12

Shannon Information and Bayesian Design for Prediction in Accelerated Life-testing

ISABELLA VERDINELLI,[a,b] NICK POLSON[b,*] & NOZER D. SINGPURWALLA[c]

[a]*University of Rome, Italy and* [b]*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890 and* [c]*Department of Operations Research, The George Washington University, Washington, DC 20052 USA.*

1. INTRODUCTION

The current literature on accelerated life-testing emphasizes issues of inference and extrapolation under a given design, rather than optimal design. In this paper we consider a Bayesian decision theoretic approach to address such an issue. We begin with an account of Shannon information, the use of which arises in information theory; we discuss how it can be interpreted in a well-posed decision problem[1,2] and show how Shannon information is a tool especially suitable for dealing with design problems.[3-5] Then we describe the setup of accelerated life-testing. Suitable model hypothesis and prior assumptions are considered to show how to apply Shannon information ideas to design accelerated life tests. We do not mean to suggest that utility functions based on Shannon information are always appropriate. Our goal is simply to describe these utility functions and illustrate their use.

2. SHANNON INFORMATION

Consider two random variables X and Y. Suppose that the random variable of interest is X and the experimenter can observe Y to learn about X. Let $p(X, Y)$ denote the joint density of the pair of random

*Now at University of Chicago.

variables and let $p(X)$, $p(Y)$ denote the marginal densities of the two random variables. In a coding theory framework, X represents the message to be sent, Y the message received, obtained passing a coded version of X through a channel disturbed by noise. Let the probability distribution $p(X|Y)$ model the channel. In this setup, a quantity known as the mutual information between the random variables X and Y, defined as

$$I\{X:Y\} = E_{(X,Y)}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right],$$

is of interest; see for example Shannon,[6] and Gallager.[7] In the above expression, $E_{(X,Y)}$ denotes the expectation with respect to the joint density $p(X, Y)$. Mutual information $I\{X:Y\}$ has the property that $I\{X:Y\} \geq 0$, with equality if and only if X and Y are independent, $p(X, Y) = p(X)p(Y)$. This occurs when the density modeling the channel is such that $p(X|Y) = p(X)$. Therefore, the minimum of the mutual information corresponds to the case in which knowledge of Y gives no information about X. The interest, however, is in obtaining a good channel to send the message. Hence, a sensible criterion is to search the family of channels and select the one that maximizes the mutual information. The maximum of $I\{X:Y\}$ with respect to $p(X|Y)$ is known as channel capacity.[6,7]

The criterion of maximizing mutual information is equivalent to the principle of maximum expected utility, within a decision theory approach. Let us examine two specific problems: the first is the case of inference, where X represents an unobservable parameter of interest θ, Y is a vector of data $(y_1, y_2, \ldots, y_n)$ and the experimenter computes $p(X|Y) = p(\theta|\mathbf{y})$. The second problem is prediction, where the random variable of interest X is the future observation y_{n+1}, Y is still a vector of data $(y_1, y_2, \ldots, y_n)$ and $p(y_{n+1}|\mathbf{y})$ is of concern.

2.1. Inference and Prediction

In the first case the marginal density $p(\theta)$ represents the experimenter's beliefs about θ. This, as noted by Lindley,[3] differs from the transmission problem. However, the mutual information can still be defined, if we consider that the conditional density representing the channel can be given by the posterior density. With the previous notation $p(X|Y) = p(\theta|y_1, \ldots, y_n)$.

Followng Shannon,[6] Lindley[3] defined the amount of information given by a probability density $p(x)$ for a random variable as:

$$Ip(\cdot) = \int p(x) \log p(x)\, dx = E_X \log p(X). \tag{2.1}$$

In the Bayesian context, the experimenter's prior and posterior knowledge about θ are represented by $p(\theta)$ and $p(\theta|\mathbf{y})$ respectively, where $\mathbf{y} = (y_1, \ldots, y_n)$ is the vector of data. Lindley[3] defined the gain in information given by an experiment by $[Ip(\cdot|\mathbf{y}) - Ip(\cdot)]$ and the expected gain of information by $E_y[Ip(\cdot|\mathbf{y}) - Ip(\cdot)]$ Using definition (2.1), an alternative expression for the expected gain in information is obtained:

$$E_y E_{\theta|y}\left[\log \frac{p(\theta|\mathbf{y})}{p(\theta)}\right]. \tag{2.2}$$

