REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
21/4/2009	Final Report	January 1, 2006 to December 31, 2008

4. TITLE AND SUBTITLE
Integrating Disparate Information

5a. CONTRACT NUMBER
-

5b. GRANT NUMBER
N00014-06-1-0037

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Nozer D. Singpurwalla
The Institute for Reliability and Risk Analysis
The George Washington University

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The George Washington University
Office of Research Services
2121 Eye Street, Washington DC 20052

8. PERFORMING ORGANIZATION REPORT NUMBER
Final Report

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
875 N. Randolph St.
One Liberty Center
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; distribution unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions, and/or findings contained in this report are those of the author's and should not be construed as official Department of the Navy position, policy, or decision.

14. ABSTRACT
The focus of our research has been a rigorous investigation of the underlying mathematics driving the integrity and survivability of systems. Included herein are approaches of integrating information from diverse sources; these serve as a paradigm for information integration. The period of research performance has resulted in nine refereed publications, two books, and one publication to appear. See Attached for highlights of significant results.

15. SUBJECT TERMS
Belief, Probability, Chance, Vague Stochastic Systems, Residual Life, Prediction Intervals, Damage Processes, Competing Risk Processes, Hazard Potential, Degradation, Mathematical Finance, Risk.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Nozer D. Singpurwalla
Unclassified	Unclassified	Unclassified	Unlimited	Unknown	19b. TELEPHONE NUMBER *(Include area code)* 202 994 7515

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Listing of all Publications

a) **Papers Published in Peer Reviewed Outlets**

1) Singpurwalla, N. D. and A. G. Wilson (2009). Probability, chance and the probability of chance. **The IIE Transactions**. Vol. 41, No. 1, pp. 12-22.

2) Sellers, K. F. and N. D. Singpurwalla (2008). Many-valued Logic in Multistate and Vague Stochastic Systems. **The International Statistical Review**. Vol. 76, No. 2, pp. 247-267.

3) Singpurwalla, N. D. (2007). Betting on residual life: The caveats of conditioning. **Statistics & Probability Letters**. Vol. 77, No. 12, pp. 1354-1361.

4) Mesbah, M. and N. D. Singpurwalla (2008). A Bayesian Ponders the Quality of Life. **Statistical Models and Methods for Biomedical and Technical Systems** (Vonta, Nikulin, Limnios, and Huber-Carol, Eds.), pp. 373-384. Birkhouser.

5) Landon, J. and N. D. Singpurwalla (2008). Choosing a Coverage Probability for Prediction Intervals. **The American Statistician**. Vol. 62, No. 2, pp. 120-124.

6) Singpurwalla, N. D. (2008). Damage Processes. **Encyclopedia of Statistics in Quality and Reliability**. (F. Ruggeri et. al. Eds.), John Wiley and Sons, Inc.

7) Singpurwalla, N. D. (2006). The Hazard Potential: Introduction and Overview. **Journal of American Statistical Association**. Vol. 101, No. 476, pp. 1705-1717.

8) Singpurwalla, N. D. (2006). Reliability and Survival in Financial Risk. **Advances in Statistical Modeling and Inference** – Essays in Honor of Kjell Doksum (V. Nair, Ed). World Scientific Publications, pp. 93-114.

9) Singpurwalla, N. D. (2006). On competing risk and degradation processes. **The Second Erich L. Lehman Symposium – Optimality. Institute of Mathematical Statistics – Monograph Series**, (J. Rojo, Ed.) Vol. 49, pp. 289-304.

b) **Books and Monographs**

1) Singpurwalla, N. D. (2006). **Reliability and Risk: A Bayesian Perspective**. John Wiley, U.K.

2) Soyer, R., T. Mazzuchi and N. D. Singpurwalla (2004). **Mathematical Reliability: An Expository Perspective**. Kluwer, Academic Publishers.

c) **Papers Accepted but Not Published**

1) Singpurwalla, N. D. and S. P. Wilson (2009). The Mathematics of Risk and Reliability: A Select History. To Appear in **The Wiley Encyclopedia of Risk**.

Summary of Significant Findings.

In a) 1, we articulate via an example from reliability, the difference between the notions of probability, chance, likelihood, vagueness, belief and plausibility. To the best of our knowledge, it is the only document that carefully makes a distinction between these intertwined notions, and states clearly what each of these terms mean and when to use them.

In a) 2, we introduce the notion of a ***vague system***; i.e. a system that can simultaneously exist in more than one state. This is done via the mathematics of many valued logic. The traditional approach in system theory is via binary logic; it is limited in scope.

In a) 3, we make the important argument that when predicting remaining life, what matters most is the likelihood, not the probability model. This paper digs deep into the meaning of ***conditional probability*** and shows how one can arrive upon different predictions.

In a) 5, we address the important practical question of what should the coverage probability for a prediction interval be. Should it be 90%, 95%, or something else? We argue that this is a problem in ***optimal decision making***, a matter that has been totally overlooked.

In a) 7, we introduce a new fundamental notion, namely that of a ***hazard potential***. We argue that items fail when suitably chosen stochastic processes hit the hazard potential. The chosen stochastic processes depend on the environment in which units and systems operate.

In a) 9, we harness the thesis of a) 7 to argue that degradation is an abstract notion, but its observable markers are things like crack growth, wear, and CDA cell counts. We then make clear the meaning of competing risks and view them as stochastic processes. This is a chance in the manner in which one thinks of competing risks and degradation.

In b) 1, we summarize our research over the past several years, much, if not all, supported by the ONR, in reliability and survival analysis, and systems survivability. This book, we think is unique because it represents a paradigm shift in how one should think about reliability and survivability, and because unlike the existing books on the subject, it dwells into uncharted territories on several fronts. The point of view taken here is Bayesian and notions like the failure rate, survival, and systems integrity are interpreted from this perspective. The book also discusses the use of expert testimonies and information theoretic notions in failure data analysis and the design of life tests.

IIE Transactions (2009) 41, 12–22
Copyright © "IIE"
ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/07408170802322630

Probability, chance and the probability of chance

NOZER D. SINGPURWALLA[1,*] and ALYSON G. WILSON[2]

[1]*The George Washington University, The Institute for Reliability and Risk Analysis, Department of Statistics, Washington, DC 20052, USA*
E-mail: nozer@gwu.edu
[2]*Department of Statistics, Iowa State University, Ames, IA 50010, USA*
E-mail: agw@iastate.edu

Received March 2006 and accepted March 2007

In our day-to-day discourse on uncertainty, words like belief, chance, plausible, likelihood and probability are commonly encountered. Often, these words are used interchangeably, because they are intended to encapsulate some loosely articulated notions about the unknowns. The purpose of this paper is to propose a framework that is able to show how each of these terms can be made precise, so that each reflects a distinct meaning. To construct our framework, we use a basic scenario upon which caveats are introduced. Each caveat motivates us to bring in one or more of the above notions. The scenario considered here is very basic; it arises in both the biomedical context of survival analysis and the industrial context of engineering reliability. This paper is expository and much of what is said here has been said before. However, the manner in which we introduce the material via a hierarchy of caveats that could arise in practice, namely our proposed framework, is the novel aspect of this paper. To appreciate all this, we require of the reader a knowledge of the calculus of probability. However, in order to make our distinctions transparent, probability has to be interpreted subjectively, not as an objective relative frequency.

Keywords: Belief functions, biometry, likelihood, plausibility, quality assurance, reliability, survival analysis, uncertainty, vagueness

1. Probability and chance

1.1. *Introduction: Statement of the problem and objectives*

Consider the following archetypal problem that commonly arises in the contexts of biomedicine, engineering and the physical sciences.

Suppose that at some reference time τ, the "now time," YOU are asked to predict the time to failure T of some physical or biological unit. The capitalized YOU is to emphasize the fact that it is a particular individual, namely yourself, that has been asked to make the prediction. To facilitate prediction, you examine the unit carefully and learn all that you can about its genesis: how, when and where it was made. You denote this information by $\mathcal{H}(\tau)$, for history at time τ. In the case of biological units, $\mathcal{H}(\tau)$ would pertain to genetic and/or medical information. Suppose, as is generally true, that based on $\mathcal{H}(\tau)$ you conclude that prediction with certainty is not possible. Consequently, you are now faced with two options: walk away from the problem, or make an informed guess about T.

Suppose that you choose the second option and are prepared to make guesses about the event $(T \geq t)$, for some $t > 0$. In reliability, $t > 0$ is known as the "mission time." There are several additional caveats to this basic problem that go into forming our overall framework; these will be presented in Sections 2 and 3. In Section 2, we introduce the caveat of data, and in Section 3 the caveat of surrogate information.

To keep the mathematics simple, you introduce a counter, say X, and adopt the convention that $X = 1$ (a "success") whenever $T \geq t$, and $X = 0$ (a "failure"), otherwise. Thus, the events $(T \geq t)$ and $(X = 1)$ are isomorphic; however, there is a loss of granularity in going from T to X. This is because X continues to equal one, even when $T \geq t + a$, for any and all $a > 0$. With the introduction of X, informed guesses about $(T \geq t)$ boil down to informed guesses about $(X = 1)$. But what do we mean by an informed guess, and how shall we make this operational? Do the terms probability, chance and likelihood constitute an informed guess, or does each of these terms connote a distinct notion? Furthermore, do these terms cover all the scenarios of uncertainty that one can possibly encounter or are there scenarios that call for additional notions such as "belief" and "plausibility"? The aim of this paper is to show that each of the above terms encapsulates a distinct notion, so that their indiscriminate use should not be a matter of course.

*Corresponding author

1.2. *Personal probability: Making guesses operational*

By informed guess, we mean a quantified measure of your uncertainty about the event $(X = 1)$ in the light of $\mathcal{H}(\tau)$, and subsequent to a thoughtful evaluation of its consequences. Now, it is generally well acknowledged that probability is a satisfactory way to quantify uncertainty, and to some, such as Lindley (1982), the only satisfactory way. There are several interpretations of probability (c.f. Good (1965)). The one we shall adopt is *personal probability*, also known as *subjective probability*. Here, you quantify your uncertainty about the event $(X = 1)$, based on $\mathcal{H}(\tau)$, by your personal probability denoted:

$$P_Y(X = 1; \mathcal{H}(\tau)). \tag{1}$$

The subscript indexing P emphasizes the fact that the specified probability is that of a particular individual, namely, you. For convenience, we set $\tau = 0$ and denote $\mathcal{H}(0)$ by simply $\mathcal{H}$. Henceforth, we also omit the subscript associated with P, so that Equation (1) is written:

$$P(X = 1; \mathcal{H}) = p, \tag{2}$$

where $0 < p < 1$. The p so specified is a personal probability because it is not unique to all persons; more important, it can change with time for the same individual. This is because the background history for this person also changes, and it is the history that plays a key role in specifying a personal probability. Thus, an informed guess is tantamount to specifying a p, where p is a personal probability.

To make an informed guess operational, that is, to make a pragmatic use of it, we need to interpret p. For this we appeal to De Finetti (1974) who proposed that p represent the amount you—the specifier of p—is willing to stake in a *two-sided bet* (or gamble) about the event $(X = 1)$. That is, should X turn out to be one, you receive as a reward one monetary unit against the p staked out by you. Should X turn out to be zero, then the amount staked, namely p, is lost. By a two-sided bet, we mean the willingness to stake p for the event $(X = 1)$, or an amount $(1 - p)$ for the event $(X = 0)$. That is, you are indifferent between the two gambles: one monetary unit in exchange for p if $(X = 1)$, or one monetary unit in exchange for $(1 - p)$ if $(X = 0)$. It is useful to bear in mind that in keeping with the spirit of the individual nature of personal probability, the amount p represents your stake. For the same event $(X = 1)$, your colleague may choose to stake a different amount $\tilde{p}$, with $\tilde{p} \neq p$. It is also important to note that with p interpreted as a gamble, the bet will only be settled when X reveals itself. Thus, bets can only be made operational for events that are ultimately observed. We do not consider here the disposition of the second party in the bet; we assume that the second party is willing to accept any bet put forth by you.

Thus, to summarize, in the context of this paper, the word "probability" is used to denote the amount an individual is prepared to stake in a two-sided bet about an uncertain event. This probability can be specified based on $\mathcal{H}$ alone, and it is not essential that $\mathcal{H}$ contain data on items judged to be similar to the item in question. That is, personal probabilities can be specified without the benefit of having observed data.

1.3. *Chance or propensity: A useful abstraction*

Whereas specifying a personal probability can be done solely by introspection considering $\mathcal{H}$, a more systematic approach, which involves breaking the problem into smaller, easier problems, begins with invoking the law of total probability on the event $(X = 1; \mathcal{H})$. Specifically, for some unknown quantity θ, $0 < \theta < 1$, and an entity $\pi(\theta; \mathcal{H})$, whose interpretation is given later in Section 1.4:

$$P(X = 1; \mathcal{H}) = \int_0^1 P(X = 1 \mid \theta; \mathcal{H})\pi(\theta; \mathcal{H})\mathrm{d}\theta, \tag{3}$$

$$= \int_0^1 P(X = 1 \mid \theta)\pi(\theta; \mathcal{H})\mathrm{d}\theta, \tag{4}$$

if you assume that X is independent of $\mathcal{H}$ given θ. That is, were you to know θ, then knowledge of $\mathcal{H}$ is unnecessary. The meaning of θ, known as a *parameter*, remains to be discussed, but for now we state that in the language of personal probability, Equation (3) implies an *extension of the conversation* from $P(X = 1; \mathcal{H})$ to $P(X = 1 \mid \theta; \mathcal{H})$. The idea here is that after invoking the assumption of independence, you may find it easier to quantify your uncertainty about $(X = 1)$ were you to know θ, than quantifying the uncertainty based on $\mathcal{H}$. Whereas the dimension of $\mathcal{H}$ can be very large, the dimension of θ is one. Thus, the role of the parameter θ is to simplify the process of uncertainty quantification by imparting to X independence from $\mathcal{H}$.

In Equation (4), the quantity $P(X = 1 \mid \theta)$ is known as a *probability model* for the binary X. Following Bernoulli, you let $P(X = 1 \mid \theta) = \theta$, where $P(X = 1 \mid \theta)$ represents your bet (personal probability) about the event $(X = 1)$ were you to know θ. This brings us to the question of what does θ mean? That is, how should we interpret θ?

The meaning of θ was made transparent by De Finetti (c.f. Lindley and Phillips (1976)) in his now famous theorem on binary exchangeable sequences. Loosely speaking, this theorem says that if a large number of units judged similar to each other (the technical term is *exchangeable*) and to the unit in question were to be observed for their survival or failure until t, and if $X_i = 1$ if the ith item survived until t ($X_i = 0$ otherwise), then:

$$\theta = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{5}$$

that is θ is the average of the X_is, when the number of X_is is infinite. De Finetti refers to this θ as a *chance* or *propensity*. Note that there is no personal element involved in defining θ, other than the fact that θ derives from the behavior of exchangeable sequences, and exchangeability is a

judgment. What you judge to be exchangeable may not sit well with your colleagues. Because θ connotes the limit of an exchangeable binary sequence, θ can be seen as an objective entity. More important, since θ cannot be actually observed (n in the Equation (5) is infinite), we claim that chance is an abstract construct. It is a useful abstraction all the same, because in writing $P(X = 1 \mid \theta) = \theta$, you are saying that your stake on the uncertain event $(X = 1)$ is θ, were you to know θ. But no one can possibly tell you what θ is, and this is what leads us to the next section. But before we do so, it may be of interest to mention a few words about two other interpretations of θ.

One is due to Laplace, who in keeping with the scientific climate of his time, and being influenced by Newton, was concerned with cause and effect relationships. Accordingly, to Laplace, θ was the cause of an effect, namely, the event $(X = 1)$. The second interpretation of θ stems from the relative frequency interpretation of probability. Indeed, here θ is taken to be the probability that $X = 1$.

Finally, even though the notion of chance introduced here has been in the context of binary variables, a parallel notion also exists for other kinds of variables.

1.4. *Probability of chance: Taking chances with chance*

Since θ is unknown, and in principle can never be known, you are uncertain about θ. In keeping with the dictum that all uncertainty be described by probability, you let $P_Y(\Theta \leq \theta; \mathcal{H})$ encapsulate your bet on the event $(\Theta \leq \theta)$. Here, in keeping with standard convention, all unknown quantities are denoted by capital letters and their realized values by the corresponding small letter; thus our use of Θ and θ. Since Θ can take all values in the continuum (0, 1), we shall assume that $P_Y(\Theta \leq \theta; \mathcal{H})$ is "absolutely continuous," so that its density at θ exists, for $0 < \theta < 1$. We denote this density by $\pi_Y(\theta; \mathcal{H})$ and interpret it as

$$\pi(\theta; \mathcal{H})\mathrm{d}\theta \approx P(\theta \leq \Theta \leq \theta + \mathrm{d}\theta; \mathcal{H}).$$

For convenience, the subscript Y has been dropped.

Thus, $\pi(\theta; \mathcal{H})\mathrm{d}\theta$ is approximately your personal probability that the unknown chance Θ is in the interval $[\theta, \theta + \mathrm{d}\theta]$. Since θ will never be known, the bet on Θ cannot be settled. However, since $\pi(\theta; \mathcal{H})$ goes into determining $P(X = 1; \mathcal{H})$—see Equation (6) below—and since bets on $(X = 1; \mathcal{H})$ can be settled, $\pi(\theta; \mathcal{H})$ can also be interpreted as a technical device that helps you specify your bet on an observable.

With the above in place, plus the fact that in our case $P(X = 1 \mid \theta) = \theta$, Equation (4) becomes:

$$P(X = 1; \mathcal{H}) = p = \int_0^1 \theta \times \pi(\theta; \mathcal{H})\mathrm{d}\theta. \qquad (6)$$

Equation (6) above is noteworthy. It embodies: (i) a personal probability about the event $(X = 1)$—the left-hand side; (ii) a chance Θ taking the value θ; and (iii) a personal probability about the chance Θ belonging to the interval $[\theta, \theta + \mathrm{d}\theta]$—the entity $\pi(\theta; \mathcal{H})\mathrm{d}\theta$. This equation helps us make transparent the difference between probability, chance and the probability of chance.

There is another angle from which Equation (6) can be viewed. This comes from the fact that the right-hand side of Equation (6) is your *expected value* of Θ, the expected value being determined by your $\pi(\theta; \mathcal{H})$. Denoting this expected value by $E_Y(\Theta)$, we have:

$$P(X = 1; \mathcal{H}) = p = E_Y(\Theta),$$

implying that your personal probability for the event $(X = 1)$ is your expected value of the chance Θ with respect to $\pi(\theta; \mathcal{H})$, your personal probability about chance.

2. The likelihood of chance

2.1. *Introducing the caveat of data*

We supplement the framework of the basic problem of Section 1.1 by introducing our first caveat. Suppose that in addition to $\mathcal{H}(\tau)$, you also have at hand the binary $x_1, \ldots, x_n$, where $x_i = 1$ if the life-length of the ith item has actually been observed to exceed t, and $x_i = 0$, otherwise. The n items that go into constituting the data $\mathbf{x} = (x_1, \ldots, x_n)$ are judged by you, prior to observing the $\mathbf{x}$, to be similar (or exchangeable) to the item in question. What can you now say about the unobserved X? In other words what is your prediction for the event $(X = 1)$ in the light of $\mathcal{H}(\tau)$ as well as $\mathbf{x}$? Certainly, the observed $\mathbf{x}$ should help you sharpen your prediction. Consequently, you are now called upon to assess $P(X = 1; \mathbf{x}, \mathcal{H})$.

One possibility would be to think hard about all that you have at hand, namely, $\mathbf{x}$ and $\mathcal{H}$, and then simply specify $P(X = 1; \mathbf{x}, \mathcal{H})$ as p^*, where $p^* \in (0, 1)$. Here p^* encapsulates your bet on the event $(X = 1)$ in the light of $\mathbf{x}$ and $\mathcal{H}$. If p^* happens to be identical to the p of Equation (2), then you are declaring the opinion that the data $\mathbf{x}$ has not had a sufficient impact on your beliefs for you to change your bet from your original p. From a philosophical point of view, there is nothing in the theory of subjective probability that stops you from specifying a p^* by introspection alone. However, from a computational point of view, it is efficient to proceed formally along the lines given below, because introspection to specify p^* subsequent to having specified p may lead to an inconsistency (technically *incoherence*). By incoherence, we mean a scenario involving a gamble in which "heads I win, tails you lose."

