the model are suggested, and are under further consideration. For example we suggest the following.

- (i) The access rate, $v(t)$, to a data base will be influenced by the presence of annoying errors detected in the received data. Consequently, it will be profitable to blend the error screening technique (e.g. the limbo file) with the present model and calculations. A reasonable amount of data screening may well increase the effective access rate to a level sufficient to justify data bank maintenance. An altered form of our model can be made to reflect this effect.
- (ii) The basic model, (5.1), is only illustrative and can be modified, either to reflect empirical reality more closely, or to represent different control strategies. For example, one might regress $dI(t)$ on $I(t)$ in order to discover the approximate system dynamics. On the other hand, we may think of $\alpha(t)$ as being in part in the control of management: if $\alpha(t)$ is increased then so is system responsiveness to attainment of the goal level \bar{I} , but changes tend then to occur thick and fast and errors in records are likely to proliferate. This in turn is likely to result in reduced data base usage, particularly if the base is primarily employed for research purposes.

REFERENCES

- [1] ANDREWS, D., BICKEL, P., HAMPEL, F., ROGERS, W., and TUKEY, J., Robust Estimates of Location: Survey and Advances. Princeton Univ. Press, 1972.
- [2] CARTER, J. O'N., Quality control and analysis of error in the naval manpower data base, Naval Postgraduate School M.S. Thesis, Sept. 1972.
- [3] COX, D. R. and LEWIS, P. A. W., The Statistical Analysis of Series of Events. Methuen, London and Barnes and Noble, New York, 1966.
- [4] COX D. R., and MILLER, H. D., The Theory of Stochastic Processes, New York, John Wiley and Sons, 1965.
- [5] GOODMAN, A., LEWIS, P. A. W., and ROBBINS, H., Simultaneous estimation of large numbers of extreme quantiles, Communications in Statistics, (forthcoming).
- [6] MARSHALL, K. T., Manpower and personnel data bases in the Navy and Marine Corps, Naval Postgraduate School Technical Report, NPS55MT73021A, February 1973.
- [7] McNEILL, D., Diffusion limits for congestion models, Journal of Applied Probability, Vol. 10, No. 2, June 1973.
- [8] NAUS, J., JOHNSON, T., and MONTALVO, R., A probabilistic model for identifying errors in data editing, Journal of the American Statistical Ass'n., Vol. 67, December 1972.
- [9] ROSS, S., Introduction to Probability Models. New York, Academic Press, 1972.
- [10] TUKEY, J. W., Exploratory Data Analysis, Addison-Wesley Publishing Co.
- [11] YAGUCHI, T. G., Quantile estimation using the maximum transformation, stochastic approximation and the jackknife, Naval Postgraduate School M.S. Thesis, March 1973.