This is also known as the expected Kullback–Leibler distance between the posterior and the prior. By Bayes theorem, we have: $p(\theta|\mathbf{y})/p(\theta) = p(\theta, \mathbf{y})/p(\mathbf{y})p(\theta)$ and, substituting this last formula into (2.2), Lindley's measure turns out to be precisely the mutual information between the random variables θ and $\mathbf{y} = (y_1, \ldots, y_n)$. Following Bernardo[1] and DeGroot[2] we show how mutual information can be viewed entirely within a Bayesian decision-theoretic framework. Let the decision space be given by the space of probabilty densities for θ, denoted by $\mathscr{P}$ and define inference as the decision problem of reporting a posterior density for θ. Thus, we consider a real-valued function $u: (\Theta \times \mathscr{P}) \to \mathbb{R}$, denoted by $u[\theta, p^*(\cdot)]$, describing the utility associated with the decision of reporting the density function $p^*(\cdot)$ for the unknown parameter θ, when θ is its true value. We emphasize that the first component of $u[\theta, p^*(\cdot)]$ is the value of the parameter and the second component is a density function that is being reported to express the uncertainty about θ. Note that p^* might not be our true distribution for θ. Let $\int u[\theta, p^*(\cdot)]p(\theta)\, d\theta$ be its expected value, $p(\theta)$ being the density representing the true beliefs about θ. From a decision theory point of view, the choice of the density $p^*(\theta)$ for θ should be such that it maximizes the expected utility.

Bernardo[1] showed that it is appropriate to consider a utility function that satisfies the following two properties, that correspond to requiring the utility function to be proper (or honest) and local:

1. The maximum of the expected utility is attained if and only if $p^*(\cdot) = p(\cdot)$.
2. $u[\theta, p^*(\cdot)] = u[\theta, p^*(\theta)]$ for all values of θ.

These assumptions characterize the logarithmic form of the utility function, in the sense that u satisfies 1. and 2. if and only if $u[\theta, p^*(\cdot)] = A \log p^*(\theta) + B(\theta)$ for some positive constant A and some function B.[1] Therefore, the gain in expected utility is easily seen to be given by (2.2) above and the expected gain in information can be interpreted as a change in expected utility.[2]

Let us now turn to the statistical problem of prediction, where the variable of interest X is the future observation y_{n+1} and Y is a vector of data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$; the mutual information between the random variables y_{n+1} and $\mathbf{y}$ is given by:

$$I\{y_{n+1}:\mathbf{y}\} = E_{(y_{n+1},\mathbf{y})}\left[\log \frac{p(y_{n+1}, \mathbf{y})}{p(y_{n+1})p(\mathbf{y})}\right]. \tag{2.3}$$

where $p(y_{n+1}) = \int p(y_1, \ldots, y_{n+1})\, dy_1 \ldots dy_n$.

By arguments similar to the ones discussed for the case of inference about θ, the use of either the information or the decision theory approach leads to (2.3). This criterion has been discussed in a prediction context by San Martini and Spezzaferri,[8] where model selection was the aim of the analysis.

2.2. Design

We now turn, more specifically, to the context of the experimental designs. Here the experimenter controls a set of variables $\mathscr{A}$, say. The information contained in Y about X varies with $\mathbf{A} \in \mathscr{A}$. Let us denote the conditional density by $p(X|Y, \mathbf{A})$ to emphasize the dependence on the control, or design variables. The experimenter can choose from a family of channels $p(X|Y, \mathbf{A})$. The criterion is to choose $\mathbf{A}^* \in \mathscr{A}$ so that the mutual information between X and Y is maximized:

$$\max_{\mathbf{A}} E_{(X,Y|\mathbf{A})}\left[\log \frac{p(X, Y|\mathbf{A})}{p(X)p(Y|\mathbf{A})}\right].$$

Here too, as in Section 2.1, it is possible to consider the two problems of inference and prediction. In the context of inference, the experimenter would choose a design $\mathbf{A}^*$ that maximizes:

$$E_{\mathbf{y}|\mathbf{A}} E_{\theta|\mathbf{y},\mathbf{A}}\left[\log \frac{p(\theta|\mathbf{y}, \mathbf{A})}{p(\theta)}\right], \tag{2.4}$$

that is the equivalent of expression (2.2). The use of (2.4) in design theory has been discussed, first by Stone[4,5] and later by Smith and Verdinelli.[9]

Giovagnoli and Verdinelli[10,11] and Verdinelli,[12] within the Lindley and Smith[13] hierarchical model. Some modifications of this criterion have been described by Verdinelli and Kadane.[14] Whittle[15] discussed theoretical and computational techniques to obtain optimal design. Chaloner[16] and Chaloner and Larntz[17] considered Bayesian designs for linear and nonlinear models.