2.2. *Bayes' law: The mathematics of changing your mind*

To address the scenario presented in Section 2.1, you start by pondering the matter of assessing your uncertainty about $(X = 1)$, in the light of $\mathcal{H}$, *were* you to know (but do not

know) the disposition of $X_1, \ldots, X_n$; here $X_i = 1$, if the ith item judged to be similar to the item in question has a life-length that exceeds t ($X_i = 0$, otherwise). That is, what would be your $P(X = 1|X_1, \ldots, X_n, \mathcal{H})$? To address this question, you follow the same line of reasoning used to arrive upon Equation (4), that is, extend the conversation to θ, and obtain

$$\begin{aligned} P(X = 1|X_1, \ldots, X_n; \mathcal{H}) &= \int_0^1 P(X = 1 \mid \theta, X_1, \ldots, X_n) \\ &\quad \times \pi(\theta|X_1, \ldots, X_n; \mathcal{H})\mathrm{d}\theta, \\ &= \int_0^1 P(X = 1 \mid \theta) \times \pi(\theta|X_1, \ldots, X_n; \mathcal{H})\mathrm{d}\theta, \\ &= \int_0^1 \theta \times \pi(\theta|X_1, \ldots, X_n; \mathcal{H})\mathrm{d}\theta. \end{aligned} \tag{7}$$

The second equality is a consequence of your judgment that X is independent of $X_1, \ldots, X_n$, were you to know θ, and the third a consequence of choosing $P(X = 1 \mid \theta) = \theta$ as a probability model for X. The quantity $\pi(\theta|X_1, \ldots, X_n; \mathcal{H})$ is the probability density at θ of your $P(\Theta \leq \theta|X_1, \ldots, X_n; \mathcal{H})$.

To obtain $\pi(\theta|X_1, \ldots, X_n; \mathcal{H})$ you invoke Bayes' law; thus:

$$\begin{aligned} \pi(\theta|X_1, \ldots, X_n; \mathcal{H}) &\propto P(X_1, \ldots, X_n \mid \theta; \mathcal{H}) \times \pi(\theta; \mathcal{H}) \\ &= \prod_{i=1}^{n} P(X_i = x_i \mid \theta) \times \pi(\theta; \mathcal{H}), \end{aligned} \tag{8}$$

by the multiplication rule, and by the independence of the X_is from each other, were you to know θ, and with $x_i = 1$ or 0. For $P(X_i = x_i \mid \theta)$, you once again choose Bernoulli's model, so that $P(X_i = x_i \mid \theta) = \theta^{x_i}(1 - \theta)^{1-x_i}$.

With the above in place, you now have:

$$\pi(\theta|X_1, \ldots, X_n; \mathcal{H}) \propto \prod_{i=1}^{n} \{\theta^{x_i}(1 - \theta)^{1-x_i}\}\, \pi(\theta; \mathcal{H}). \tag{9}$$

Since $\pi(\theta; \mathcal{H})$ encapsulates your uncertainty about Θ in the light of $\mathcal{H}$ alone, and $\pi(\theta|X_1, \ldots, X_n; \mathcal{H})$ your uncertainty about it were you to be provided additional information via the $X_1, \ldots, X_n$, we say that Bayes' law provides a mathematical prescription for changing your mind about the unobservable Θ. Once Equation (9) is at hand we may incorporate it in Equation (7) to write:

$$\begin{aligned} &P(X = 1|X_1, \ldots, X_n; \mathcal{H}) \\ &\quad \propto \int_0^1 \theta \prod_{i=1}^{n} \{\theta^{x_i}(1 - \theta)^{1-x_i}\}\pi(\theta; \mathcal{H})\mathrm{d}\theta, \end{aligned} \tag{10}$$

as a prescription of how to change your mind about the event $(X = 1)$ itself.

2.3. *Likelihood function: The weight of evidence*

There are two aspects of Equations (8) to (10) that need to be emphasized. The first is that the left-hand sides of these equations pertain to conditional events, namely the proposition that "were you to know the disposition of the X_is, $i = 1, \ldots, n$"; that is, supposing you were provided with the realizations of each X_i. The second feature is that they inform the reader as to how you express your uncertainties (or bets) about Θ and X respectively, once the X_is reveal themselves as x_i. Implicit to this bet is your particular choice of probability models $P(X = x \mid \theta)$ and $P(X_i = x_i \mid \theta)$, $i = 1, \ldots, n$.

In actuality, however, the X_is have indeed revealed themselves in the form of data, as $\mathbf{x} = (x_1, \ldots, x_n)$, where each x_i is known to you as being one or zero. In view of this, the left-hand sides of Equations (8) to (10) should be rewritten as $\pi(\theta; x_1, \ldots, x_n, \mathcal{H})$ and $P(X = 1; x_1, \ldots, x_n, \mathcal{H})$ respectively. But more significant is the fact that the quantity $P(X_i = x_i \mid \theta)$ of Equation (8) can no longer be interpreted as a probability. This is because the notion of probability is germane only for events that have yet to occur, or for events that have occurred but whose disposition is not known to you. In our case, X_i is known to you as $x_i = 1$ or $x_i = 0$, thus $P(X_i = x_i \mid \theta)$ is not a probability. So what does the quantity $P(X_i = x_i \mid \theta) = \theta^{x_i}(1 - \theta)^{1-x_i}$, with x_i fixed as zero or one, and θ unknown, mean? Similarly, in the context of Equation (9) with $r = \sum_{i=1}^{n} x_i$, what does the quantity:

$$\prod_{i=1}^{n} \{\theta^{x_i}(1 - \theta)^{1-x_i}\} = \theta^r(1 - \theta)^{n-r}, \tag{11}$$

with n and r known, but θ unknown, mean? Note that r is the total number of successes.

As a function of θ, with n and r fixed, the quantity $\theta^r(1 - \theta)^{n-r}$ is called the *likelihood function* of θ; it is denoted, $\mathcal{L}_Y(\theta; n, r)$, the subscript, which will henceforth be dropped, signaling the fact that like probability, the likelihood function is also personal. Since $\mathcal{L}(\theta; n, r)$ is not a probability, the likelihood function, even though it is derived from a probability model, is not a probability. It can be viewed as a function that assigns weights to the different values θ that Θ can take, in the light of the known n and r; these latter quantities can be viewed as *evidence*. Thus, the likelihood function can be interpreted as a function that prescribes the weight of evidence provided by the data for the different values that chance Θ can take. For example, with $n = r = 1$, $\mathcal{L}(\theta; n = r = 1) = \theta$; this suggests—see Fig. 1—that with $n = r = 1$, more weight is given by the likelihood function to the large values of θ than to the smaller values.

To summarize, the expression $P(X_i = x_i \mid \theta) = \theta^{x_i}(1 - \theta)^{1-x_i}$, specifies a probability of the event $(X_i = x_i)$ when X_i is unknown, and θ is assumed known; whereas with X_i known as x_i, it specifies a likelihood for the unknown θ. With $\mathbf{x}$ known, Equation (10) when correctly written becomes:

$$P(X = 1; \mathbf{x}, \mathcal{H}) \propto \int_0^1 \theta(\theta^r(1 - \theta)^{n-r}) \times \pi(\theta; \mathcal{H})\mathrm{d}\theta. \tag{12}$$

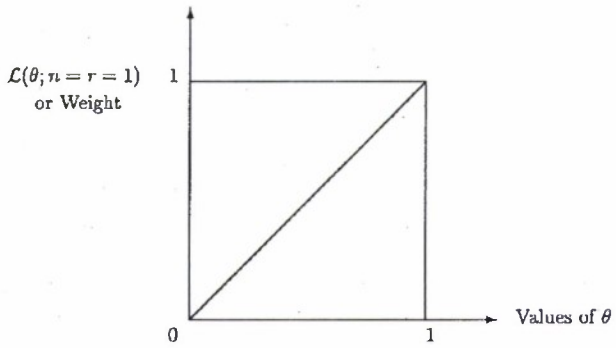

Fig. 1. The likelihood function with $n = r = 1$.

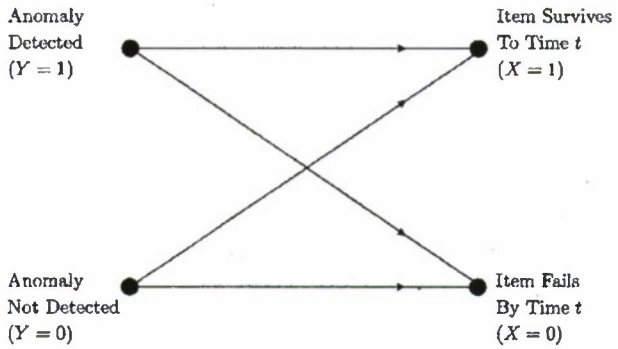

Fig. 2. Effect of anomalies on survival.

Equation (12) is interesting. It encapsulates, as we read from left to right, the four notions we have introduced thus far: personal probability (the left-hand side); chance (the parameter θ); the likelihood of chance (the quantity $\theta^r(1-\theta)^{n-r}$); and the probability of chance (the quantity $\pi(\theta; \mathcal{H})$).

Note also that the right-hand side of Equation (12) is the expected value of a function of Θ, namely, the function $\Theta^{r+1}(1-\Theta)^{n-r}$. Thus, we may say that the effect of the data $\mathbf{x}$ is to change your bet on the event $(X = 1)$ from $E_Y(\Theta)$ to $E_Y(\Theta^{r+1}(1-\Theta)^{n-r})$.

3. Imprecise surrogates: motivation for vagueness and belief

In Section 1 we outlined a problem that is the focus of our discussion, and in Section 2 we added a feature to it by bringing in the role of data. The notions used in Sections 1 and 2 are probability, chance and likelihood. Are these the only ones needed to address all problems pertaining to uncertainty? Are there circumstances that pose a challenge to us in terms of being able to lean on these notions alone? If so, what are these, and under what scenarios do we need to go beyond what has been introduced and discussed? The purpose of this section is to address the above and related questions. But first we bring into play our second caveat and explore the circumstances under which the notions of probability, chance and likelihood will suffice to address this caveat. The caveat in question pertains to the presence or not of detectable anomalies during inspection, quality control and other diagnostic testing functions.

3.1. *Anomalies: A surrogate of failure*

To keep our discussion simple, suppose that in order to assess your uncertainty about the event $(X = 1)$, you have at your disposal $\mathcal{H}$ and also a knowledge of the presence or the absence of a detectable anomaly. An anomaly could be a visible defect, or noticeable damage, or some other suitable indicator of imperfection. Anomalies could be present and yet not be detected. We denote the presence of a detected anomaly by letting a binary variable Y take the value one; the absence of a detectable anomaly by letting $Y = 0$. The presence of an anomaly does not necessarily imply that X will be zero; similarly, its absence is no assurance (to you) that X will be one; see Fig. 2. Rather, like the $X_1, \ldots, X_n$ of Section 2, the presence or absence of a detectable anomaly helps you sharpen your assessment of the uncertainty about $(X = 1)$.

Suppose then, that $Y = y$ has been observed, with $y = 1$ or 0, and that you are required to assess $P(X = 1; y, \mathcal{H})$. A simple way to proceed would be to treat y as a part of $\mathcal{H}$, and upon careful introspection specify:

$$P(X = 1; y, \mathcal{H}) = \widehat{p}, \quad 0 < \widehat{p} < 1,$$

as your bet on the event $(X = 1)$. The $\widehat{p}$ above is like the p of Section 1, in the sense that if $\widehat{p} = p$, then y has had no effect on your disposition about $(X = 1)$. There is, of course a more systematic way to incorporate the effect of y into your analysis, and this involves a use of the likelihood. To see how, start by pondering the matter of assessing your uncertainty about the event $(X = 1)$, in the light of $\mathcal{H}$, *were* you to know (but do not know) the disposition of Y. This is what was also done in Section 2.2. That is, you ask yourself what $P(X = 1 | Y; \mathcal{H})$ should be? By Bayes' law:

$$P(X = 1 | Y; \mathcal{H}) \propto P(Y = y | X = 1; \mathcal{H}) \times P(X = 1; \mathcal{H}),$$

$y = 1$ and 0. For $P(X = 1; \mathcal{H})$ you may use your p of Equation (2). To proceed further, you need to specify a probability model for Y, conditional on $(X = 1)$. That is, you need to specify $P(Y = 1 | X = 1; \mathcal{H})$ and $P(Y = 0 | X = 1; \mathcal{H})$; this is tantamount to specifying a joint distribution for X and Y. Once this can be done, you have:

$$P(X = 1 | Y; \mathcal{H}) \propto P(Y = y | X = 1; \mathcal{H}) \times p. \qquad (13)$$

However, in actuality, Y has been observed as $y = 1$ or $y = 0$. Consequently, Equation (13) becomes

$$P(X = 1; y, \mathcal{H}) \propto \mathcal{L}(X = 1; y, \mathcal{H}) \times p, \qquad (14)$$

where $\mathcal{L}(X = 1; y, \mathcal{H})$ is your likelihood function for the unknown event $(X = 1)$ in the light of the evidence y and $\mathcal{H}$. The probability model $P(Y = y|X = 1; \mathcal{H})$ helps you specify the likelihood. Equation (14) says that your bet on the event $(X = 1)$ in the light of y and $\mathcal{H}$, is proportional to your bet on $(X = 1)$ based on $\mathcal{H}$ alone, multiplied by your likelihood. The approach prescribed above is more systematic than the one involving the specification of $\hat{p}$ based on introspection alone, because it incorporates the p of Equation (2). A key point to note is that $\mathcal{L}(X = 1; y, \mathcal{H})$ is the likelihood of an observable event; it is not the likelihood of chance Θ discussed in Section 2.3. Should you prefer to work with the likelihood of chance, then you must introduce chance into your pondering. To do so, you may proceed as follows:

$$P(X = 1|Y; \mathcal{H}) = \int_0^1 P(X = 1 \mid \theta, Y; \mathcal{H}) \times \pi(\theta|Y; \mathcal{H})\mathrm{d}\theta,$$

which extends the conversation to θ, as was done to arrive at Equation (3). If you now assume that $(X = 1)$ is independent of both Y and $\mathcal{H}$, were you to know θ, and assume Bernoulli's model, then:

$$P(X = 1|Y; \mathcal{H}) = \int_0^1 \theta \times \pi(\theta|Y; \mathcal{H})\mathrm{d}\theta. \qquad (15)$$

But by Bayes' law:

$$\pi(\theta|Y; \mathcal{H}) \propto P(Y = y \mid \theta; \mathcal{H}) \times \pi(\theta; \mathcal{H}). \qquad (16)$$

Consequently, to proceed further, you need to specify a probability model for the anomaly Y, were you to know θ, and also $\pi(\theta; \mathcal{H})$, an entity that has already appeared in Sections 1 and 2. Since Y has in actuality been observed (as $y = 1$ or $y = 0$), Equation (16) becomes:

$$\pi(\theta; y, \mathcal{H}) \propto \mathcal{L}(\theta; y, \mathcal{H}) \times \pi(\theta; \mathcal{H}),$$

where $\mathcal{L}(\theta; y, \mathcal{H})$ is the likelihood function of the chance Θ, in the light of $\mathcal{H}$ and evidence about the anomaly y. With the above in place Equation (15) becomes:

$$P(X = 1; y, \mathcal{H}) \propto \int_0^1 \theta \times \mathcal{L}(\theta; y, \mathcal{H}) \times \pi(\theta; \mathcal{H})\mathrm{d}\theta.$$

To compare the above equation with Equation (14) (their left-hand sides are the same), we note that since $p = E(\Theta)$, Equation (14) may also be written as

$$P(X = 1; y, \mathcal{H}) \propto \int_0^1 \theta \times \mathcal{L}(X = 1; y, \mathcal{H}) \times \pi(\theta; \mathcal{H})\mathrm{d}\theta.$$

The last two equations signal the fact that in order to incorporate the effect of the detected anomalies into the assessment of your uncertainty about $(X = 1)$, you should be prepared to either specify the likelihood of $(X = 1)$ in the light of y (and $\mathcal{H}$), or the likelihood of θ in the light of y (and $\mathcal{H}$), whichever is more convenient. To specify these likelihoods, you may want to specify $P(Y = y|X = 1; \mathcal{H})$ or $P(Y = y \mid \theta; \mathcal{H})$, probability models for Y, were you to know X or θ, respectively. Of these, the former may be easier to assess than the latter, since it is based only on observables. We shall therefore focus on the case $P(Y = y|X; \mathcal{H})$, and refer to it as a *postmortem probability model.*

3.2. *Eliciting postmortem probabilities: Potential obstacles*

The material of Sections 1 and 2 required of you the specification of $P(X = x \mid \theta)$ and $\pi(\theta; \mathcal{H})$, for $x = 1$ or 0. For the former, Bernoulli's model is a natural choice; for the latter, a beta density with parameters α and β is a choice with much flexibility. Thus, for $0 < \theta < 1$:

$$P(X = x \mid \theta) = \theta^x(1 - \theta)^{1-x},$$

and

$$\pi(\theta; \mathcal{H}) = \pi(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1 - \theta)^{\beta-1}.$$

Coming to the scenario of Section 3, you are required to specify the above, and also a model for the postmortem probability $P(Y = y|X = x; \mathcal{H})$, for $x, y = 1$ or 0. The latter could pose two difficulties. The first is that you should be able to probabilistically relate detectable anomalies and failure; Fig. 2 with the direction of the arrows reversed could provide guidance. The second—a bigger problem—can arise because of the fact that the absence or the presence of any trait which qualifies as an anomaly may not be easily determined. For example, both a surface scratch and a dent qualify as defects, but the former could be less deleterious to an item's survival than the latter. Also, at what point does a rough scratch get labeled as a dent? The classification of an anomaly is therefore not crisp, so that the event "anomaly" is not well defined. It is this lack of crispness that motivates a consideration of "vagueness" as another aspect of uncertainty quantification; more on this will be said in Section 4.

One manifestation of this absence of crispness is that responses to questions for eliciting postmortem probabilities tend to be unhelpful. The following two responses from an actual scenario are illustrative.

1. "If the unit works, there is a less than 20% chance that we would have detected an anomaly. If it does not, we would be seeing something 20–40% of the time."
2. "If it works, that means that it was well manufactured. If it does not, then it means that it was handled poorly when it was shipped."

Clearly, pinning down postmortem probabilities from statements like the two above is not possible. At best statement 1 can provide bounds on the postmortem probabilities, and statement 2 has no probabilistic content whatsoever. Yet statements 1 and 2 provide information, albeit not in the form required by the calculus of probability.

To summarize, as long as the event "anomaly" is well defined so that one is able to precisely specify the postmortem probabilities, the development of Section 3.1 can be used,

and to do so all that one needs are the notions of probability, chance and likelihood. Once difficulties of the type discussed above come into play, postmortem probabilities cannot be elicited. When such is the case, the notions of "vagueness" and "belief" enter the arena of uncertainty quantification. We emphasize that we do not see these notions as a prelude to supplanting probability; rather, they enhance probability by making its use more encompassing. However, to some, like Zadeh (1978), the notion of vagueness invites alternatives to probability, a matter upon which we disagree.

4. Harnessing vagueness: Uncertainty quantification under imprecision

What do we mean by the term "vagueness"? Is it synonymous with the term "imprecision"? How do vagueness and imprecision enter the arena of uncertainty quantification? These are some of the questions that we aim to address in this section. We shall use the scenario of anomalies discussed in Section 3 as a point of discussion.