In the context of prediction, the predictive density of the future observation y_{n+1}, depends on the design variables $\mathscr{A}$. The mutual information in (2.3), between y_{n+1} and $\mathbf{y}$, can be written as:

$$E_{\mathbf{y}|\mathbf{A}} E_{y_{n+1}|\mathbf{y},\mathbf{A}}\left[\log \frac{p(y_{n+1}|\mathbf{y}, \mathbf{A})}{p(y_{n+1})}\right]. \tag{2.5}$$

Note that the density $p(y_{n+1}|\mathbf{y}, \mathbf{A})$ is the usual predictive density of the future observation y_{n+1} given the vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and $\mathbf{A}$, while $p(y_{n+1})$ is the marginal density of the future observation y_{n+1} before the vector of data is observed and does not depend on $\mathbf{A}$. These densities will be referred to, respectively, as the posterior-predictive and prior-predictive.

3. MODEL FOR ACCELERATED LIFE TESTS

Let x_{ij} denote the observed time to failure of the j-th item under a stress $S_i (i = 1, 2, \ldots, k; j = 1, 2, \ldots, n_i)$, where S_i is considered less severe than $S_{i'}$ if $i \leqslant i'$. Let S_u be an environmental condition at which it is not convenient to conduct a life test and suppose that it is of interest to gather information about x_u, the time to failure under S_u, from testing items at stress condition S_i's, more severe than use stress.

Several quantities can be controlled by the experimenter, namely, the set of variables $\mathscr{A}$ will include the number k of levels of testing, the actual values of stresses $S_1, S_2, \ldots, S_k$, the number n_i of items to be tested at the i-th level, the number r_i of failures to be observed and, sometimes, the time T_i of termination of the life test under S_i. The model should consider them all, even if to solve a given design problem, some of them will be assumed fixed.

A typical distributional assumption in life testing is that the observations x_{ij} are lognormal with parameters μ_i and σ_i^2 and the time transformation function follows the power law, commonly used in both biometry and reliability (see for example Sethuraman & Singpurwalla.[18]

We consider the time transformation function in terms of $E(x_{ij})$. More specifically, for unknown constants $C>0$ and $P\geqslant 0$ it is assumed that:

$$E(x_{ij})=\exp\left\{\mu_i+\frac{\sigma_i^2}{2}\right\}=\frac{C}{S_i^P}. \tag{3.1}$$

We further assume that the variances σ_i^2 are known. There are several ways to help their specification, consistent with the physical aspects of the life testing setup. In particular it might be reasonable to expect that σ_i^2 decreases with S_i, but we will not deal with this aspect any further.

Let us consider the transformation $z_{ij}=\log x_{ij}$ so that $z_{ij}\sim N(\mu_i,\sigma_i^2)$; then, letting $a=\log C$, $b=-P$ and $V_i=\log S_i$, from expression (3.1) our model takes the form:

$$z_{ij}=a+bV_i-\frac{\sigma_i^2}{2}+\varepsilon_{ij}, \qquad \varepsilon_{ij}\sim N(0,\sigma_i^2).$$

Note that, as variances σ_i^2 are assumed known, the model above can be further simplified letting $y_{ij}=z_{ij}+\sigma_i^2/2$ and, denoting by $\mathbf{y}$ the vector of all y_{ij}, we have, in matrix notation, $\mathbf{y}=\mathbf{A}\beta+\varepsilon$, where $\mathbf{A}$ is the design matrix, to be determined:

$$\mathbf{A}^T=\begin{bmatrix} 1\ldots 1, & 1\ldots 1, & \ldots, & 1\ldots 1 \\ \underbrace{V_1\ldots V_1}_{n_1}, & \underbrace{V_2\ldots V_2}_{n_2}, & \ldots, & \underbrace{V_k\ldots V_k}_{n_k} \end{bmatrix},$$