4.1. *Fuzzy sets and the uncertainty of classification*

As a preamble, recall that in Section 3.1, Y was a binary variable taking values $y = 0$ or $y = 1$, with $Y = 0(1)$ denoting the absence (presence) of a detectable anomaly. Declaring that $Y = 0$ or 1 is often a judgment call, which does not encapsulate the degree of the anomaly. In this section we refine the above process by introducing some granularity to the values y that Y can take. To do so, we let Y denote some undesirable characteristic of the item in question that can be quantified—for instance the depth of a scratch—and allow Y to take a continuum of values y in some well-defined range, say $\mathcal{R} = [0, M]$, where M is specified. Let $\tilde{A}$, a subset of $\mathcal{R}$, be the set of all ys that lead to the assessment that the item in question has an anomaly. Now if there exists a value y^* such that for any $y \geq y^*$ an anomaly is declared, then $\tilde{A}$ is called a *crisp* (or a sharp) set; crisp to reflect the fact that $\tilde{A}$ has well-defined boundaries. Consequently, any y can be placed with precision in the set $\tilde{A}$, or its complement. Crisp sets are said to adhere to the *law of the excluded middle*, in the sense that any y either does belong or does not belong to $\tilde{A}$. However, if it is not possible to identify a y^* of the kind described above, then a boundary of $\tilde{A}$ is not well defined. Consequently, we are unable to classify the membership of certain ys in $\tilde{A}$ with definitiveness (or precision). Such ys can *simultaneously* belong and not belong to $\tilde{A}$. Sets which exhibit the property of having boundaries that are not sharp are said to be *fuzzy*. Fuzzy sets do not adhere to the law of the excluded middle. In the context of the scenario considered here, one may not be able to classify, with definiteness, certain defects as being anomalies. That is, there could arise, in practice, scenarios in which there is an uncertainty (in a subject matter specialist's mind) about classifying a defect as being an anomaly or not, and also an unwillingness (of the specialist) to assign probabilities to the uncertainty of classification.

To summarize, fuzzy sets are those whose boundaries are not well defined, and imprecision pertains to an inability to place with certainty every element of a set, such as $\mathcal{R}$, into its fuzzy subset such as $\tilde{A}$. That is, imprecision is a consequence of vagueness.

The Kolmogorov axiomatization of probability is developed on the premise that probability measures be defined on sharp sets ((c.f. Billingsley (1985), p. 20)). Thus, the appearance of fuzzy sets requires of us ways to develop approaches whereby probabilities can be endowed to fuzzy sets as well. A strategy for doing so is via the introduction of "membership functions" which, though not probabilistic, can be seen as a subject matter specialist's classification "probabilities." Membership functions are discussed in Section 4.2 and their use for inducing probabilities on fuzzy sets discussed in Section 4.3. As a final reminder, it is important to keep in mind that the material of Sections 4.2 and 4.3 will not come into play if the event "anomaly" can be well defined.

4.2. *The membership function of a fuzzy set*

The *membership function* of a fuzzy set $\tilde{A}$ encapsulates the degree to which any $y \in \mathcal{R}$ belongs to $\tilde{A}$. It is denoted by $\mu_{\tilde{A}}(y)$, for every y. It is important to note that $\mu_{\tilde{A}}(y)$ is not a probability, because $\sum_y \mu_{\tilde{A}}(y)$ need not be one; however, it is often the case that $0 \leq \mu_{\tilde{A}}(y) \leq 1$, for all y. Operations with fuzzy sets, such as unions, intersections and complements are facilitated by the membership function. Like probability, the membership function is subjectively specified, and may change from person to person. The membership function of a crisp set is an identity function; i.e., if $\tilde{A}$ is a crisp set, then $\mu_{\tilde{A}}(y) = 0$ for $y < y^*$ and $\mu_{\tilde{A}}(y) = 1$, otherwise. For the scenario of anomalies considered here, with y encapsulating the magnitude of a defect, $\mu_{\tilde{A}}(y)$ would be of the form illustrated in Fig. 3. Small values of y would certainly not be viewed as an anomaly and large values certainly would. For the intermediate values of y, $\mu_{\tilde{A}}(y)$ shows the extent to which y would be judged (by one particular individual) to be an anomaly.

4.3. *Endowing probabilities to fuzzy sets*

By endowing probabilities to fuzzy sets we mean assessing *our* personal probability that Y belongs to $\tilde{A}$ in the light of the membership function $\mu_{\tilde{A}}(y)$. For this we first need to assess our personal probability that Y reveals itself as y—that is our probability that the outcome of Y is y—and our personal probability that the revealed y belongs to $\tilde{A}$. Supposing Y to take discrete values, we denote the above personal probabilities by $P_y(Y = y)$ and $P_y(y \in \tilde{A})$ respectively. The need for this latter probability

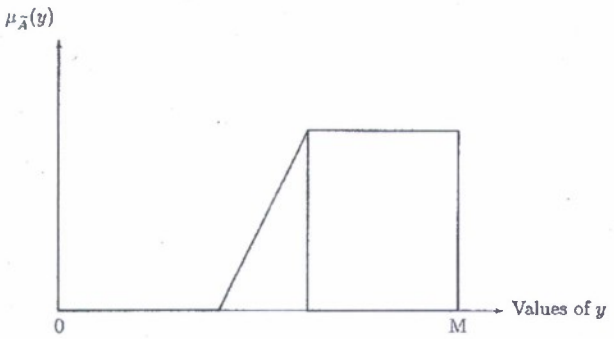

Fig. 3. Membership function of a fuzzy set $\tilde{A}$.

entails a philosophical argument whose roots can be traced to Laplace. By interpreting $\mu_{\tilde{A}}(y)$ as a likelihood function and invoking Bayes' law, Singpurwalla and Booker (2004) go through some standard technical manipulations to evaluate the constants of proportionality and to argue that:

$$P_y(Y \in \tilde{A}; \mu_{\tilde{A}}(y)) = \sum_y \left[1 + \frac{1 - \mu_{\tilde{A}}(y)}{\mu_{\tilde{A}}(y)} \times \frac{P_Y(y \notin \tilde{A})}{P_Y(y \in \tilde{A})}\right]^{-1} P_y(Y = y). \quad (17)$$

See Equation (10) of Singpurwalla and Booker (2003).

4.4. *Assessing failure probability with imprecisely specified anomalies*

With Equation (17) in place, it is a relatively straightforward matter to obtain an analogue of the postmortem probability when the classification of anomalies is imprecise, as

$$P(Y \in \tilde{A} | X; \mu_{\tilde{A}}(y)) = \sum_y \left[1 + \frac{1 - \mu_{\tilde{A}}(y)}{\mu_{\tilde{A}}(y)} \times \frac{P(y \notin \tilde{A})}{P(y \in \tilde{A})}\right]^{-1} P(Y = y | X), \quad (18)$$

where for convenience the subscripts associated with all the Ps have been omitted. The key difference between Equations (17) and (18) is in the last term. The former entails an unconditional probability for Y; the latter, a conditional probability that Y reveals itself as y, given X, the disposition of an item's status—surviving or failed. Note that $P(Y = y \mid X)$ is like the postmortem probability of Section 3.1, save for the fact that Y can now take a range of values y, instead of it being zero or one.

To assess an item's survival probability were an imprecisely specified anomaly to be declared as $Y \in \tilde{A}$, we consider the analogue of Equation (13). Specifically, we have:

$$P(X = 1 | Y \in \tilde{A}; \mathcal{H}) \propto P(Y \in \tilde{A} \mid X = 1; \mu_{\tilde{A}}(y)) \times p, \quad (19)$$

where the middle term is given by Equation (18), and as before, p is our prior probability that $(X = 1)$.

Equation (19) forms the basis of assessing the item's survival probability when the presence of an anomaly is actually declared, but not the extent of the defect that is believed to result in an anomaly. That is, we are not given the value of y. In this case $P(Y \in \tilde{A} | X = 1; \mu_{\tilde{A}}(y))$ is viewed as the likelihood and the left-hand side of Equation (19) becomes $P(X = 1; Y \in \tilde{A}, \mathcal{H})$, the required probability. Consequently, Equation (19) leads us to

$$P(X = 1; Y \in \tilde{A}, \mathcal{H}) \propto \mathcal{L}(X = 1; Y \in \tilde{A}, \mu_{\tilde{A}}(y)) \times p, \quad (20)$$

which is our personal probability that $(X = 1)$, given the presence of an anomaly that is vaguely specified.

5. A reason to believe

Sections 3 and 4 required of us the specification of a conditional probability $P(Y = y \mid X = x; \mathcal{H})$ and the membership function $\mu_{\tilde{A}}(y)$, $y \in [0, M]$, as a way of dealing with vagueness and anomalies. What if vagueness and other reasons create an *unwillingness* to specify the conditional probability but a willingness to specify a marginal probability $P(Y = y; \mathcal{H})$?

The notion of "belief" was introduced by Dempster (1967) as a way of dealing with such partial specifications. Dempster's development is articulated via a key feature of axiomatic probability theory, namely, that in order to induce probability measures from a probability measure space to another measure space it is necessary that the mapping from the former to the latter be a many-to-one map. As an example, a random variable is a many-to-one map. Consequently, its probability distribution function can be induced from the probability measure space on which the random variable is defined. When the mapping is a one-to-many map—as is the case with our anomaly (see Fig. 2)—the induced measure will no more be a probability measure. For a more detailed appreciation of this argument, we refer the reader to Wasserman's (1990) excellent exposition; parts of it are reproduced in the Appendix. The induced measure not being a probability measure, alternate labels for it become germane. Dempster's choice of a label is *Basic Probability Assignment* (BPA).

With respect to the problem at hand, suppose that we are able to elicit personal probabilities of the type $P(Y = y; \mathcal{H})$, $y = 1$ or 0, as p_a and $(1 - p_a)$ respectively. Given p_a, and the mapping of Fig. 2, how may we describe our uncertainty about the survival (or failure) of the item to time t? That is, how may we express our uncertainty about the event $(X = x)$ for $x = 1$ or 0?

The "belief function" approach of Dempster starts by noting that the mapping from $Y = y$ to $X = x$ is a one-to-many map. In particular, if Γ denotes the mapping from the Y-space to the X-space, then $\Gamma(Y = 1) = \{X = 1, X = 0\}$. That is, the singleton $(Y = 1)$ maps into the set $\{X = 1, X = 0\}$ via the map Γ; in other words, Γ is a *set-valued* map,

similarly with $\Gamma(Y = 0)$. However, in order to make the essence of our development more transparent, we suppose that $\Gamma(Y = 0) = (X = 1)$. This means that the absence of an anomaly is tantamount to the item's success. In other words, the mapping from $Y = 0$ to the X-space is a one-to-one map. Consequently, in Fig. 2, the arc joining the nodes $(Y = 0)$ and $(X = 0)$ needs to be removed.

With the above in place, the next step in the development of the belief function approach is to induce measures of uncertainty from the Y-space to the X-space. Recall, that it is only the Y-space that has been endowed with probability as the measure of uncertainty. Since the X-space has only two elements, $(X = 1)$ and $(X = 0)$, $\mathcal{F}(X)$, the measure space (i.e., the set of all sets) generated by X, has four elements, namely:

$$\mathcal{F}(X) = \{\{\phi\}, \{X = 1\}, \{X = 0\}, \{X = 1, X = 0\}\}.$$

With $\Gamma(Y = 1) = \{X = 1, X = 0\}$ and $\Gamma(Y = 0) = (X = 1)$, the induced measure, say m, on $\mathcal{F}(X)$ will be of the form: $m(\phi) = 0$, $m(X = 1) = P(Y = 0) = 1 - p_a$, $m(X = 0) = 0$ and $m\{X = 1, X = 0\} = P(Y = 1) = p_a$. Recall that in Dempster's terminology, the $m(\bullet)$s constitute a BPA. It is easy to verify that m possesses the following two properties: $m(\phi) = 0$, and for $F \in \mathcal{F}(X)$, $\sum_{F \in \mathcal{F}(X)} m(F) = 1$. However, m is not countably additive and thus is not a probability measure. To make m a probability measure we should be prepared to apportion p_a between the events $(X = 1)$ and $(X = 0)$.

Once the BPAs are in place, the *belief function* induced by the map Γ on $\mathcal{F}(X)$ is defined , for any $F, G \in \mathcal{F}(X)$ as

$$\text{bel}(F) = \sum_{G \subseteq F} m(G),$$

and bel(F) is then considered as a quantified measure of uncertainty about F. Thus for our problem at hand $\text{bel}(X = 1) = 1 - p_a$, whereas $\text{bel}(X = 0) = 0$; also, $\text{bel}\{X = 1, X = 0\} = 1 - p_a$.

Dempster has also introduced the dual of the belief function, called the *plausibility function*, where for any $F \in \mathcal{F}(X)$:

$$\text{pl}(F) = 1 - \text{bel}(F^c);$$

F^c is the complement of F. For our problem at hand $\text{pl}(X = 1) = 1$, whereas $\text{pl}(X = 0) = p_a$.

To make these ideas operational, that is, to make a pragmatic use of them, we need to interpret bel($\bullet$) and pl($\bullet$). Using bets, $\text{bel}(X = 1)$ is the most you are willing to pay for a bet on $(X = 1)$: if $\text{bel}(X = 1) = 1 - p_a$, you are willing to pay at most $1 - p_a$ to receive one monetary unit if $(X = 1)$. $\text{pl}(X = 1)$ is (1—the most you are willing to pay for a bet on $(X = 1)^c$): if $\text{pl}(X = 1) = 1$, you are not willing to pay anything to bet on $(X = 1)^c = (X = 0)$. However, as pointed out by a referee, Walley (1991) has argued that it is misleading to interpret the belief and plausibility functions as betting rates.

5.1. *Summarizing "beliefs"*

By way of a closure, we claim that the notion of belief, or its dual plausibility, comes into play when joint probabilities of the type $P(Y = 1, X = 1; \mathcal{H})$ cannot be elicited, and when the marginal probabilities of the type $P(Y = 1; \mathcal{H}) = p_a$ cannot be apportioned in a one-to-many map. Intuitively, the uncertainty measure bel($\bullet$) seems reasonable; it can be seen as a lower bound on probability. When the mapping under discussion is a one-to-one or a many-to-one, belief and probability agree, and thus the belief function will obey the rules of probability. We may conclude by saying that there is a price to be paid for not being able to elicit the required conditional probabilities, and the price is to forsake the notion of probability and its accompanying virtues. Dempster has also proposed rules for combining uncertainties, the details about which can be found in Shafer (1976) or in Wasserman (1990).

Acknowledgements

Sallie Keller-McNulty (now Dean of Engineering at Rice University) played an instrumental role in sponsoring this work. Nozer D Singpurwalla's research has been suported in part by grants N00014-06-1-037, Office of Naval Research, and W911NF-05-01-2009 by the US Army Research Office.

References

Billingsley, P. (1985) *Probability and Measure*, second edition, Wiley, New York, NY.

De Finetti, B. (1974) *Theory of Probability*, Wiley, New York.

Dempster, A.P. (1967) Upper and lower probabilities induced for a multivariate mapping. *Annals of Mathematical Statistics*, **38**, 325–339.

Good, I.J. (1965) *The Estimation of Probabilities*, The MIT Press, Cambridge, MA.

Lindley, D.V. (1982) Scoring rules and the inevitability of probability. *International Statistical Review*, **50**, 1–26.

Lindley, D.V. and Phillips, L.D. (1976) Inference for a Bernoulli process (a Bayesian View). *The American Statistician*, **30**, 112–119.

Shafer, G. (1976) *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, NJ.

Singpurwalla, N.D. and Booker, J.M. (2004) Membership functions and probability measures of fuzzy sets. *Journal of the American Statistical Association*, **99**(467), 867–877.

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London.

Wasserman, L.A. (1990) Belief functions and statistical inference. *The Canadian Journal of Statistics*, **18**(3), 183–196.

Zadeh, L.A. (1978) Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, **1**, 3–28.

Appendix

Belief and plausibility

In order to gain an appreciation of the notion of "belief" and its dual "plausibility," it is best that we start off with a

look at the essentials of how to measure theoretic probability. This we do below via the following seven steps, each of which serves as a prelude to the next step. We assume of the reader some familiarity with these steps. From Step 8 and onwards, our discussion highlights arguments necessary to motivate the notions of belief and plausibility.

Step 1. Let $(\Omega, \mathcal{F}(\Omega), \mu)$ be a probability measure space, with ω as an element of Ω, and μ assessed for all members A of $\mathcal{F}(\Omega)$.

Step 2. Let $(X, \mathcal{F}(X))$ be some measure space with x as an element of X. This is our space of interest.

Step 3. Let $B \subset X$; since $\mathcal{F}(X)$ is a σ-field generated by X, $B \in \mathcal{F}(X)$.

Step 4. Our aim is to endow the space $(X, \mathcal{F}(X))$ with a measure that encapsulates our uncertainty about any B, where $B \subset X$, or about a singleton x, where $x \in X$, should X have countable elements. Ideally, our measure of uncertainty should be a probability.

Step 5. The measure that we endeavor to endow $(X, \mathcal{F}(X))$ with, should bear some relationship to the measure μ. This is because we have been able to assess probabilities on the space $(\Omega, \mathcal{F}(\Omega))$; i.e., we are prepared to place bets only on members of $\mathcal{F}(\Omega)$.

Step 6. In order to be able to do the above, we should connect the spaces $(\Omega, \mathcal{F}(\Omega), \mu)$ and $(X, \mathcal{F}(X))$. This connection can be made in several ways, *two* of which are indicated below:

(i) a mapping from Ω as the domain, to X as the range, or

(ii) a mapping from Ω as the domain, to $\mathcal{F}(X)$ as the range.

Step 7. The standard approach is 6 (i) above; this is what leads us to the notion of a *real-valued random variable*, say Z.

Specifically, we take X to be the real line $\mathbb{R}$, or a countably infinite set of integers $I = \{0, \pm 1, \pm 2, \ldots\}$, or a countably finite set of integers $I_N = \{0, \pm 1, \ldots, \pm N\}$. When $X = \mathbb{R}$, $\mathcal{F}(X) = \mathcal{B}(X)$—the Borel sets of $\mathbb{R}$. When $X = I_N$, then $\mathcal{F}(X)$ is the power set of I_N.

Suppose that $X = \mathbb{R}$. Then Z is a mapping with domain Ω and range $\mathbb{R}$. Furthermore, Z is a *many-to-one* map from Ω to $\mathbb{R}$. Specifically, for every $\omega \in \Omega$, there is one and only one $Z(\omega)$, and $Z(\omega) \in \mathbb{R}$. However, we do allow for the possibility that for any two (or more) $\omega_1, \omega_2 \in \Omega$, $Z(\omega_1) = Z(\omega_2)$.

Now, a (fortunate) consequence of the many-to-one map Z is that such a map is able to induce a probability measure, say μ^*, on $(X, \mathcal{F}(X))$ (or to put it more correctly on $(\mathbb{R} < \mathcal{F}(\mathbb{R}))$). Specifically, for any $a \in \mathbb{R}$, the set $(Z(\omega) \leq a) \in \mathcal{F}(X)$, and

$$\mu^*(Z(\omega) \leq a) = \mu\{\omega \in \Omega : Z(\omega) \leq a\},$$

is a probability measure of the set $(Z(\omega) \leq a)$. Consequently, we now have a probability measure space $(X, \mathcal{F}(X), \mu^*)$ in addition to our original probability measure space $(\Omega, \mathcal{F}(\Omega), \mu)$.

Thus with a many-to-one map, we are able to describe our uncertainties about events of interest in $\mathcal{F}(X)$ via a probability μ^*, with μ^* being based on μ.

Step 8. Suppose now that the connection between the spaces $(\Omega, \mathcal{F}(\Omega), \mu)$ and $(X, \mathcal{F}(X))$ is established via a mapping Γ whose domain is Ω (as before) but whose range is $\mathcal{F}(X)$ instead of X. That is, $\Omega\Gamma \rightarrow \mathcal{F}(X)$. More specifically, for every $\omega \in \Omega$, $\Gamma(\omega) = B$, where $B \in \mathcal{F}(X)$.