$\beta^T=(a,b)$, $\varepsilon\sim N(0,\Sigma)$, with Σ the diagonal matrix: $diag\{\sigma_1^2 I_{n_1},\sigma_2^2 I_{n_2},\ldots,\sigma_k^2 I_{n_k}\}$, I_{n_i} denoting the n_i identity matrix. Let us now assume that the prior knowledge about the vector β can be expressed as: $\beta|\beta_0,\Sigma_0\sim N(\beta_0,\Sigma_0)$ where $\beta_0^T=(a_0,b_0)$ is known and so is

$$\Sigma_0=\begin{bmatrix}\sigma_a^2, & \sigma_{ab}\\ \sigma_{ab}, & \sigma_b^2\end{bmatrix}.$$

Note that problems can arise in cases where P is close to zero and positive; the assumption that $b=-P$ is normal should then be considered together with a choice of b_0 close to zero and a small value for σ_b^2.

4. DESIGN FOR ACCELERATED LIFE TESTS

Let $y_u=\log x_u+\sigma_u^2/2$, where x_u, as described earlier, is the time to failure under stress S_u and σ_u^2 is the known variance under stress S_u. Inference is required on x_u or, equivalently, on y_u. Criterion (2.5) consists in selecting the $\mathbf{A}^*$ that maximizes:

$$E_y\left[\int p(y_u|\mathbf{y},\mathbf{A})\log\frac{p(y_u|\mathbf{y},\mathbf{A})}{p(y_u)}\,dy_u\right], \tag{4.1}$$

where $p(y_u|\mathbf{y},\mathbf{A})$ and $p(y_u)$ are, respectively, the posterior-predictive and the prior-predictive densities of y_u. With the model hypothesis of Section 3, these densities are univariate normal. Specifically, let:

$$y_u\sim N(m_1,s_1^2) \qquad y_u|\mathbf{y},\mathbf{A}\sim N(m_2,s_2^2)$$

where:

$$m_1=[\,1,V_u]\beta_0, \qquad s_1^2=[1,V_u]\Sigma_0\begin{bmatrix}1\\V_u\end{bmatrix}+\sigma_u^2,$$

and

$$m_2=[1,V_u][\mathbf{A}^T\Sigma^{-1}\mathbf{A}+\Sigma_0^{-1}]^{-1}[\mathbf{A}^T\Sigma^{-1}\mathbf{y}+\Sigma_0^{-1}\beta_0],$$

$$s_2^2=[1,V_u][\mathbf{A}^T\Sigma^{-1}\mathbf{A}+\Sigma_0^{-1}]^{-1}\begin{bmatrix}1\\V_u\end{bmatrix}+\sigma_u^2,$$

It can be shown that expression (4.1) reduces to $\log s_1/s_2$. Hence maximizing (4.1) with respect to $\mathbf{A}$ is equivalent to minimizing the predictive variance s_2^2 with respect to $\mathbf{A}$ since s_1^2 depends neither on $\mathbf{A}$ nor on $\mathbf{y}$. Some examples will be given in the next section.

5. EXAMPLES

We consider now simple special cases to show how the fully Bayesian approach describes the solution to the optimal design for prediction in accelerated life-testing. Suppose first that C is known, let C be 1, say, so that the power law in (3.1) simplifies to $E(x_{ij})=S_i^{-P}$, $a=0$ and the transformed linear model reduces to $y_{ij}=bV_i+\varepsilon_{ij}$. Let us further assume that $k=1$. In other words it is only possible to test items at a single stress

point $S \geqslant S_u$. Let $\Sigma = \sigma^2 I_n$ and $\Sigma_0 = \sigma_b^2$. It can be easily seen that the posterior variance becomes:

$$s_2^2(V) = n^{-1}\sigma^2 \frac{V_u^2}{V^2 + \delta_b} + \sigma_u^2,$$

where $\delta_b = \sigma^2(n\sigma_b^2)^{-1}$. Hence $s_2^2(V)$ as a function of $V = \log S$, is minimized for values of V as large as possible. This conclusion might seem contrary to the intuition, but it is not so. The linear model we are considering is simply a straight line through the origin and to estimate y_u precisely we need to take observations as far away from V_u as possible. In connection with the above, it is of interest to plan the optimal number of items n to test at V. To see this, Fig. 1 plots the behavior of expression (4.1) -- the expected gain in Shannon Information—as a function of n, for $\sigma^2 = \sigma_b^2 = 1$ and $V = 2, 4, 6$ and 8.