If we assume that the above mapping is many-to-one, in the sense that every $\omega \in \Omega$ gets mapped to one and only one set B (where B may or may not be a singleton), then this mapping is known as a many-to-one *set-valued* map. When such is the case Γ is also able to induce a probability measure, say μ^{**}, on the space $(\mathcal{F}(X), \mathcal{F}(\mathcal{F}(X)), \mu^{**})$, where $\mathcal{F}(\mathcal{F}(X))$ is a σ-field of sets generated by $\mathcal{F}(X)$. Consequently, for any set $C \in \mathcal{F}(\mathcal{F}(X))$:

$$\mu^{**}(C) = \mu\{\omega \in \Omega : \Gamma(\omega) = C\}.$$

Thus, to summarize, a many-to-one set-valued map is also able to induce a probability measure μ^{**} on the space $(\mathcal{F}(X), \mathcal{F}(\mathcal{F}(X)))$, assuming that the latter space is of interest to us. But what about the space $(X, \mathcal{F}(X))$? This after all, is our space of interest.

Step 9. The fact that Γ is a many-to-one set-valued map on $\mathcal{F}(X)$ is tantamount to the fact that Γ is a *many-to-many* point-valued map on X. In particular, if $X = \mathbb{R}$ and $\mathcal{F}(X) = \mathcal{B}(\mathbb{R})$, then Γ is a many-to-many real-valued map on $\mathbb{R}$. Consequently, for every $\omega \in \Omega$, $\Gamma(\omega)$ can take any and *all* values in an interval, say $\mathcal{I}$, where $\mathcal{I} \in \mathcal{B}(\mathbb{R})$. Inducing a probability measure on $\mathcal{I}$ or any subset of $\mathcal{I}$ boils down to *smearing* $\mu(\omega)$, the probability measure on ω, over $\mathcal{I}$. How should this measure be smeared? What if one is unwilling to specify a strategy for smearing (or distributing) $\mu(\omega)$ over $\mathcal{I}$? When such is the case we are unable to induce a probability measure from the space $(\Omega, \mathcal{F}(\Omega), \mu)$ to $(X, \mathcal{F}(X))$. As a consequence, an alternative measure called *plausibility*, abbreviated pl(•), has been proposed on $\mathcal{F}(X)$. But before examining pl(•), it may be useful to better articulate this matter of smearing $\mu(\omega)$ by looking at a special case of $\mathcal{I}$, namely an $\mathcal{I}$ consisting of a countable number of elements, say two; denote these by $\{x_1, x_2\}$. Suppose that $\Gamma^{-1}\{x_1, x_2\} = \omega$; then $\mu(\omega)$ is the induced probability measure of $\{x_1, x_2\}$. However, to induce a probability measure on x_1 or x_2, we

need to split (apportion) $\mu(\omega)$ in some logical and meaningful manner.

To summarize, whenever the map connecting two measure spaces is a many-to-one set-valued map, or a many-to-one point-valued map, a probability measure can always be induced from the domain space to the range space. Probability measures cannot be induced when the mapping is a one-to-many, or a many-to-many, point-valued map, unless additional assumptions are made. When such assumptions cannot be made, a compromise has to be struck and upper and lower probabilities enter the foray of uncertainty assessment. These are discussed below.

Step 10. Consider the subset B of X. Suppose that there does *not* exist an induced probability measure from $(\Omega, \mathcal{F}(\Omega), \mu)$ to B. That is, $\nexists$ and $\omega \in \Omega$, such that $\Gamma(\omega) = B$.

Now consider a set $C \in \mathcal{F}(\mathcal{F}(X))$ with the feature that $C \cap B \neq \phi$; suppose that C is the only set in $\mathcal{F}(\mathcal{F}(X))$ that intersects with B. Since $C \in \mathcal{F}(\mathcal{F}(X))$, $\mu^{**}(C)$ is known. Let $\omega_1, \omega_2, \ldots, \omega_n$ be such that $\Gamma(\omega_i) = C$, $i = 1, \ldots, n$. Then, the plausibility of B, denoted $\text{pl}(B)$ is the (probability) measure $\text{pl}(B) = \mu\{\omega_1, \ldots, \omega_n\}$. Alternatively put

$$\text{pl}(B) = \mu\{\omega \in \Omega; \Gamma(\omega) = C \text{ and } B \cap C \neq \phi\}.$$

The above expression generalizes when more than one set intersects B. For example, suppose that $B \cap C_i \neq \phi$, for $i = 1, \ldots, k$, with $C_i \in \mathcal{F}(\mathcal{F}(X))$. Then:

$$\text{pl}(B) = \mu\{\omega \in \Omega; \Gamma(\omega) = C_i \text{ and } B \cap C_i \neq \phi,\ i = 1, \ldots, k\}.$$

Since there are several sets C_i that intersect with B, there are overlapping ωs in the definition of $\text{pl}(B)$. Consequently, it is also called an "upper probability."

Step 11. A notion dual to $\text{pl}(\bullet)$—in a sense to be explained later—is $\text{bel}(\bullet)$; here

$$\text{bel}(B) = \mu\{\omega \in \Omega; \Gamma(\omega) = C_i,\ C_i \subset B,\ i = 1, \ldots, k\}.$$

$\text{Bel}(B)$ is a lower probability, with $0 \leq \text{bel}(B) \leq \text{pl}(B) \leq 1$. Also, $\text{bel}(B) = 1 - \text{pl}(B^c)$.

The measures $\text{pl}(\bullet)$ and $\text{bel}(\bullet)$ are not probability measures in the sense that:

$$\text{bel}(A \cup B) \geq \text{bel}(A) + \text{bel}(B);$$

i.e., because of an overlap of ωs, $\text{bel}(\bullet)$ is superadditive.

Biographies

Nozer D Singpurwalla is a Professor of Statistics and a Professor of Decision Sciences at The George Washington University. He is a Fellow of The American Statistical Association, The Institute of Mathematical Statistics and The American Association for the Advancement of Sciences. He has authored several papers in reliability, statistics, quality control and related topics and has written three books on these subjects, the last one being *Reliability and Risk: A Bayesian Perspective*, published by Wiley (in the UK) in 2006.

Alyson Wilson is an Associate Professor in the Department of Statistics at Iowa State University. Prior to her move to Iowa State, she was a Project Leader in the Statistical Sciences Group at Los Alamos National Laboratory. Her research focuses on Bayesian methods, with emphasis on reliability, complex systems and data integration. These interests developed out of the challenge of science-based stockpile stewardship, LANL's approach to understanding and certifying the reliability and performance of the enduring US nuclear stockpile without full-system testing. Prior to her move to Los Alamos, she was a senior operations research systems analyst working in support of the US Army Operational Evaluation Command, Air Defense Artillery Evaluation Directorate. She also spent two years at the National Institutes of Health performing research in the biomedical sciences. She is currently the Chair of the American Statistical Association Section on Statistics in Defense and National Security. She received her Ph.D. in Statistics from the Institute of Statistics and Decision Sciences at Duke University.

International Statistical Review (2008), 76, 2, 247–267 doi:10.1111/j.1751-5823.2008.00049.x

Many-valued Logic in Multistate and Vague Stochastic Systems

Kimberly F. Sellers[1] and Nozer D. Singpurwalla[2]

[1]*306 St. Mary's Hall, Department of Mathematics, Georgetown University, Washington DC 20057, USA. E-mail: kfs@georgetown.edu*
[2]*Department of Statistics, The George Washington University, 2140 Pennsylvania Avenue, NW, Washington DC 20052, USA*

Summary

The state of the art in coherent structure theory is driven by two assertions, both of which are limiting: (1) all units of a system can exist in one of two states, failed or functioning; and (2) at any point in time, each unit can exist in only one of the above states. In actuality, units can exist in more than two states, and it is possible that a unit can *simultaneously* exist in more than one state. This latter feature is a consequence of the view that it may not be possible to precisely define the subsets of a set of states; such subsets are called *vague*. The first limitation has been addressed via work labeled 'multistate systems'; however, this work has not capitalized on the mathematics of many-valued propositions in logic. Here, we invoke its truth tables to define the structure function of multistate systems and then harness our results in the context of vagueness. A key contribution of this paper is to argue that many-valued logic is a common platform for studying both multistate and vague systems but, to do so, it is necessary to lean on several principles of statistical inference.

Key words: Consistency profile; likelihood function; membership functions; reliability; probability; maintenance management; natural language; degradation modeling; decision making and utility.

1 Introduction and Overview

The calculus of coherent systems, innovated by Birnbaum *et al.* (1961) has served as a mathematical foundation for a theory of systems. Here, one explores the effect that a system's components have on the system. The bulk of the effort, however, has been devoted to the case of binary states with precise classification. That is, the components and the system can (at any point in time) be in one of two unambiguously defined states, functioning or failed. In actuality, items can function in degraded states, and these could be a discrete set or a continuum of states. An example of the former is a load-sharing system, like a transmission line for power with r strands. As the strands break, the rope transitions from its ideal load carrying capability to its complete disintegration (Smith, 1983). An example of the latter is a precipitator for reducing air pollution whose cleaning efficiency ranges from (almost) 100 to 0% (Matland & Singpurwalla, 1981). Systems that can exist in more than two states are called *multistate systems*.

There are two interrelated aims to this paper. The first is to contribute to the mathematics of multistate systems with precise classification via many-valued logic. To set the stage for this, we overview some key notions and results in the reliability theory of binary systems.

Section 1.2 is archival; however, Section 1.3 is current in the sense that it incorporates the view that, when discussing system reliability, one needs to distinguish between probability (which is personal) and propensity (which is physical), and that the assumption of the independence is conditional upon propensities. The second aim of this paper is to argue that multivalued logic also provides a framework for assessing the reliability of binary or multistate systems with imprecise classification. Imprecision (or vagueness) is articulated in Section 1.4; Section 1.5 is a guide to the rest of this paper.

1.1 Preamble: Notation and Terminology

Consider a system with n components. The system and each of its components can exist in several states in $\mathcal{S} \subseteq [0, 1]$. Let X_i, $i = 1, \ldots, n$ denote the state of component i at time $\tau \geq 0$, and denote $\mathbf{X} = (X_1, \ldots, X_n)$. Binary systems are those for which $\mathcal{S} = \{0, 1\}$, where 1 (0) denotes a functioning (failed) state. The state of the system is a function of $\mathbf{X}$, called the 'structure function'. We denote by $\phi(\mathbf{X})$ the structure function for a binary system. The structure function for a system with multiple states will be denoted by $\psi(\mathbf{X})$. We assume that the component and system states belong to the same set $\mathcal{S}$; e.g. $X_i \in \mathcal{S}$ and $\phi(\mathbf{X}) \in \mathcal{S}$. However, it is possible that the X_i's belong to [0,1] whereas $\phi(\mathbf{X})$ can only take values in $\{0,1\}$.

1.2 The Calculus of Binary Systems with Precise Classification

The following is an overview of the calculus of binary systems (Barlow & Proschan, 1975); we generalize this construction in Sections 3 and 4. Let $\mathcal{S} = \{0, 1\}$ with $X_i = 1$ (0) if component i functions (fails), $i = 1, \ldots, n$; similarly, $\phi(\mathbf{X}) : \mathcal{S}^n \to \mathcal{S}$ equals 1 (0) if the system functions (fails). ϕ is a binary coherent system if (1) ϕ is non-decreasing in each argument of $\mathbf{X}$, and (2) each component is relevant. Examples of binary coherent systems are a series system, a parallel redundant system, and a k-out-of-n system. The dual of a binary coherent system $\phi(\mathbf{X})$ is defined as $\phi^D(\mathbf{X}) = \mathbf{1} - \phi(\mathbf{1} - \mathbf{X})$, where $\mathbf{1} - \mathbf{X} = (1 - X_1, 1 - X_2, \ldots, 1 - X_n)$. Any binary structure function ϕ with n components can be decomposed as $\phi(\mathbf{X}) = X_i\, \phi(1_i, \mathbf{X}) + (1 - X_i)\, \phi(0_i, \mathbf{X})$, for all $\mathbf{X}$, $i = 1, \ldots, n$; this is later referred to as the *pivotal decomposition*. The following notation, definitions and theorems are conventional (Barlow & Proschan, 1975):

$$\mathbf{X} \cdot \mathbf{Y} = (X_1 \cdot Y_1,\ X_2 \cdot Y_2,\ \ldots,\ X_n \cdot Y_n),$$
$$\mathbf{X} \amalg \mathbf{Y} = (X_1 \amalg Y_1,\ X_2 \amalg Y_2,\ \ldots,\ X_n \amalg Y_n),$$

where $X_i \amalg Y_i = 1 - (1 - X_i)(1 - Y_i), i = 1, 2, \ldots, n$.

THEOREM 1: *For any binary coherent system* ϕ, $\phi_S(\mathbf{X}) \stackrel{\text{def}}{=} \prod_{i=1}^n X_i \leq \phi(\mathbf{X}) \leq \coprod_{i=1}^n X_i \stackrel{\text{def}}{=} \phi_P(\mathbf{X})$.

THEOREM 2: *For any binary coherent system* ϕ,

$$\phi(\mathbf{X} \amalg \mathbf{Y}) \geq \phi(\mathbf{X}) \amalg \phi(\mathbf{Y}) \tag{1}$$

and

$$\phi(\mathbf{X} \cdot \mathbf{Y}) \leq \phi(\mathbf{X}) \cdot \phi(\mathbf{Y}), \tag{2}$$

with equality holding in equation (1) *(equation 2) if and only if the structure function* ϕ *is* $\phi_P(\phi_S)$. *Proofs of Theorems 1 and 2 can be found in Barlow & Proschan* (1975).

1.3 Reliability of Binary Systems

Suppose that the X_i's are *exchangeable*, and that p_i is the *propensity* of X_i being 1; that is, $p_i = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n}$ [cf. Lindley & Singpurwalla (2002) or Spizzichino (2001)]. Then, conditional on p_i, our subjective probability that $X_i = 1$ is p_i, $i = 1, \ldots, n$. Unconditionally, $P(X_i = 1) = \int_0^1 p_i \pi(p_i) \mathrm{d}p_i = \mathrm{E}(p_i)$, where $\pi(p_i)$ encapsulates our uncertainty about the propensity p_i; i.e. $\pi(p_i)$ is our subjective probability of p_i. The notions of propensity and subjective probability are articulated in de Finetti's theorem on exchangeable Bernoulli sequences; see Lindley & Phillips (1976).

Much of the literature on the reliability of binary coherent systems is conditional on p_i. An exception is Lynn *et al.* (1998), in which the analysis is based on averaging out $p_1, \ldots, p_n$ with respect to a joint distribution.

Conditional on $\mathbf{p} = (p_1, \ldots, p_n)$, the reliability of the system is a function of $\mathbf{p}$, say $h(\mathbf{p})$, but only if the X_i's are (conditionally) independent; i.e. (1) given $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, X_i and X_j are independent, $\forall\, i \neq j$, and (2) given p_i, X_i is independent of p_j, $\forall\, j \neq i$. Consequently, $P(\phi(\mathbf{X}) = 1 \mid \mathbf{p}) = E(\phi(\mathbf{X}) \mid \mathbf{p}) = h(\mathbf{p})$.

Analogues of the pivotal decomposition and Theorems 1 and 2 follow, asserting that the reliability of any binary coherent system is bounded below (above) by that of a series (parallel) system, if the X_i's are conditionally (given $\mathbf{p}$) independent, and redundancy at the component level is superior to redundancy at the system level when the systems are connected in parallel; vice versa if in series; see Barlow & Proschan (1975).

1.4 Vagueness or Imprecision

For purposes of discussion, consider a generic element of $\mathcal{S} = [0, 1]$, say x. At any point, we may be able to inspect the system and declare that $\psi(\mathbf{X}) = x$. If we are able to place this x in a well-defined subset of $\mathcal{S}$, then we say that the states of the system can be *classified with precision.* There are scenarios, however, where the identification of a state can be done unambiguously, but the classification cannot; this is the case of *classification with 'vagueness'.*

In the context of coherent systems, vagueness is not synonymous with uncertainty of performance. Uncertainty of performance is lack of knowledge about the future state of the system, e.g. will the system be functioning 5 hours from now? Vagueness pertains to uncertainty about classification, i.e. an inability to place any outcome x in a subset of $\mathcal{S}$ because the boundaries of the subset cannot be sharply delineated. Some examples illustrate this point.

Suppose that $\mathcal{S} = \{0, 1, \ldots, 10\}$, with each element representing a state in which the system can exist, ranging from the ideal at 10, to the undesirable at 0. Then what is the subset of 'good states' in $\mathcal{S}$? This subset is not well defined; for example, is 7 a good state? If $\mathcal{S}$ were to be partitioned into 'good' and 'bad' states, such partitioning being a feature of *natural language* (Zadeh, 1965), would 5 qualify as a good state or a bad state? More likely, 5 qualifies as both a good state and a bad state. Thus if $\psi(\mathbf{X}) = 5$, then the state of the system is simultaneously good and bad. As another scenario, consider an automobile that has 3000 miles on it. Should this automobile be classified as a 'new' or a 'used' car? The question of classification arises in the contexts of setting insurance rates, taxation and warranties. The subset of miles that go into classifying a car as being 'new' is not sharply defined; it is *imprecise.* Most cars sold as being new have anywhere from 20 to 100 miles—perhaps even more—on them. In actual practice, decisions are often made on the basis of vague knowledge that is relevant, e.g. decisions about health care, maintenance and replacements (see Section 6). As another illustration, medical treatments are based on classification of 'high blood pressure' or 'bad cholesterol,' and such classifications fluctuate due to the subjectivity of interpretation between 'good' and 'bad'. The

philosopher Black (1939) gives examples from other sciences. Of historical note is the famous example of Schrodinger's Cat [cf. Pagels (1982), p. 125] from quantum physics. Schrodinger's thought experiment pertains to a cat in a sealed radioactive box in outer space which, according to one school of thought, is simultaneously alive and dead. Examples from the statistical sciences wherein vague knowledge is relevant are most likely to arise from the behavioral and social contexts, such as inferences based on political polling, and medical decisions based on a quality of life questionnaire (Cox *et al.*, 1992), wherein responses almost always tend to be vague.

The existing theory of both binary and multistate coherent systems with precise classification as its underlying premise is unable to deal with the types of scenarios mentioned above. Some other concerns have been voiced by Marshall (1994). One idea, namely to classify states by more than one criterion, precedes ours and we applaud him for this foresight; it makes a case for the viewpoint espoused here.

1.5 Overview of Paper

In Section 2, we give a synopsis of many-valued logic to include its connectives of negation, conjunction, disjunction, implication, and equivalence. In Section 3, we extend the material of Section 1.2 to the case of multistate systems; i.e. for those components and systems where S consists of more than two elements. Here, we invoke Lukasiewicz's (1930) many-valued logic to define the structure function of multistate systems, and arrive upon results that are in agreement with those currently available. The material of Section 3 serves two purposes. One, it shows how many-valued logic provides a common platform via which the material on multistate systems can be seen. Second, it sets the stage for developing the material of Sections 4 and 5, which is entirely new. A use of many-valued logic is unlike that used by Baxter (1984), El-Neweihi *et al.* (1978) and Griffith (1980), whose development centres around binary logic.

Sections 4 and 5 pertain to the scenarios wherein the classification of component and system states is vague. In both sections, S consists of two vague subsets, and these serve as an analogue to binary state systems with precise classification. A key tool here is the 'consistency profile' introduced by Black (1939). Zadeh's (1965) 'membership function' parallels the notion of a consistency profile. The harnessing of Lukasiewicz's many-valued logic with Black's consistency profile provides a vehicle for the treatment of vague coherent systems. To do so, however, we need to lean on aspects of statistical inference and the statistical treatment of expert testimonies.

Section 6 relates the material of Sections 4 and 5 to decision making in maintenance management using natural language. Section 7 concludes the paper.

2 Many-valued Logic: An Overview

Binary logic, upon whose foundation the theory of coherent structures has been developed, pertains to propositions that adhere to the 'Law of Bivalence' (or the 'Law of the Excluded Middle'): all propositions are either true or false. Lukasiewicz (1930) recognized the existence of propositions that can be both true and false simultaneously, and thus modified the calculus of binary propositions to develop a calculus of three-valued propositions. Alternatives exist to Lukasiewicz's three-valued logic; however, for us, Lukasiewicz's proposal is most appealing.