One possible approach is to choose the value of n after which (4.1) shows little or no improvement. It is interesting to note that the optimum value of n decreases as V increases, as it is intuitively sensible from the previous considerations.

Let us now examine the more realistic case in which both P and C are unknown, but where we are still only allowed one stress point $k = 1$ to test items. This case might be of practical concern, for example under budget constraints. Even if the two parameters a and b in the linear model are not identifiable, the Bayesian framework gives us an appealing solution. Specifically, let us assume $\Sigma = \sigma^2 I_n$ and $\Sigma_0 = diag\{\sigma_a^2, \sigma_b^2\}$ and let us consider $s_2^2(V)$ as a function of V. Straightforward algebra shows that:

$$s_2^2(V) = \frac{(V - V_u)^2 + \delta_b + V_u^2\delta_a}{(1 + \delta_a)(V^2 + \delta_b) - V^2} + \sigma_u^2,$$

where $\delta_a = \sigma^2/n\sigma_a^2$ and $\delta_b = \sigma^2/n\sigma_b^2$. Hence, the value $V^* \geqslant V_u$ that minimizes $s_2^2(V)$ is readily seen to be $V^* = (1 + \delta_a)V_u$. The value V^* is thus independent of the prior variance σ_b^2 and it is as far away from V_u as the value of δ_a increases or, equivalently, as the values of the prior variance σ_a^2 decreases. Thus, if the prior knowledge of the intercept a is more precise, we choose values of V^* far away from V_u, in accordance with our previous remarks.

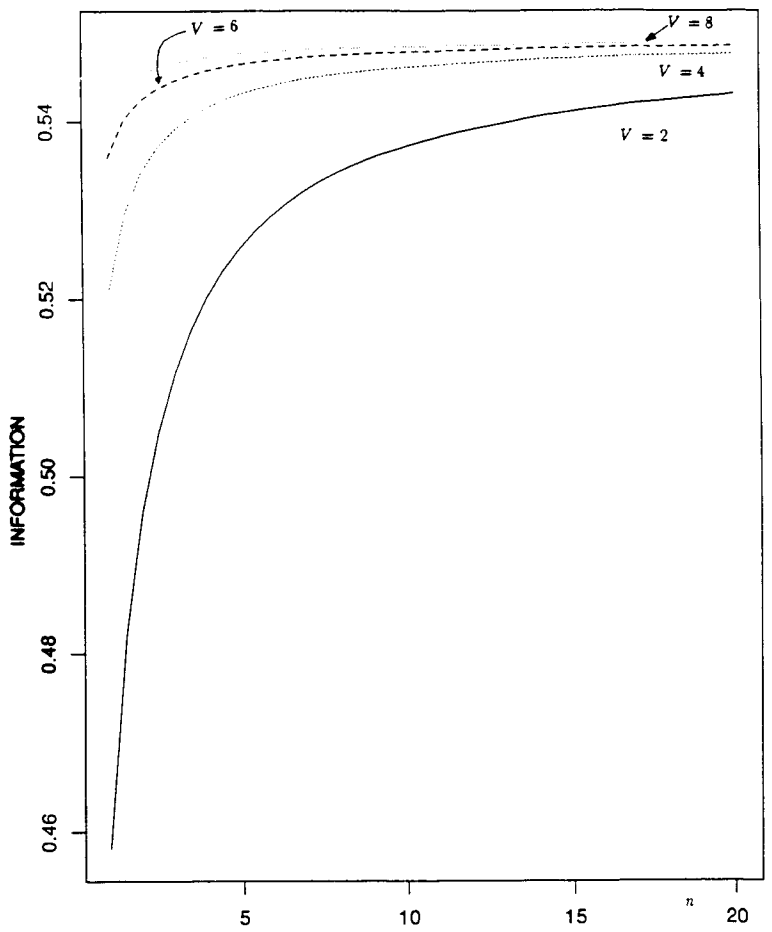

Fig. 1. Expected gain in Shannon Information as function of n. $V = 2, 4, 6, 8$ and $C = 1$.

ACKNOWLEDGMENTS

We would like to thank Larry Wasserman for useful discussions and suggestions and the Italian Research Council (C.N.R.) and the University of Siena for having made the conference possible.

REFERENCES

1. Bernardo, J. M., Expected information as expected utility. *Ann. Stat.*, **7** (1979) 686–90.
2. DeGroot, M. H., Changes in utility as information. In *Recent Developments in the Foundations of Utility and Risk Theory*, eds L. Daboni *et al.*, Dordrecht, Reidel, 1986, pp. 265–75.
3. Lindley, D. V., On the measure of information provided by an experiment. *Ann. Math. Stat.*, **27** (1956) 986–1005.
4. Stone, M., Application of a measure of information to the design and comparison of regression experiment. *Ann. Math. Stat.*, **30** (1959) 55–70.
5. Stone, M., Discussion of Kiefer. *J. R. Stat. Soc. B*, **21** (1959) 313–15.
6. Shannon, C. E., A mathematical theory of communication. *Bell System Tech. J.*, **27** (1948) 379–423 and 623–56.
7. Gallager, R. G., *Information Theory and Reliable Communication*, Wiley, New York, 1968.
8. San Martini, A. & Spezzaferri, F., A predictive model selection criterion. *J. R. Stat. Soc. B*, **46** (1984) 296–303.
9. Smith, A. F. M. & Verdinelli, I., A note on Bayesian design for inference using a hierarchical linear model. *Biometrika*, **67** (1980) 613–19.
10. Giovagnoli, A. & Verdinelli, I., Bayes D-optimal and E-optimal block designs. *Biometrika*, **70** (1983) 695–706.
11. Giovagnoli, A. & Verdinelli, I., Optimal block designs under a hierarchical linear model. In *Bayesian Statistics 2*. eds J. M. Bernardo *et al.*, North Holland, 1985, pp. 655–661.
12. Verdinelli, I., Computing Bayes D- and E-optimal designs for a two-way model. *The Statistician*, **32** (1983) 161–7.
13. Lindley, D. V. & Smith, A. F. M., Bayes estimates for the linear model (with Discussion). *J. R. Stat. Soc. B*, **34** (1972) 1–41.
14. Verdinelli, I. & Kadane, J. B., Bayesian designs for maximizing information and outcome. *J. Am. Stat. Assoc.*, (1992) (to appear).
15. Whittle, P., Some general points in the theory of optimum experimental designs. *J. R. Stat. Soc. B*, **35** (1973) 135–50.
16. Chaloner, K., Optimal Bayesian experimental designs for linear models. *Ann. Stat.*, **12** (1984) 283–300.
17. Chaloner, K. & Larntz, K., Optimal Bayesian designs applied to logistic regression experiments. *J. Stat. Plann. Inference*, **21** (1989) 191–208.
18. Sethuraman, J. S. & Singpurwalla, N. D., Testing of hypothesis for distributions in accelerated life testing. *J. Am. Stat. Assoc.*, **77** (1982) 204–8.

13

The Bayesian Approach to Quality

RICHARD E. BARLOW & TELBA Z. IRONY

IEOR Department, University of California, Berkeley, CA 94720, USA

1. STATISTICAL CONTROL

The control chart for industrial statistical quality control was invented by Dr. Walter A. Shewhart[1] in 1924 and was the foundation for his 1931 book *Economic Control of Quality of Manufactured Product*. (A highly recommended recent reference is Deming.[2] On the basis of Shewhart's industrial experience, he formulated several basic and important ideas. Recognizing that all production processes will show variation in product if measurements of quality are sufficiently precise, Shewhart described two sources of variation; namely

(i) variation due to *chance causes* (called 'common causes' by Deming[2]);
(ii) variation due to *assignable causes* (called 'special causes' by Deming[2]).

Chance causes are inherent in the system of production while assignable causes, if they exist, can be traced to a particular machine, a particular worker, a particular material, etc. According to both Shewhart and Deming, if variation in product is only due to chance causes, then the process is said to be in *statistical control*. Duncan[3] describes chance variations: 'If chance variations are ordered in time or possibly on some other basis, they will behave in a random manner. They will show no cycles or runs or any other defined pattern. *No specific variation to come can be predicted from knowledge of past variations.*' Duncan, in the