It is important to distinguish between the calculus of probability and the calculus of three-valued logic. Probability pertains to the quantification of uncertainty about events (or propositions) that adhere to the Law of Bivalence. Thus we have, as a part of the calculus of probability, the axiom of additivity. On the other hand, the calculus of many-valued logic is based on a rejection of the Law of Bivalence. The two are therefore different constructs.

Table 1

(a) Truth Table for Lukasiewicz's $Y \wedge Z$.

$Y \wedge Z$		Values of Proposition Z		
		0	1/2	1
Values of Proposition Y	0	0	0	0
	1/2	0	1/2	1/2
	1	0	1/2	1

(b) Truth Table for Lukasiewicz's $Y \vee Z$.

$Y \vee Z$		Values of Proposition Z		
		0	1/2	1
Values of Proposition Y	0	0	1/2	1
	1/2	1/2	1/2	1
	1	1	1	1

Consider two propositions Y and Z, each taking one of three values: 0, $\frac{1}{2}$ and 1. The negation of Y is $Y' = 1 - Y$, as proposed by Lukasiewicz (1930). When the proposition Y takes the value 1 (0) in a truth table, it signals the fact that the proposition is true (false) with certainty. Values of Y intermediate to 1 and 0 signal an uncertainty about the truth or the falsity of Y. The value $\frac{1}{2}$ is chosen arbitrarily for convenience; any value between 0 and 1 could have been chosen. The other logical connectives in the three-valued logic of Lukasiewicz are conjunction, disjunction, implication and equivalence, denoted $(Y \wedge Z)$, $(Y \vee Z)$, $(Y \rightarrow Z)$ and $(Y \equiv Z)$, respectively. The truth tables for the first two are given in Table 1, and we refer the interested reader to Malinowski (1993) for further details. Generalizations from the three-valued to the many-valued case to incorporate propositions that are true or false with various degrees of uncertainty are straightforward.

3 Invoking Many-Valued Logic for Multistate Systems

3.1 Introduction

The aim of this section is to generalize the case of binary systems with precise classification to systems that can exist in multiple ($m + 1$ with $m > 1$) states. The states are labeled $\frac{j}{m}$, $j = 0, 1, 2, \ldots, m$, with 1 representing a perfect state and 0, the state of total collapse. The intermittent states of degradation range from $\frac{m-1}{m}$ to $\frac{1}{m}$, where $\frac{1}{m}$ is the state which is penultimate to the total failure of the system. Thus, the range of states now takes the form $\mathcal{S} = \{\frac{j}{m}; j = 0, 1, 2, \ldots, m\}$ and, by allowing m to be infinite, we are able to consider a continuum of degraded states, in which case, $\mathcal{S} \subseteq [0, 1]$. With $\mathcal{S}$ so defined for both the components and the system, what would be the meaningful choices for the structure function when the system has a series, parallel, or k-out-of-n architecture?

In the past, several proposed definitions of multistate systems have been made. An overview of these is in El-Neweihi *et al.* (1978) and in Baxter (1984), which to the best of our knowledge represents the latest endeavors. Considering the fact that these papers appeared over 20 years ago, one may sense that a satisfactory answer to the above question is available. This may not be true, however, because all the proposed approaches reduce to a representation in terms of binary states and, thus, an adherence to binary logic. As an example, Baxter (1984), following Barlow & Proschan (1975), defines the structure function of a multistate system in terms of the system's 'min-path' and 'min-cut' sets, notions which can have an interpretation only within the context of binary systems. By contrast, our proposal here is to use Lukasiewicz's many-valued logic as a basis for defining the structure function of multistate systems.

Lukasiewicz's motivation for introducing a third value, namely $\frac{1}{2}$, and his calculus of three-valued logic was prompted by an uncertainty about the truth or the falsity of a proposition. The number $\frac{1}{2}$ did not reflect—in any sense—a degree of uncertainty. Whereas Lukasiewicz did not appear to have any motivation for his many-valued logic other than the need to generalize, the

degree of uncertainty interpretation provides a vehicle for extending the three-valued logic. With this in mind, we may ask whether Lukasiewicz's calculus can be directly imported to the scenario of multistate systems when the degraded states can be specified with precision? Our examples of Table 1 illustrating the three-valued logic suggest that this can be done. More importantly, our results are consistent with those given in El-Neweihi *et al.* (1978). Consequently, the Lukasiewicz logic can be seen as providing a rationale for the existing results on multistate systems, a rationale that has been missing.

3.2 Definition and Structural Properties

Let X_i denote the state of component i, $i = 1, \ldots, n$, and $\psi = \psi(\mathbf{X})$ the state of the multistate system ; $\mathbf{X} = (X_1, \ldots, X_n)$. The X_i's and $\psi(\mathbf{X})$ take values in $S = \{\frac{j}{m}, j = 0, 1, \ldots, m\}$.

Definition 1: (Griffith, 1980) ψ is a multistate coherent system if

1. ψ is non-decreasing in each argument of $\mathbf{X}$,
2. for each $i = 1, 2, \ldots, n$, there exist states $0 \leq a_i < b_i \leq m$ and a state vector $(\bullet_i, \mathbf{X})$ such that

$$\psi\left(\frac{a_i}{m}, \mathbf{X}\right) < \psi\left(\frac{b_i}{m}, \mathbf{X}\right);$$

that is, each component is relevant, and

3. $\psi(\frac{j}{m}) = \frac{j}{m}$ where $\frac{j}{m} = (\frac{j}{m}, \frac{j}{m}, \ldots, \frac{j}{m})$.

Properties 1 and 3 of Definition 1 are consistent with those of Barlow & Wu (1978), El-Neweihi *et al.* (1978) and Natvig (1982). Property 2 generalizes the notion of relevance.

To use the logic of many-valued propositions for multistate systems, it is necessary to order the state vector $\mathbf{X}$. Since each $X_i \in \{\frac{j}{m}, j = 0, 1, \ldots, m\}$, we order the X_i's by the values they take. Specifically, let $0 \leq X_{(1:n)} \leq X_{(2:n)} \leq \cdots \leq X_{(i:n)} \leq \cdots \leq X_{(n:n)} \leq 1$ denote the ordered vectors, i.e. $X_{(1:n)}$ is the weakest of all the n components and $X_{(n:n)}$ the strongest. Consequently, from Table 1(a), the structure function of a series system is $\psi_S = \min_i X_i = X_{(1:n)}$; that is, the performance of a multistate series system is no better than the performance of its weakest component. If $n = 2$, and if each X_i can take only three values $\{0, \frac{1}{2}, 1\}$ with $\frac{1}{2}$ denoting the degraded state, then Table 1(a) with $Y \wedge Z$ replaced by $\psi_S(\mathbf{X})$ and Y (Z) replaced by $X_1(X_2)$ gives us a table for the states of the system, given the states of the components. Figure 1(a) displays the state of $\phi_S(\mathbf{X}) = \phi_S(X_1, X_2)$ when X_1 and X_2 take binary values, 0 and 1. In contrast, Figure 1(b) shows the behaviour of $\psi_S(\mathbf{X})$ when X_1 and X_2 are allowed to take all values in the unit interval, showing the effect of continuously degrading components on the structure function. Clearly, $\psi_S(\mathbf{X})$ provides more granularity than $\phi_S(\mathbf{X})$.

For a parallel redundant system, $\psi_P(\mathbf{X}) = \max_i X_i = X_{(n:n)}$; see Table 1(b). This suggests that the performance of a multistate parallel system is no worse than the performance of its strongest component. In the three-valued case, the entries of Table 1(b) provide us with a table for the states of the system given the states of the components, when $n = 2$. The state of $\phi_P(\mathbf{X})$ when X_1 and X_2 take binary values, 0 and 1, is displayed in Figure 2(a). In contrast, Figure 2(b) shows the behaviour of $\psi_P(\mathbf{X})$ when X_1 and X_2 take all values in [0,1]. Again, $\psi_P(\mathbf{X})$ provides more granularity than $\phi_P(\mathbf{X})$.

For multistate k-out-of-n systems, we define $\psi_K(\mathbf{X}) = X_{(n-k+1:n)}$; this definition ensures consistency among systems, i.e. n-out-of-n systems are denoted $\psi_S(\mathbf{X})$ and 1-out-of-n systems are denoted $\psi_P(\mathbf{X})$. Interestingly, our set-up and definition of a multistate coherent system

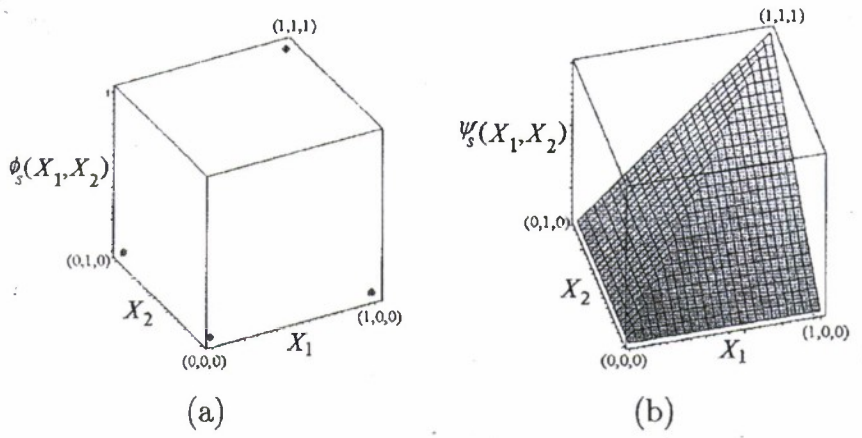

Figure 1. *(a) Two-component binary system, $\phi_S(\mathbf{X})$. (b) Two-component system, $\psi_S(\mathbf{X})$, with continuously degrading components. The coordinates are labeled $(X_1, X_2, \phi_S(\mathbf{X}))$ and $(X_1, X_2, \psi_S(\mathbf{X}))$, respectively.*

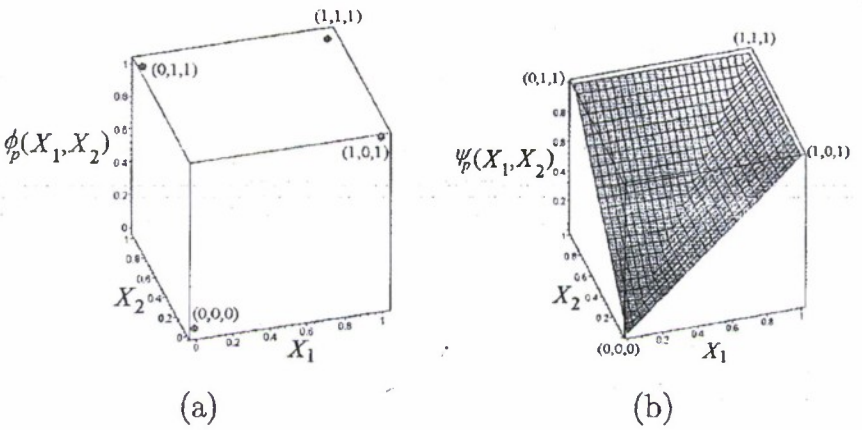

Figure 2. *(a) Two-component binary system, $\phi_P(\mathbf{X})$. (b) Two-component system, $\psi_P(\mathbf{X})$, with continuously degrading components. The coordinates are labeled $(X_1, X_2, \phi_P(\mathbf{X}))$ and $(X_1, X_2, \psi_P(\mathbf{X}))$, respectively.*

permits the definition of a dual of a binary coherent system to hold. The dual of a k-out-of-n system is $\psi_K^D(\mathbf{X}) = \psi_{(n-k+1:n)}(\mathbf{X})$, an $(n-k+1)$-out-of-n system.

In Lemma 1, the pivotal decomposition for binary structure functions is generalized for $(m+1)$ precise categories through consideration of their associated indicator variables.

LEMMA 1: *The following identity holds for every n-component multistate structure function ψ with precise classification: $\psi(\mathbf{X}) = \sum_{j=0}^{m} \psi[(\frac{j}{m})_i, \mathbf{X}] I_{[X_i=\frac{j}{m}]}$, for $i = 1, \ldots, n$ where $I_{[X_i=\frac{j}{m}]} =$* 1(0) if $X_i = \frac{j}{m} (X_i \neq \frac{j}{m})$.

Proof. Any multistate structure function, $\psi(\mathbf{X})$ can be decomposed into a representation that considers the i-th component separately from the remaining $(n-1)$ components. In particular for the multistate component, X_i takes only one value from $\{0, \frac{1}{m}, \frac{2}{m}, \cdots, \frac{m-1}{m}, 1\}$. The result follows.

Theorems 1 and 2 of Section 1 can be generalized for multistate coherent systems. To do so, we introduce the following additional notation. For $\mathbf{X} = \{X_1, \ldots, X_n\}$ and $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$, $\mathbf{X} \leq \mathbf{Y}$ if $X_i \leq Y_i$ for each $i = 1, \ldots, n$. As a generalization of Theorem 1, we have:

THEOREM 3: *Let ψ be a multistate coherent system of order n; i.e. ψ has n components. Then* $X_{(1:n)} \leq \psi(\mathbf{X}) \leq X_{(n:n)}$.

THEOREM 4: *Let ψ be a multistate coherent system of order n. Then*

$$\psi(\mathbf{X} \vee \mathbf{Y}) \geq \psi(\mathbf{X}) \vee \psi(\mathbf{Y}), \tag{3}$$

and

$$\psi(\mathbf{X} \wedge \mathbf{Y}) \leq \psi(\mathbf{X}) \wedge \psi(\mathbf{Y}). \tag{4}$$

The equality in (3) and (4) hold for all $\mathbf{X}$ and $\mathbf{Y}$ if and only if the system's architecture is parallel and series, respectively.

Thus, for a multistate coherent system, equation (3) reiterates the result that, structurally, component-level redundancy is superior to system level redundancy, and vice versa in equation (4). Theorems 3 and 4 and Lemma 1 are also in El-Neweihi & Proschan (1984). They are stated here for completeness.

Since $X_{(1:n)} = \psi_S(\mathbf{X})$ and $X_{(n:n)} = \psi_P(\mathbf{X})$, we have the result that the structure function of any multistate coherent structure is bounded by the structure functions of multistate series and parallel systems.

3.3 Multistate System Reliability under Precise Classification

Suppose that the component state vectors $X_1, \ldots, X_n$ are (conditionally) independent and identically distributed with $P(X_i = \frac{j}{m} \mid \tilde{p}_{j+1}) = \tilde{p}_{j+1}$, for $i = 1, \ldots, n$ and $j = 0, \ldots, m$, where $\tilde{p}_{j+1} \geq 0$ and $\sum_{j=0}^{m} \tilde{p}_{j+1} = 1$. That is, each X_i has a multinomial distribution over $\{\frac{j}{m}; j = 0, 1, 2, \ldots, m\}$ with parameter $\tilde{p}_{j+1}$, $j = 0, \ldots, m$. Let $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_{m+1})$. Clearly for each j, $P(\psi(\mathbf{X}) = \frac{j}{m})$ depends on $\tilde{\mathbf{p}}$ alone, since the X_i's are assumed to be conditionally (given $\tilde{\mathbf{p}}$) independent. Thus, we let $P(\psi(\mathbf{X}) = \frac{j}{m} \mid \tilde{\mathbf{p}}) = h_j(\tilde{\mathbf{p}})$, where h_j is some function of $\tilde{\mathbf{p}}$. Suppose that the architecture of ψ is a $(n - k + 1)$-out-of-n system. Then

$$h_j(\tilde{\mathbf{p}}) = P\left(\psi_{n-k+1}(\mathbf{X}) = \frac{j}{m} \middle| \tilde{\mathbf{p}}\right)$$
$$= \sum_{a=k}^{n} \binom{n}{a} \left\{ \left(\sum_{b=1}^{j+1} \tilde{p}_b \right)^a \left(\sum_{b=j+2}^{m+1} \tilde{p}_b \right)^{n-a} - \left(\sum_{b=1}^{j} \tilde{p}_b \right)^a \left(\sum_{b=j+1}^{m+1} \tilde{p}_b \right)^{n-a} \right\}.$$

Example 1: Let $m, n = 2$. Therefore, we consider a two-component system with three possible states: total failure (0), degradation ($\frac{1}{2}$), and perfect functioning (1), with associated probabilities $\tilde{p}_1$, $\tilde{p}_2$, and $\tilde{p}_3$, respectively. Then, the probability that the parallel system is totally failed is $h_0(\tilde{\mathbf{p}}) = P(\psi_P(\mathbf{X}) = 0 \mid \tilde{\mathbf{p}}) = \tilde{p}_1^2$. i.e. the parallel system is totally failed when all its components are totally failed. The probability that a series system totally fails is $h_0(\tilde{\mathbf{p}}) = P(\psi_S(\mathbf{X}) = 0 \mid \tilde{\mathbf{p}}) = 2\tilde{p}_1\tilde{p}_2 + 2\tilde{p}_1\tilde{p}_3 + \tilde{p}_1^2$; thus, a series system fails completely when at least one component is totally failed.

When $X_1, \ldots, X_n$ are independent but not identically distributed, we may generalize the above properties by introducing $P(X_i = \frac{j}{m} \mid p_{i_{j+1}}) = p_{i_{j+1}}$, $j = 0, \ldots, m$ where for each i, $p_{i_{j+1}} \geq 0$ and $\sum_{j=0}^{m} p_{i_{j+1}} = 1$. We define $\mathbf{p_i} = (p_{i_1}, \ldots, p_{i_{m+1}})$ to be the reliability vector associated with the i-th component and $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_n)$. Given the conditional independence of the X_i's, a

$(n-k+1)$-out-of-n system has

$$h_j(\mathbf{p}) = P\left(\psi_{n-k+1}(\mathbf{X}) = \frac{j}{m} \,\middle|\, \mathbf{p}\right)$$
$$= \sum_a \left(\prod_{i\in J_a} \sum_{b=1}^{j+1} p_{i_b}\right)\left(\prod_{i\in J_a'} \sum_{b=j+2}^{m+1} p_{i_b}\right) - \sum_a \left(\prod_{i\in (J-1)_a} \sum_{b=1}^{j} p_{i_b}\right)\left(\prod_{i\in (J-1)_a'} \sum_{b=j+1}^{m+1} p_{i_b}\right),$$

where J_a is the subset of $(1, 2, \ldots, n)$ where at least k components are performing within level $\frac{j}{m}$ and J_a' is the complement of J_a. Similarly, $(J-1)_a$ is the subset of $(1, 2, \ldots, n)$ where at least k components function within level $\frac{j-1}{m}$ and $(J-1)_a'{}_a$ is the complement of $(J-1)_a$.

Lemma 2 provides the pivotal decomposition for the reliability function, $h_j(\mathbf{p})$.

LEMMA 2: *The following identity holds for the pivotal decomposition of* $h_j(\mathbf{p})$:

$$h_j(\mathbf{p}) = \sum_{a=0}^{m} h_j\left[\left(\frac{a}{m}\right)_i, \mathbf{p}\right] \cdot p_{i_{a+1}}, \quad \textit{for } j = 0, \ldots, m; i = 1, \ldots, n, \tag{5}$$

where $h_j[(\frac{a}{m})_i, \mathbf{p}] = P(\psi(\mathbf{X}) = \frac{j}{m} \mid X_i = \frac{a}{m}, \mathbf{p})$.

Proof. Follows from the Law of Total Probability.

4 Components with Imprecise State Classification

Binary state systems with precise classification were overviewed in Section 1.2, and the concept of vagueness introduced in Section 1.4. Sections 4 and 5 serve to combine these two notions to develop a mechanism for the treatment of vague coherent systems, with Section 4 devoted to the case of components in vague states, and Section 5 to the case of coherent systems in vague states.

The terms 'coherence' and 'vagueness' may seem contradictory; however, they do not pertain to the same object. The first is associated with the truth values of logical connectives, whereas the second pertains to the partitioning of a set into subsets. We start with some background on vagueness and then discuss approaches for quantifying it.

4.1 Vagueness: General Background

Vagueness has been discussed by philosophers like Bertrand Russell, and by physicists like Albert Einstein. To Russell (1923), 'all language is more or less vague' so that the Law of the Excluded Middle 'is true when precise symbols are employed but it is not true when symbols are vague, as, in fact, all symbols are.' Black (1939) recognized the inability of binary logic to satisfactorily represent propositions that are neither perfectly true nor false. He attempted to rectify this by analyzing the concept of vagueness in order to establish an 'appropriate symbolism' by which binary logic can be viewed as a special case. Unlike Lukasiewicz (1930), who was also concerned about the Law of the Excluded Middle, Black did not introduce three-valued propositions. Rather, he defined a vague proposition as one where the possible states of the proposition are not clearly defined with respect to inclusion, and introduced the mechanism of 'consistency profiles' as a way of treating vagueness. Black's consistency profile is a graphical portrayal of the degree of membership of some proposition in a set of imprecisely defined states, with 1 representing absolute membership in a state and 0 an absolute lack of membership. Precise propositions are treated via step functions as consistency profiles, and vague propositions

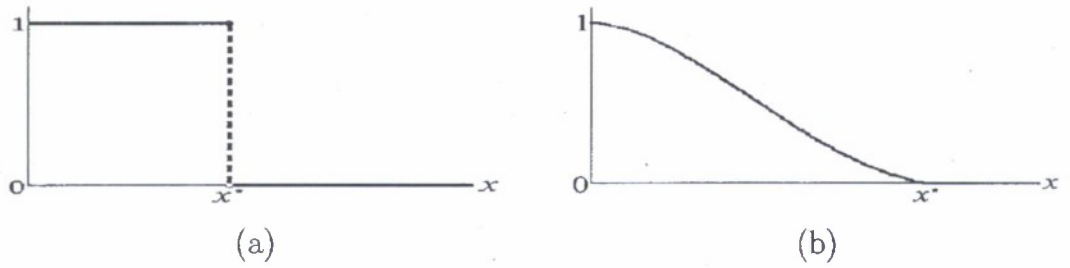

Figure 3. *Example of Consistency Profiles: (a) for a precise set. (b) for a vague set. The consistency profile is 0 after x^*.*

Table 2
Membership table for precise set, A_1, versus fuzzy set, A_2.

x	1	2	3	4	5	6	7	8	9	10
$\mu_{A_1}(x)$	0	0	0	0	0	0	1	1	1	1
$\mu_{A_2}(x)$	0	0	0	0	0	0.2	0.5	0.9	1	1

by consistency profiles that tend gradually from one extreme to another; see Figure 3. The scaling between 0 and 1 is arbitrary; other convenient limits could have been used. Further, the consistency profile which is specified by an individual, or a group of individuals, need not be unique.

4.2 Membership Functions and Probabilities of Fuzzy Sets

Black's (1939) consistency profile precedes Zadeh's (1965) membership function. For each x, a normalized membership function $0 \leq \mu_A(x) \leq 1$ describes a belief of containment of x in a set A. When $\mu_A(x) = 1$ or 0, A is a crisp (or precise) set; when $0 \leq \mu_A(x) \leq 1$, A is a *fuzzy set*. To illustrate the concept of a fuzzy set, consider

Example 2: Let $A_1 = \{x \in \{1, 2, \ldots, 10\} \mid x \geq 7\}$. For any specified x, there is no ambiguity as to whether x belongs to A_1 or not. By definition, $\mu_{A_1}(x) = 1$ when $x = 7, 8, 9$, or 10; otherwise, it is zero (see Table 2). Thus A_1 is a precise set, since $\mu_{A_1}(x) = 1$ or 0. By contrast, consider the set $A_2 = \{x \in \{1, 2, \ldots, 10\} \mid x \text{ is large}\}$. The term 'large' is vague; thus, we cannot precisely ascertain the containment of any x in A_2. A possible membership function for A_2, $\mu_{A_2}(x)$, is given in Table 2; this assignment is not unique.

For fuzzy sets, A and B in a basic set M, with membership functions $\mu_A(x)$ and $\mu_B(x)$ respectively, Zadeh (1965) defined set operations that parallel those of precise sets. For any x in a given basic set M,

1. $\mu_{A \cup B}(x) = \max[\mu_A(x), \mu_B(x)]$,
2. $\mu_{A \cap B}(x) = \min[\mu_A(x), \mu_B(x)]$,
3. $\mu_{A'}(x) = 1 - \mu_A(x)$,
4. $A \subseteq B \Leftrightarrow \mu_A(x) \leq \mu_B(x)$, and
5. $A \equiv B \Leftrightarrow \mu_A(x) = \mu_B(x)$.

Thus, the union of fuzzy sets A and B is the fuzzy set $A \cup B$, whose membership function is max $[\mu_A(x), \mu_B(x)]$; similarly for the intersection and the complement. There is a parallel between operations with fuzzy sets and the conjunction and disjunction connectives of Lukasiewicz (1930). In Section 5.1, we use these operations to define structure functions of vague binary state systems. Thus, we claim that Lukasiewicz's logic provides a unifying framework via which both multistate as well as vague systems can be studied.

4.2.1 Probabilities of fuzzy sets

In the context of this paper, statistical inference plays a key role. This role comes into effect when we endow a probability measure for a fuzzy set, say A. There are two key ideas that drive this development, namely that (1) vague sets are a consequence of one's uncertainty about the boundaries of sharp sets, and (2) the membership function $\mu_A(x)$ is to be interpreted as data (or information) whose role is to help induce a likelihood function, just like the role of an observation in traditional statistical inference. The above ideas can be best exposited by envisioning the scenario of expert testimonies and information integration that has gained current popularity in statistical practice (cf. Reese *et al.* 2004).

Accordingly, we consider the actions of D, an assessor of probabilities (or a decision maker), who quantifies his (her) uncertainty about any outcome of X, say x, being classified in A via a prior probability $\pi_D(x \in A)$. The thesis here is that all uncertainties, including those of classification, be quantified via probability. In order to sharpen the prior probability, D consults an expert, say Z, and elicits from Z a membership function $\mu_A(x)$. This $\mu_A(x)$ can be seen as additional information about the nature of x's membership in A, and de facto serves a role analogous to that of observed data in statistical inference about outcomes. In essence, observed data are evidence about outcomes whereas membership functions are evidence about classification. In principle, D may consult several experts and elicit from each membership functions as a way to further sharpen the analysis.

With $\mu_A(x)$ at hand, D constructs his (her) likelihood function that $x \in A$; we denote this likelihood by $\mathcal{L}[x \in A; \mu_A(x)]$. The construction of this likelihood follows standard statistical procedures for formally incorporating expert testimonies, and should include things such as D's view of the expertise of Z and, in the case of several experts, correlations between them (cf. Lindley, 1991; Clarotti & Lindley, 1988). Since $\mathcal{L}[x \in A; \mu_A(x)]$ is D's likelihood that Z declares $\mu_A(x)$ when $x \in A$, the specification of this likelihood is a subjective exercise on the part of D. Conventionally, in statistical inference, likelihoods for unknown parameters are prescribed via probability models (for outcomes) using the observed data as fixed quantities. By contrast, what we have done here is prescribed a likelihood about classification using the membership as a fixed entity, but without the benefit of a probability model. In so doing, we have interpreted the likelihood in a broader sense, namely as a weighting function (Basu, 1975). In addition to $\mathcal{L}[x \in A; \mu_A(x)]$, D also needs to specify $\mathcal{L}[x \notin A; \mu_A(x)]$, which is D's likelihood that $x \notin A$ when Z declares a $\mu_A(x)$, and $P_D(x)$ which is D's subjective probability that an outcome x will occur. Thus D needs to specify two probability measures $\pi_D(x)$ and $\pi_D(x \in A)$, one for outcomes and one for classification, and two likelihoods, $\mathcal{L}[x \in A; \mu_A(x)]$ and $\mathcal{L}[x \notin A; \mu_A(x)]$.

With the above in place, D uses standard statistical methodology involving Bayes' Law, Bayes' Factors, and prior to posterior odds (cf. Kass, 1993) to obtain a probability measure for a fuzzy set A (cf. Singpurwalla & Booker, 2004) as

$$P_D[X \in A; \mu_A(x)] = \sum_x \left[1 + \frac{\mathcal{L}[x \notin A; \mu_A(x)]}{\mathcal{L} \in A; \mu_A(x)} \cdot \frac{\pi_D(x \notin A)}{\pi_D(x \in A)}\right]^{-1} P_D(x). \tag{6}$$

Equation (6) above is the essence of the material of this section; it is to play a key role in what is to follow. In obtaining the above, we have leaned heavily on the statistical notion of likelihood and the likelihood ratio. Equation (6) simplifies if D chooses to use Z's declared $\mu_A(x)$ as the sole basis for constructing his (her) likelihood, so that $\mathcal{L}[x \in A; \mu_A(x)] = \mu_A(x)$, and $\mathcal{L}[x \notin A; \mu_A(x)] = 1 - \mu_A(x)$. In this case,

$$P_D[X \in A; \mu_A(x)] = \sum_x \left[1 - \left(1 - \frac{1}{\mu_A(x)}\right) \cdot \frac{\pi_D(x \notin A)}{\pi_D(x \in A)}\right]^{-1} P_D(x). \tag{7}$$

4.2.2 The role of precise and fuzzy data in vague systems

In equations (6) and (7), $P_D(x)$ encapsulates D's prior uncertainty about an outcome x. Were D to have at his (her) disposal $\mathbf{x} = (x_1, \ldots, x_n)$, data on X, then $P_D(x)$ would get replaced by a posterior probability, say $P_D(x; \mathbf{x})$. The calculation of this posterior would be a routine exercise were D to invoke a probability model for outcomes, and were the actual observations $x_1, \ldots, x_n$ sharp (i.e. precisely stated). What must D do to update $P_D(x)$ if the data $\mathbf{x}$ is itself fuzzy?

To address this question, we first need to clarify as to what one means by *fuzzy data*, a term that has appeared in several book and article titles; see, for example Bertoluzza *et al.* (2002), and Viertl (2006). If by fuzzy data, we mean imprecision of observation (i.e. observation error), then the treatment of such data can be routinely handled via standard statistical technology, provided that an error distribution can be specified. The literature on 'calibration' adequately deals with this issue; see, for example, Huang (2002). If by fuzzy data, we mean a statement such as 'the outcome does or does not belong to the fuzzy set A', then the incorporation of such information for updating $P_D(x)$ is no more a standard matter. In other words, when the actual value taken by X, say x_i, is not declared, but what is declared is whether the actual value belongs or not to A, an assessment of $P_D(x;$ observed value belongs (does not belong) to A) poses a challenge. This can be addressed if a likelihood for $X = x_i$ with the knowledge that the 'observed value belongs (does not belong) to A' can be specified by D. The specification of such a likelihood will entail several issues such as who provides D the said knowledge, Z or someone other than Z. If it is Z, then $\mu_A(x)$ provides some guidance to D about specifying the likelihood. If it is someone other than Z, then D needs to contemplate the knowledge provider's actions. These and other issues remain to be addressed, including the matter of calibrating Z and updating membership functions.

4.3 Components in Vague Binary States

The notion that units can exist in states that are vaguely defined was introduced in Section 1.4. Specifically, let X denote the state of a component at some time $\tau > 0$, and let X take values in $\mathcal{S} = \{x; 0 \leq x \leq 1\}$, with one representing the perfectly functioning state. Consider $\mathcal{G} \subset \mathcal{S}$, where $\mathcal{G} = \{x; x$ is a 'desirable' state$\}$. Suppose that interest centres around $X \in \mathcal{G}$. Suppose also that we are unable to specify an x^* such that $X \geq x^*$ implies that $X \in \mathcal{G}$ and, otherwise, $X \notin \mathcal{G}$. Thus, the boundary of $\mathcal{G}$ is not sharp; i.e. $\mathcal{G}$ is a fuzzy set. Let $\mu_{\mathcal{G}}(x)$ be the membership function of $\mathcal{G}$. Figure 4 illustrates plausible forms for $\mu_{\mathcal{G}}(x)$. Interest may centre around $\mathcal{G}$ for several reasons, a relevant one being a desire to use 'natural language' for communication with others on matters such as repair and replacement. Another possibility is that it may not be possible to observe the actual value of x, but one may be able to make a general statement about the state of the component.

The complement of $\mathcal{G}$, say $\mathcal{G}^C$, is that fuzzy set whose membership function is $1 - \mu_{\mathcal{G}}(x)$. It is important to note that, if another subset $\mathcal{B} \subset \mathcal{S}$ was defined as $\mathcal{B} = \{x; x$ is an 'undesirable' state$\}$, then $\mathcal{G}^C$ may or may not be $\mathcal{B}$ unless $\mu_{\mathcal{B}}(x)$, the membership function of $\mathcal{B}$, was such that $\mu_{\mathcal{B}}(x) = 1 - \mu_{\mathcal{G}}(x)$. In principle, one is free to choose a $\mu_{\mathcal{B}}(x)$ that need not bear a relationship to $\mu_{\mathcal{G}}(x)$. For example, in Figure 4(a), $\mu_{\mathcal{B}}(x)$ is symmetric to $\mu_{\mathcal{G}}(x)$, whereas in Figure 4(b), $\mu_{\mathcal{B}}(x)$ and $\mu_{\mathcal{G}}(x)$ are not symmetric. There is precedent in the statistical sciences for choosing asymmetric likelihood functions. For example, one need not specify likelihood functions that are symmetrical for competing hypotheses.

Example 3: An assessor D wants to assess the probability that a component will be in a 'desirable' state $\mathcal{G}$ at some future time τ. That is, D wishes to specify $P_D[X \in \mathcal{G}; \mu_{\mathcal{G}}(x)]$, where a membership function of the form $\mu_{\mathcal{G}}(x) = x^4$, $0 \leq x \leq 1$ has been elicited by D from an expert,

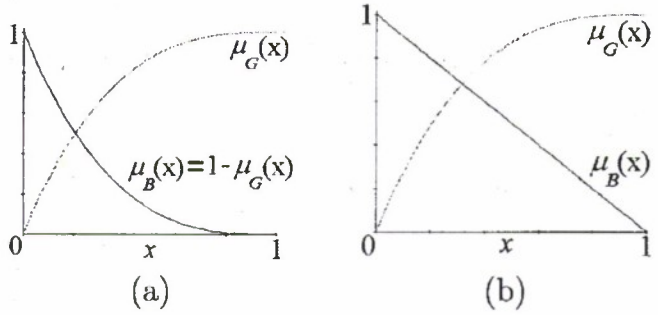

Figure 4. *Membership functions of $\mathcal{G}$ and $\mathcal{B}$: (a) Symmetric case. (b) Asymmetric case.*

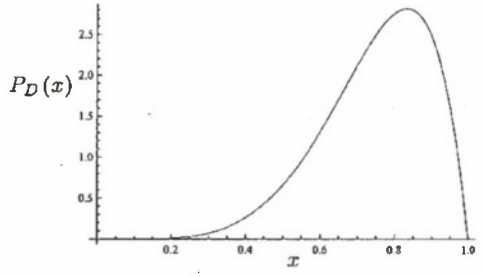

Figure 5. *Component state at time τ, $P_D(x)$.*

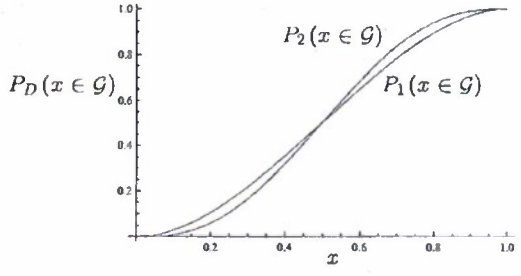

Figure 6. *Two possible prior forms of classifying $X = x$, $P_1(x \in \mathcal{G})$ and $P_2(x \in \mathcal{G})$, supplied by the assessor D.*

Z. Suppose that $P_D(x)$, D's personal probability that the state of the component at time τ will be x is of the form given in Figure 5; it is a Beta(6,2) density. Furthermore, suppose that D's belief that nature will classify any x in $\mathcal{G}$, namely $P_D(x \in \mathcal{G})$, is of the general form illustrated in Figure 6 with the label, $P_1(X \in \mathcal{G})$. Then, it can be seen—via equation (7)—that $P_D[X \in \mathcal{G}; \mu_{\mathcal{G}}(x)] = 0.6605$. As a consequence, $P_D[X \notin \mathcal{G}; \mu_{\mathcal{G}}(x)] = 1 - 0.6605 = 0.3395$. By contrast, suppose now that, if D were to specify $P_D(x \in \mathcal{G})$ via the label $P_2(x \in \mathcal{G})$ of Figure 6 and keep everything else the same; then $P_D[X \in \mathcal{G}; \mu_{\mathcal{G}}(x)]$ would increase to 0.7486. Thus, even a small change in the form of $P_D(x \in \mathcal{G})$ produces a noticeable change in D's final answer.

4.4 Reliability of Components in Vague Binary States

We say that a component's state is 'vague and binary' if interest centres around a single vague set of the kind $\mathcal{G}$ or $\mathcal{B}$ in our illustrations. As was mentioned before, we should bear in mind that, in general, $\mathcal{G}^C$ need not be $\mathcal{B}$ and vice versa, unless of course $\mathcal{G}$ and $\mathcal{B}$ are precise sets. For $\mathcal{G} = \{x; x \text{ is a 'desirable' state}\}$ and $\mu_{\mathcal{G}}(x)$ specified, it is reasonable to define the *reliability* of the

component as $P_D[X \in \mathcal{G}; \mu_{\mathcal{G}}(x)]$. Equation (6) can now be used to evaluate this probability. With $\mathcal{B} = \{x; x \text{ is an 'undesirable' state}\}$, and $\mu_{\mathcal{B}}(x)$ specified, we may define the *unreliability* of the component as $P_D[X \in \mathcal{B}; \mu_{\mathcal{B}}(x)]$. We could have also defined the unreliability of the component as $P_D[X \in \mathcal{G}^C; \mu_{\mathcal{G}}(x)]$, where $\mathcal{G}^C$ is that fuzzy set whose membership function equals $1 - \mu_{\mathcal{G}}(x)$. With either choice for the definition of unreliability, we see that, when a component's state is vague and binary, *its unreliability is not necessarily the complement of its reliability*! This result is in contrast to that of binary coherent systems.

Example 4: The case of components that can exist in precise binary states can be encompassed within the above framework; $\mu_{\mathcal{G}}(x) = 1$ for $x \geq x^*$ and $\pi_D(x \in \mathcal{G}) = 1$ if $x \geq x^*$, and zero otherwise. Furthermore, $\mathcal{B} = \mathcal{G}^C$, thus $P_D[X \in \mathcal{G}; \mu_{\mathcal{G}}(x)] = 1 - F_D(x^*)$ and $P_D[X \in \mathcal{B}; \mu_{\mathcal{G}}(x)] = P_D[X \notin \mathcal{G}; \mu_{\mathcal{G}}(x)] = F_D(x^*)$, where $F_D(x^*)$ is the cumulative distribution function (cdf) associated with $p_D(x)$ evaluated at x^*.

5 Binary State Systems with Imprecise Classification

The purpose of this section is to extend the development of Section 4.3 on binary state components with imprecise classification to the case of binary state, n-component systems with imprecise classification. By 'binary state systems with imprecise classification', we mean those systems whose component states are vague and binary, and whose structure functions satisfy the logical connectives of Lukasiewicz; see Section 2. Our motivation for choosing this as a definition of structure functions is that the structure functions of binary state coherent systems with precise classification are exactly the membership functions of certain precise sets. The case of multistate systems with imprecise classification, though not discussed here, follows by analogy.

5.1 Structure Functions as Membership Functions of Precise Sets

Let X_i be the state of component i taking a particular value x_i, $i = 1, \ldots, n$. Suppose that each X_i can take values in $\mathcal{S} = \{x; 0 \leq x \leq 1\}$. Let $\mathcal{G}_i = \{x_i; x_i \text{ is a 'desirable' state}\}$, $\mathcal{G}_i \subset \mathcal{S}$. Let $\mu_{\mathcal{G}_i}(x_i)$ denote the membership function of $\mathcal{G}_i$, $i = 1, \ldots, n$. For now, suppose that $\mathcal{G}_i$ is precise for all i. That is, for each i, there exists an x_i^* such that $\mu_{\mathcal{G}_i}(x_i) = 1(0)$ when $x_i \geq x_i^*$ ($x_i < x_i^*$). For ease of notation, this section focuses solely on the subspace $\mathcal{G}_i$; therefore, we use $\mu_i(x_i)$ to denote the representation of the above membership functions, with the understanding that the membership function assigned is dependent on the fuzzy classification, $\mathcal{G}_i$, which itself depends on component i. For the remainder of this paper, we let $\mathcal{L}[X \notin \mathcal{G}_i; \mu_i(x)] = 1 - \mu_i(x)$ and $\mathcal{L}[X \notin \mathcal{G}_{\phi(\mathbf{X})}; \mu_{\phi(\mathbf{X})}(x)] = 1 - \mu_{\phi(\mathbf{X})}(x)$, where $\phi(\mathbf{X})$ is as defined in Section 1.2.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ and suppose that the n components are in series. Thus the system's structure function is 1 if and only if $x_i \geq x_i^*$ for all $i = 1, \ldots, n$. However, $x_i \geq x_i^*$ implies that $\mu_i(x_i) = 1$ for each i. Thus we may write

$$\phi_S(\mathbf{X}) = \prod_{i=1}^{n} \mu_i(X_i) = \min_i [\mu_i(X_i)] \doteq \mu_{(1:n)}(\mathbf{X}), \tag{8}$$

where $\mu_{(1:n)}(\mathbf{X})$ is the membership function of the intersection of the n precise sets $\mathcal{G}_i$, $i = 1, \ldots, n$. Thus, the structure function of a series system with precise classification can also be interpreted as the membership function of the intersection of n precise sets. Similarly, if the n components were to be connected in parallel redundancy, then the structure function of the

system would be

$$\phi_P(\mathbf{X}) = \coprod_{i=1}^{n} \mu_i(X_i) = \max_i[\mu_i(X_i)] \doteq \mu_{(n:n)}(\mathbf{X}), \tag{9}$$

which is the membership function of the union of $\mathcal{G}_i, i = 1, \ldots, n$. Finally, for a k-out-of-n system, we could write

$$\phi_K(\mathbf{X}) = \begin{cases} 1, & \text{if } \sum_{i=1}^{n} \mu_i(X_i) \geq k \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Whereas the relationships of equations (8) and (9) have an interpretation within the calculus of fuzzy sets, equation (10) does not. Sums of membership functions are not a part of the calculus of fuzzy sets. We therefore seek an alternate way of expressing $\phi_K(\mathbf{X})$. We do this as follows.

Suppose that the $\mu_i(X_i)$ terms are relabeled so that $\mu_{(1:n)}(\mathbf{X})$ is the minimum and $\mu_{(n:n)}(\mathbf{X})$ is the maximum; i.e. $\mu_{(1:n)}(\mathbf{X}) \leq \mu_{(2:n)}(\mathbf{X}) \leq \cdots \leq \mu_{(n-k+1:n)}(\mathbf{X}) \leq \cdots \leq \mu_{(n:n)}(\mathbf{X})$. Since each $\mu_i(X_i)$ is either zero or one, the above ordering will result in equalities for many of the above terms. Once the above is done, we see that $\phi_K(\mathbf{X}) = \mu_{(n-k+1:n)}(\mathbf{X})$. Thus, in general, the structure function of a k-out-of-n system is the membership function of the precise set intersecting the k smallest $\mathcal{G}_i$ sets.

5.2 Structure Functions of Vague Binary State Systems

Motivated by the material of the previous section, we define the structure function of series, parallel, and k-out-of-n systems whose component states are vague and binary as

$$\begin{aligned} \phi_S(\mathbf{X}) &= \min_i[\mu_i(X_i)] = \mu_{(1:n)}(\mathbf{X}), \\ \phi_P(\mathbf{X}) &= \max_i[\mu_i(X_i)] = \mu_{(n:n)}(\mathbf{X}), \text{ and} \\ \phi_K(\mathbf{X}) &= \mu_{(n-k+1:n)}(\mathbf{X}). \end{aligned}$$

These structure functions are identical to those for the case of binary precise sets, except that now, $\mu_i(X_i)$ is a membership function of an associated vague set $\mathcal{G}_i, i = 1, \ldots, n$.

Finally, if $\pi_D(x_i \in \mathcal{G}_i)$ denotes D's probability that a particular x_i gets classified in $\mathcal{G}_i$, then by analogy with equation (7), we have

$$P_D[X_i \in \mathcal{G}_i; \mu_i(x_i)] = \int_{x_i} \left[1 - \left(1 - \frac{1}{\mu_i(x_i)}\right) \cdot \frac{\pi_D(x_i \notin \mathcal{G}_i)}{\pi_D(x_i \in \mathcal{G}_i)}\right]^{-1} \mathrm{d}P_D(x_i), \tag{11}$$

where $P_D(x_i)$ is D's probability that $X_i \leq x_i$.

Our development thus far has assumed that the membership functions $\mu_i(x_i), i = 1, \ldots, n$, are all distinct. Simplification occurs if $\mu_i(x_i) = \mu(x)$ for $i = 1, \ldots, n$. We limit our attention to the case of series and parallel systems because more complicated systems, such as networks can be represented as a combination of series-parallel systems.

5.3 Reliability of Vague Binary State Systems

If the state of each component in a system is a desirable state, will the system itself be in a desirable state? The answer to this question need not be in the affirmative. This is because requirements on the system could be more stringent than those on each component of the system. This is unlike the case of binary state systems with precise classification wherein a series system is judged to be reliable if all its components are reliable. Thus, there are two possible ways in which the reliability of a vague coherent system can be defined. The first is to assume that a

series system is reliable if all its components are in a desirable state. The second is to require that for a system to be judged reliable, its state—say x—be a desirable state. Specifically, we require that $x \in \mathcal{G}_{\phi(\mathbf{X})}$, where $\mathcal{G}_{\phi(\mathbf{X})} = \{x; x \text{ is a 'desirable' system state}\}$ and $\mathcal{G}_{\phi(\mathbf{X})} \subset S$. Associated with $\mathcal{G}_{\phi(\mathbf{X})}$ is its membership function, $\mu_{\mathcal{G}_{\phi(\mathbf{X})}}(x)$. Similarly, in the case of a parallel system, we have two possibilities for defining reliability—the first one being that the system is reliable if at least one of its components is in a desirable state, and the second being the requirement that its state $x \in \mathcal{G}_{\phi(\mathbf{X})}$. We simplify notation by letting $\mu_{\phi(\mathbf{X})}(x) = \mu_{\mathcal{G}_{\phi(\mathbf{X})}}(x)$ and focusing the discussion on the subspace $\mathcal{G}_{(\cdot)}$.

For assessing reliability, let us consider the first case for series and parallel systems. Assuming the X_i's independent, the reliability of a series system would be $\prod_{i=1}^{n}[P_D[X_i \in \mathcal{G}_i; \mu_i(x_i)]]$ where $P_D[X_i \in \mathcal{G}_i; \mu_i(x_i)]$ is given by equation (11). The reliability of a parallel redundant system is $P_D(\bigcup_{i=1}^{n}\{X_i \in \mathcal{G}_i\}; \mu_i(x_i), i = 1, \ldots, n)$; it can be evaluated by the Inclusion-Exclusion formula of probability (Feller, 1968). The computations simplify when the X_i's are assumed identically distributed. The case of k-out-of-n systems follows along similar lines.

With regard to the above, a question arises as to what we mean by independence of the X_i's, when the X_i's take values in a vague set. In the context of equation (11), X_i and X_j, $i \neq j$, will be judged independent if

$$\begin{aligned} P_D(X_i \leq x_i, X_j \leq x_j) &= P_D(X_i \leq x_i) \cdot P_D(X_j \leq x_j), \text{ and if} \\ P_D(x_i \in \mathcal{G}_i, x_j \in \mathcal{G}_j) &= P_D(x_i \in \mathcal{G}_i) \cdot P_D(x_j \in \mathcal{G}_j) \text{ and} \\ P_D(x_i \notin \mathcal{G}_i, x_j \notin \mathcal{G}_j) &= P_D(x_i \notin \mathcal{G}_i) \cdot P_D(x_j \notin \mathcal{G}_j). \end{aligned}$$

The more interesting case is the second one, wherein a system is reliable if the state in which it resides is a desirable one. We start with the case of a series system with structure function $\phi_S(\mathbf{X})$. Its reliability is $P_D(\phi_S(\mathbf{X}) \in \mathcal{G}_{\phi_S(\mathbf{X})}; \mu_{\phi_S[\mathbf{X})}(x)]$ which, from equation (11), is of the form

$$P_D(\phi_S(\mathbf{X}) \in \mathcal{G}_{\phi_S(\mathbf{X})}; \mu_{\phi_S[\mathbf{X})}(x)] = \int_x \left[1 - \left(1 - \frac{1}{\mu_{\phi_S(\mathbf{X})}(x)}\right) \cdot \frac{\pi_D(x \notin \mathcal{G}_{\phi_S(\mathbf{X})})}{\pi_D(x \in \mathcal{G}_{\phi_S(\mathbf{X})})}\right]^{-1} \mathrm{d}P_D(x), \tag{12}$$

where $\pi_D(x \in \mathcal{G}_{\phi_S(\mathbf{X})})$ is D's probability that x is classified in $\mathcal{G}_{\phi_S(\mathbf{X})}$ were $\phi_S(X) = x$, and $P_D(x)$ is D's probability that $\phi_S(\mathbf{X}) \leq x$.

Since $\phi_S(\mathbf{X}) = \min_i \mu_i(X_i) = \mu_{(1:n)}(\mathbf{X})$, we obtain $P_D(x)$ as follows:

$$\begin{aligned} P_D(\phi_S(\mathbf{X}) \geq x) &= P_D(\mu_{(1)}(\mathbf{X}) \geq x) \\ &= P_D(\mu_i(X_i) \geq x, \ i = 1, \ldots, n) \\ &= P_D(X_i \geq \mu_i^{-1}(x), \ i = 1, \ldots, n), \\ &= \prod_{i=1}^{n} P_D[X_i \geq \mu_i^{-1}(x)], \text{ if } X_i\text{'s are assumed independent,} \end{aligned} \tag{13}$$

where $\mu_i^{-1}(\cdot)$ denotes the inverse of $\mu_i(\cdot)$. Subsequently, $\mathrm{dP_D}(x)$ can be obtained. If the X_i's cannot be judged independent with respect to D's distribution for the X_i's, we need to specify a joint distribution for these, such as Marshall & Olkin's (1967) multivariate exponential, or any of its variants. In the case of parallel systems, the development will proceed along similar lines, save that now $P_D(x)$ will be obtained via $\prod_{i=1}^{n} P_D[X_i \leq \mu_i^{-1}(x)]$. Finally, the case of $(n - k + 1)$-out-of-n would follow by considering the distribution of the k-th order membership function, $\mu_{(k:n)}(x)$.

Example 5: Consider a two-component series system where the component performances are independent and identically distributed. D wishes to assess $P_D[\phi_S(\mathbf{X}) \in \mathcal{G}_{\phi_S(\mathbf{X})}; \mu_{\phi_S(\mathbf{X})}(x)]$. The

first option is to compute the product of the component probabilities. Let $\mu_{\mathcal{G}_i}(x) = x^2$, and $P_D(x)$ and $P_D(x \in \mathcal{G}_i)$ be as shown in Figures 5 and 6, respectively, for $i = 1, 2$. Then, $P_D[\phi_S(\mathbf{X}) \in \mathcal{G}_{\phi_S(\mathbf{X})}; \mu_{\phi_S(\mathbf{X})}(x)] = 0.6232$. The second option is to compute the system reliability directly, through the use of Z's membership function for the entire system. Supposing that the expert holds a stronger standard for the system to be in a desirable state than that for the components, we let $\mu_{\phi_S(\mathbf{X})}(x) = x^{10}$. Meanwhile, D considers $P_D(x)$ and $P_D(x \in \mathcal{G}_i)$ as specified in Figures 5 and 6 for $\phi_S(\mathbf{X})$, implying that $P_D[\phi_S(\mathbf{X}) \in \mathcal{G}_{\phi_S(\mathbf{X})}; \mu_{\phi_S(\mathbf{X})}(x)] = 0.4321$. Thus, by holding the system to a more stringent standard, D's assessment of the system reliability is lower when considered directly, as opposed to that when using a more relaxed membership function to represent belief at the component level.

6 Maintenance Management in a Vague Environment

Examples 3–5 illustrate how D is able to assess the probability that the state of a unit will be in a 'desirable' state, or its complement. Why would D be interested in such a probability instead of the probability that the state of the unit will be x, $0 \leq x \leq 1$? Reasons were given in Section 4.3, the one pertaining to communication using 'natural language' being the most relevant. This point is best underscored via the scenario of maintenance wherein one must decide whether to repair, replace, or simply continue to monitor the unit. In practice, judgments about maintenance are not based on assessments of uncertainty about x; they are based on conjectures about whether or not the unit will be in a 'desirable' state.

Consider the following: a unit is required to perform service for some time period. The unit can exist in one of three states: $\mathcal{G}$ (for good), $\mathcal{B}$ (for bad), and $\mathcal{A}$ (for acceptable). When the unit is in state $\mathcal{G}$, the utility to D provided by the unit is $\mathcal{U}(\mathcal{G})$; analogously, we define $\mathcal{U}(\mathcal{A})$ and $\mathcal{U}(\mathcal{B})$. It is reasonable to suppose that $\mathcal{U}(\mathcal{A}) < \mathcal{U}(\mathcal{G})$ and, in principle, $-\mathcal{U}(\mathcal{B})$ could be greater than $\mathcal{U}(\mathcal{G})$, i.e. the cost for being in state $\mathcal{B}$ could dominate the reward for being in state $\mathcal{G}$. With the above in place, D's problem is to make a decision whether to replace the unit, denoted $\mathcal{R}$, or to repair the unit, denoted $\mathcal{M}$, or do nothing, denoted $\mathcal{N}$. There is a cost associated with each of these three actions, and these are denoted $-\mathcal{U}(\mathcal{R})$, $-\mathcal{U}(\mathcal{M})$, and $-\mathcal{U}(\mathcal{N})$, respectively. Presumably, $-\mathcal{U}(\mathcal{N}) < -\mathcal{U}(\mathcal{M}) < -\mathcal{U}(\mathcal{R})$. Which of the above three actions should D take?

The problem is solved by using *maximization of expected utility* (MEU) [cf. Lindley (1991), p. 58]. The decision tree of Figure 7 facilitates an implementation of this recipe; the rectangle represents D's decision node and the three circles denoted R_1, R_2, and R_3 represent the three nodes corresponding to the three actions $\mathcal{R}$, $\mathcal{M}$ and $\mathcal{N}$, respectively. Each (random) node results in one of three outcomes, $\star = \mathcal{G}, \mathcal{A}$ or $\mathcal{B}$, and these are portrayed in Figure 7 only for the node R_3. At the terminus of the tree are the utilities. For example, $\mathcal{U}(\mathcal{N}, \mathcal{G})$ denotes the utility to D, when D's decision is to monitor the unit and the outcome is $\mathcal{G}$.

The MEU principle requires that, at each random node, D compute an expected utility of an action that leads to that node. For this, D needs to assess the probabilities that at τ, the state of the unit will be in $\mathcal{G}$, $\mathcal{A}$, and $\mathcal{B}$, respectively. These probabilities would depend on three ingredients: membership functions of the kind $\mu_\star(x)$, $\mu_\mathcal{A}(x)$, and $\mu_\mathcal{B}(x)$; D's prior probability that an x is classified (by nature) in $\mathcal{G}$, $\mathcal{A}$, and $\mathcal{B}$ (i.e. $P_D(x \in \star)$, $\star = \mathcal{G}, \mathcal{A}, \mathcal{B}$), and $P_D(x)$, D's subjective probability that the state of the unit will be x. Since $\sum_{\star=\mathcal{G},\mathcal{A},\mathcal{B}} P_D(x \in \star) = 1$, D need only specify any two probabilities. Once these are at hand, D invokes equation (7) to obtain the required probabilities. All of the above is straightforward except that $P_D(x)$ depends on the action that D takes. Both repair and replacement actions tend to right-skew the form of $P_D(x)$ toward one. Thus, with respect to the illustration of Figure 5, a repair action will tend to shift the probability mass closer to one, and moreso with replacement. To summarize, the impact of D's

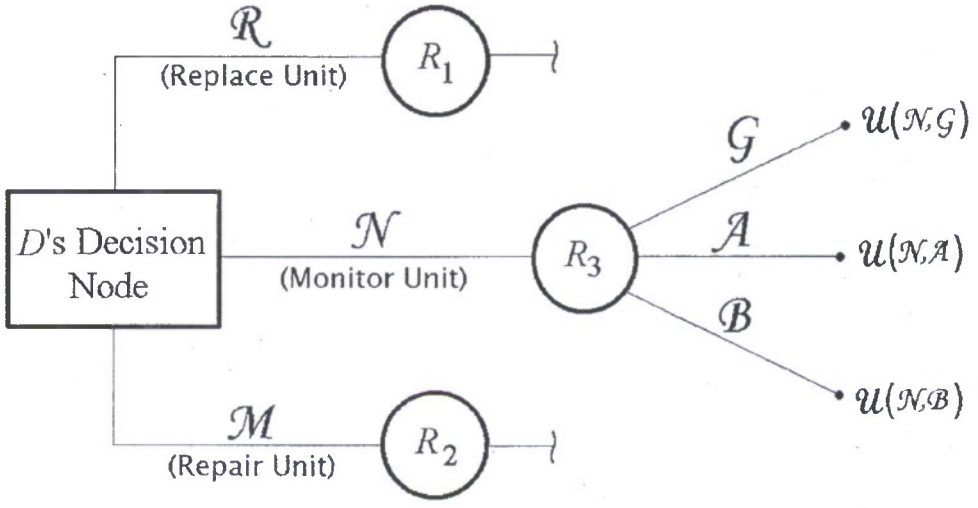

Figure 7. *D's decision tree for maintenance actions.*

actions on D's probabilities of the state of the unit are reflected only in $P_D(x)$. The membership functions and the classification probabilities are unaffected. To denote such a dependence, we shall replace the $P_D(x)$ of equation (7) by $P_D(x;\bullet)$, and $P_D[X \in \star;\ \mu_\star(x)]$ by $P_D(X \in \star;\ \mu_\star(x),\ \bullet)$ for $\bullet = \mathcal{R}, \mathcal{M}$ and $\mathcal{N}$; $\star = \mathcal{G}, A, B$.

Whereas the development in Sections 4 and 5 pertained to the binary case involving two vague sets $\mathcal{B}$ and $\mathcal{G}$, our example here involves three vague sets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{G}$, and their respective membership functions, $\mu_\bullet(x)$, $\bullet = \mathcal{A}, \mathcal{B}$ and $\mathcal{G}$. Of these, only $\mu_{\mathcal{A}}(x)$ warrants comment since the general nature of the other two has been discussed before; see Figures 4(a) and (b). It is reasonable to suppose that the general form of $\mu_{\mathcal{A}}(x)$ is either bell-shaped or an inverted U.

Finally, a question arises as to whether $\mu_{\mathcal{A}}(x)$, $\mu_{\mathcal{B}}(x)$, and $\mu_{\mathcal{G}}(x)$ can take any arbitrary form independent of each other. The answer to this question is in the negative because the membership functions go to determine the quantities $P_D[X \in \mathcal{A};\ \mu_{\mathcal{A}}(x)]$, $P_D[X \in \mathcal{B};\ \mu_{\mathcal{B}}(x)]$ and $P_D[X \in \mathcal{G};\ \mu_{\mathcal{G}}(x)]$, and these must sum to one. Thus, D needs to ensure coherence of the membership functions just like how D needs to ensure a coherence of the classification and state probabilities. Since D elicits membership functions from Z, it is incumbent on D to ensure that membership functions do not lead to results that violate the countable additivity axiom of probability. This important point has not been addressed in Singpurwalla & Booker (2004).

The utilities at the terminus of a tree, $\mathcal{U}(\mathcal{N}, \mathcal{G})$, $\mathcal{U}(\mathcal{N}, \mathcal{A})$ and $\mathcal{U}(\mathcal{N}, \mathcal{B})$ are straightforward to write out. Thus, for example, $\mathcal{U}(\mathcal{N}, \mathcal{G}) = \mathcal{U}(\mathcal{N}) + \mathcal{U}(\mathcal{G})$, which is the sum of the disutility due to monitoring and the utility of the unit being in state $\mathcal{G}$. Similarly, $\mathcal{U}(\mathcal{N}, \mathcal{B}) = \mathcal{U}(\mathcal{N}) + \mathcal{U}(\mathcal{B})$, and $\mathcal{U}(\mathcal{N}, \mathcal{A}) = \mathcal{U}(\mathcal{N}) + \mathcal{U}(\mathcal{A})$. With this in place, we compute the expected utility at each node. Thus, for example, $\mathcal{U}(\mathcal{N})$, the *expected utility* at node R_3 is $\mathcal{U}(N) = \sum_{\star=\mathcal{G},\mathcal{A},\mathcal{B}} \mathcal{U}(N, \star) \cdot P_D(X \in \star;\ \mu_\star(x), \mathcal{N})$, where $P_D(X \in \star;\ \mu_\star(x), \mathcal{N})$ is the right-hand side of equation (7) with $P_D(x)$ replaced by $P_D(x;\mathcal{N})$; similarly, the other terms of $\mathcal{U}(\mathcal{N})$. The expected utilities at nodes R_1 and R_3 are analogously computed as $\mathcal{U}(\mathcal{R})$ and $\mathcal{U}(\mathcal{M})$, respectively, *mutatis mutandis*. Once the above are done, D's maintenance decision is to choose that action for which the expected utility is a maximum. Thus, for example, if $\mathcal{U}(\mathcal{N}) > \mathcal{U}(\mathcal{R}) > \mathcal{U}(\mathcal{M})$, then D's decision would be simply to do nothing.

How does the above material differ from that which is currently available in the literature on maintenance planning? The current literature would require each node to be binary and,

to compute the expected utility at each node, all we need is the probability that $x \geq x^*$. This probability can be had once D specifies $P_D(x; \bullet)$, $\bullet = \mathcal{R}, \mathcal{N}$ and $\mathcal{M}$. By contrast, we allow an x to exist in three vaguely defined sets, and allow x to simultaneously exist in more than one of these. The advantage is flexibility and a facility to entertain an analysis that facilitates natural language communication. Further, in the existing literature, uncertainties are assessed about times to failure via probabilistic failure models, and failure is viewed as a sharply defined event. Consequently, the analysis is forced into a binary framework. By contrast, our uncertainties are focused on x which can encapsulate degradation of a unit.

7 Summary and Conclusions

The term 'complex stochastic systems' is well entrenched into the vocabulary of statisticians, though it generally pertains to a use of the Markov Chain Monte Carlo method. This paper takes a broader view of this term by embedding within it the theory of vague coherent structures. This theory, which is generally associated with work in applied probability and reliability is germane to statisticians, especially those whose focus is on biostatistics, genetics, graphical models, and neural nets. With that in mind, we have devoted Section 1 to an overview of the key notions and ideas of binary state systems whose two states can be precisely delineated. The mathematics which drives the development of results for such systems is binary logic. In Section 1, we also set the stage for the material of Sections 4 and 5 by introducing the idea of imprecise or vague sets. The need for such sets has been acknowledged by physicists, philosophers, and logicians. More recently, their need has also been recognized by those involved in decision making and natural language processing. Section 2 is devoted to multivalued logic in the context of multivalued propositions. The focus here is on the connectives of conjunction and disjunction; these connectives can be used to define the structure function of multistate systems, a topic treated in Section 3. In Section 3, it is assumed that the classification of states is precise. This topic has been covered before via the literature on multistate reliability; however, what is new here is the departure from binary logic to multivalued logic.

Sections 4 and 5 impart to this paper a feature that is novel. Specifically, they pertain to the development of reliability for components and systems whose state space is vague. In actuality, vague state spaces are more realistic than the usual zero-one states, which are an idealization. In Sections 4 and 5, we also show that the usual notions of reliability do not always hold when the state space is vague. For example, the unreliability of a unit is not one minus its reliability, and that there is more than one way to define system reliability.

There is another aspect of this paper that warrants comment. In the existing theory of coherent structures with precise classification, statistical principles have no role to play. All that is needed is the calculus of probability. By contrast, when dealing with vague systems, membership functions and consistency profiles create a role for the likelihood function and, in so doing, mandate a consideration of the principles of Bayesian statistical inference.

The illustrative examples of Sections 4 and 5, and the maintenance management architecture of Section 6 should give the reader an inkling of the practical import of the material here. For example, in maintenance and replacement actions pertaining to decision making uncertainty, the usual strategy is to assume that the state space is binary—functioning and failed. In actuality, functioning can occur at different levels whose boundaries cannot be sharply delineated. Thus, it makes more sense to study maintenance and replacement when the state space is vague for, in actuality, this is how such decisions are made.

Acknowledgements

The authors wish to thank the reviewers for their helpful comments. Nozer Singpurwalla was supported by the Office of Naval Research Contract N00014-06-1-0037 and by the Army Research Office Grant W911NF-05-1-0209 with The George Washington University.

References

Barlow, R. & Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*. New York: Holt, Rinehart and Winston, Inc.

Barlow, R. & Wu, A. (1978). Coherent systems with multi-state components. *Math. Oper. Res.*, 3(4), 275–281.

Basu, D. (1975). Statistical information and likelihood. *Sankhyā, Ser. A*, 37, 1–71.

Baxter, L. (1984). Continuum structures I. *J. Appl. Probab.*, **21**, 802–815.

Bertoluzza, C., Gil, M.A. & Ralescu, D.A. (2002). *Statistical Modeling, Analysis and Management of Fuzzy Data*. New York: Springer-Verlag.

Birnbaum, Z., Esary, J. & Saunders, S. (1961). Multicomponent systems and structures, and their reliability. *Technometrics*, **3**, 55–77.

Black, M. (1939). Vagueness: an exercise in logical analysis. *Philos. Sci.*, 3(1), 427–455.

Clarotti, C. & Lindley, D. (1988). *Accelerated Life Testing and Expert Opinion in Reliability*. New York: Elsevier Science Publishing.

Cox, D.R., Fitzpatrick, R., Fletcher, A.E., Gore, S.M., Spiegelhalter, D.J. & Jones, D.R. (1992). Quality-of-life assessment: Can we keep it simple? *J. Roy. Statist. Soc. A*, **155**(3), 353–393.

El-Neweihi, E. & Proschan, F. (1984). Degradable systems: A survey of multistate system theory. *Commun. Stat. Theory*, **13**, 405–432.

El-Neweihi, E., Proschan, F. & Sethuraman, J. (1978). Multistate coherent systems. *J. Appl. Probab.*, **15**, 675–688.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, **1**, 3rd ed. New York: John Wiley & Sons, Inc.

Griffith, W. (1980). Multistate reliability models. *J. Appl. Probab.*, **17**, 735–744.

Huang, Y. (2002). Calibration regression of censored lifetime medical cost. *J. Am. Statist. Assoc.*, **97**, 318–327.

Kass, R.E. (1993). Bayes factors in practice. *Statistician*, **42**, 551–560.

Lindley, D. & Phillips, L. (1976). Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.*, **30**, 112–119.

Lindley, D. & Singpurwalla, N. (2002). On exchangeable, causal and cascading failures. *Statist. Sci.*, **17**(2), 209–219.

Lindley, D.V. (1991). *Making Decisions*. New York: John Wiley.

Lukasiewicz, J. (1930). Philosophische Bemerkungen zu mehrwertigen Systemen des Aussagenkalkūls. English tr. (in 1967) Philosophical remarks on many-valued systems of propositional logic. In *Polish Logic 1920-1939*, Ed. S. McCall, pp. 40–65. Oxford: Clarendon Press.

Lynn, N., Singpurwalla, N. & Smith, A. (1998). Bayesian assessment of network reliability. *SIAM Rev.*, **40**(2), 202–227.

Malinowski, G. (1993). *Many-valued Logics*. Oxford: Clarendon Press.

Marshall, A. & Olkin, I. (1967). A multivariate exponential distribution. *J. Am. Statist. Assoc.*, **62**, 30–44.

Marshall, A.W. (1994). A systems model for reliability studies. *Statist. Sinica*, 4(2), 549–565.

Matland, R. & Singpurwalla, N. (1981). A reliability model for chapteralized precipitators. *J. Air Pollut. Contl. Assoc.*, **31**(2), 144–147.

Natvig, B. (1982). Two suggestions of how to define a multistate coherent system. *Adv. Appl. Probab.*, **14**, 434–455.

Pagels, H. (1982). *The Cosmic Code*. New York: Bantam Books.

Reese, C., Wilson, A., Hamada, M., Martz, H. & Ryan, K. (2004). Integrated analysis of computer and physical experiments. *Technometrics*, **46**(2), 153–164.

Russell, B. (1923). Vagueness. *Aust. J. Philos.*, **1**, 88.

Singpurwalla, N. & Booker, J. (2004). Membership functions and probability measures of fuzzy sets. *J. Am. Statist. Assoc.*, **99**(467), 867–877.

Smith, R. (1983). Limit theorems and approximations for the reliability of load-sharing systems. *Adv. Appl. Probab.*, **15**(2), 304–330.

Spizzichino, F. (2001). *Subjective Probability Models for Life-times*. Boca Raton, FL: Chapman and Hall/CRC.

Viertl, R. (1996). *Statistical Methods for Non-Precise Data*. Boca Raton, FL: CRC Press.

Viertl, R. (2006). Univariate statistical analysis with fuzzy data. *Comput. Statist. Data Anal.*, **51**(1), 133–147.

Zadeh, L. (1965). Fuzzy sets. *Inform. Control*, **8**(3), 338–353.

Résumé

L'état de l'art dans la théorie de structure cohérente est guidé par deux assertions qui sont tous deux limitants : (1) toutes les unités d'un système peuvent exister dans un de deux états, défaillant ou fonctionnant; et (2) à n'importe quel moment, chaque unité peut seulement exister dans un des susdits états. En réalité, les unités peuvent exister dans plus de deux états et c'est possible qu'une unité puisse *simultanément* exister dans plus d'un état. Cette dernière caractéristique est une conséquence de l'opinion qu'il ne soit peut-être pas possible de définir avec précision les sous-ensembles d'un ensemble d'états; on appelle de tels sous-ensembles vagues. La première restriction a été adressée par les méthodes appelées "systèmes multi-états"; pourtant, ces méthodes n'ont pas pris avantage des mathématiques sur les propositions multivalues en logique. Ici, nous invoquons ses tables de vérité pour définir la fonction des systémes multi-états et exploiter ensuite nos résultats dans le contexte d'ambiguïté. Une contribution clé de ce papier est d'argumenter que la logique de plusieurs values est une plateforme commune pour étudier tant les systèmes multi-états que les systémes vagues, mais pour faire ceci, il est nécessaire de se baser sur plusieurs principes d'inférence statistique.

[*Received June 2007, accepted March 2008*]

Available online at www.sciencedirect.com
ScienceDirect
Statistics & Probability Letters 77 (2007) 1354–1361
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

Betting on residual life: The caveats of conditioning

Nozer D. Singpurwalla*

The George Washington University, Washington, DC 20052, USA

Available online 24 March 2007

Abstract

Assessing conditionals based on any specified probability model is straightforward and unique when the conditioning event is in the subjunctive mood; that is, supposing that the conditioning event were to occur. The matter becomes problematic, however, when the conditioning event actually does occur as observed data, and thus becomes a reality. We illustrate this point by considering a commonly occurring scenario in the actuarial sciences, engineering reliability, survival analysis, and in general, any type of an activity that involves filtering. We argue that there could be more than one way to bet on residual life. Our message is that it is the likelihood—not Bayes' Law—which is the tail that wags the dog!

This paper should appeal to both probabilists and statisticians who are interested in foundational issues. It has been written to honour Richard Johnson whose Editorship of *Statistics and Probability Letters* has provided a platform for dialogue between probabilists, statisticians, and those who strive to be both.

Keywords: Actuarial science; Conditionalization principle; Double slit experiment; Filtering; Forecasting; Likelihood; Reliability; Survival analysis

1. Introduction

In the process of using marker data to assess the lifetime of an item experiencing failure due to ageing, we were confronted by a dilemma that sneaked upon us as a matter of course (see Singpurwalla, 2006a). It turns out that the scenario leading to the dilemma is quite common and can arise when addressing practical issues of conditioning in the actuarial, the engineering, and the biomedical sciences. Stripped to its essentials, the scenario goes as follows.

Suppose that an item's lifetime X is judged to have a distribution function $G(x) = P(X \leqslant x)$, and a survival function $\overline{G}(x) \stackrel{\text{def}}{=} 1 - G(x) = P(X > x)$. We suppose that lifetime can be continuously monitored so that $x \geqslant 0$. Were this item supposed to survive until x, its *residual* (or *remaining*) *lifetime* will be $X - x$. We are required to make statements of uncertainty about $(X - x)$, so that actuarial, engineering, or medical decisions about the item can be made. That is, we are required to specify $P(X - x > u | X > x)$, for all $u > 0$. Our interpretation of probability is de Finnetian (see de Finetti, 1937), in the sense that probability reflects one's disposition to a two-sided bet. Thus, probability assessments can be seen as a device for hedging our bets on the item's survival, or some other unknown quantity of interest, such as parameters in probability models.

*Tel.: +1 202 994 7515; fax: +1 202 994 7508.
E-mail address: nozer@gwu.edu.

0167-7152/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2007.03.021

A solution to the problem posed is elementary and unique, given a distribution function G. Specifically, for any $u>0$

$$P(X-x>u|X>x)=P(X>x+u|X>x)=\frac{P(X>x+u)}{P(X>x)}=\frac{\overline{G}(x+u)}{\overline{G}(x)}. \tag{1.1}$$

Suppose now, that instead of the subjunctive, "*were* the item to survive until x", we are told that the item actually *did* survive to x. That is, the event $(X>x)$ is no more an uncertain event; $(X>x)$ has now become observed data. What then would our assessment of the uncertainty about the residual life $(X-x)$ be? In other words, how would we bet on the event $(X-x>u)$, for $u>0$? Would it continue to be $\overline{G}(x+u)/\overline{G}(x)$, or could it be something else? If the latter, would the number to bet be unique? For a discussion of these and related questions, one may visit Freedman and Purves (1969). A more recent discourse on the different kinds of conditional beliefs is in Joyce (1999, Chapter 6).

Intuitively, it seems that there ought to be some distinction between looking at $(X>x)$ as a possibility, versus looking at it as a fact that is revealed as data. Thus, $\overline{G}(x+u)/\overline{G}(x)$ need not be the correct answer. Yet many individuals when faced with this problem would simply mimic the steps leading to Eq. (1.1) and continue to declare $\overline{G}(x+u)/\overline{G}(x)$ as their answer. In doing so they do not appear to be making a distinction between $(X>x)$ as a supposition versus a reality. Alternatively put, they may be failing to recognize the connotation that in a conditional probability statement, the word "given" does not indicate a fact; rather it indicates a supposition that the conditioning event is true. Thus, are those who declare $\overline{G}(x+u)/\overline{G}(x)$ as their answer—irrespective of the character of the conditioning event—in error, or is there a rationale for their answer?

We claim that the rationale cannot completely be within the calculus of probability, because the notion of probability—at least from a subjectivistic point of view—is germane only when the disposition of *all* events in question is unknown. Thus, for example, it may not make sense to say that the probability that a coin with heads on both faces when flipped will land heads, is one. This is because the disposition of the outcome is known before the flip. Consequently, a two-sided bet on the outcome heads has to be \$1, which will be exchanged for a \$1 when the coin lands heads, which it will. The two-sided bet of \$1 is thus meaningless. The rationale therefore must come from concepts in statistics wherein the notion of a likelihood plays a signal role. By all accounts the notion of a likelihood appears to be alien to probability theory.

In what follows we point out that there are both philosophical and technical arguments which support $\overline{G}(x+u)/\overline{G}(x)$ as an answer, but that this answer is one among other possible answers. This is the main point of this article. Arguments about conditioning are common among philosophers of science. That such arguments could also be relevant to reliability, survival analysis, filtering, and forecasting seems to not have been recognized.

2. Answer(s) to the question

2.1. *Reassessment and the principle of conditionalization*

Some individuals when faced with the matter of assessing $P(X-x>u)$ with $(X>x)$ as observed data, may chose to re-assess all probabilities treating the factual event $(X>x)$ as a part of background history; that is, they would start from ground zero, even if the observed $(X>x)$ is not a surprise. Diaconis and Zabell (1982) label a process like this, *complete reassessment*; however, the driving premise considered by the above authors is different from the one we are discussing here, in the sense that the observed event is considered to be a surprise. In a re-assessment one essentially starts all over again from scratch and possibly even rejects G as the underlying probability model. The answer that one obtains may therefore not necessarily be $\overline{G}(x+u)/\overline{G}(x)$. Reassessment is a perfectly legitimate step; its main danger is the risk of incoherence (i.e. a lack of consistency). We therefore do not pursue here this line of reasoning and do not advocate reassessment as a strategy.

To ensure coherence one may proceed formally by invoking Bayes' Law as an inferential mechanism, using $(X>x)$ as data. These are two directions from which this can be approached, one general, the other specific. These we describe in Sections 2.2 and 2.3, respectively, wherein we point out that there need not be a unique

answer to the question posed, and that under a certain assumption, $\overline{G}(x+u)/\overline{G}(x)$ will indeed be one of several possible answers.

But there is another, more philosophical, argument that supports $\overline{G}(x+u)/\overline{G}(x)$ as a correct answer. This argument, known as the *Principle of conditionalization* (cf. Howson and Urbach, 1989, p. 68), proceeds as follows:

Prior to observing $(X>x)$ as factual data, we had declared that $\overline{G}(x+u)/\overline{G}(x)$ would represent our bet (or personal probability) on the event $(X-x>u)$, for some $u>0$, were the event $(X>x)$ turns out to be a fact. Now that $(X>x)$ has revealed itself as being actually true, we shall act as we had declared, and thus $\overline{G}(x+u)/\overline{G}(x)$ would continue to be our bet. As suggested by a reviewer, another way to articulate the principle of conditionalization is, to assert that "if I say I am going to do something, I will do it".

Those who subscribe to a complete reassessment by starting all over from scratch, may reject the principle of conditionalization on grounds that the *actual* occurrence of the event $(X>x)$ has changed their psychological disposition so dramatically from their disposition under the supposition that $(X>x)$, that they can no more subscribe to G as their model of uncertainty. They then seek an alternate to G, say H as a model for assessing $(X-x)$. This point was made by Ramsey (1931) (cf. Diaconis and Zabell, 1982) who stated that

> [The degree of belief in p given q] is not the same as the degree to which [a subject] would believe p, if he believed q for certain; for knowledge of q might for psychological reasons profoundly alter his whole system of beliefs.

Diaconis and Zabell (1982) also cite other, more modern, references that mention the above issue; these are Hacking (1967), de Finetti (1972, p. 150; 1975, p. 203), Teller (1976), and Freedman and Purves (1969).

Additionally, there also happens to be empirical evidence from quantum mechanics that rejects the conditionalization principle vis-a-vis the "double slit experiment". This experiment has now become a classic thought experiment for its clarity in expressing the central puzzles of quantum mechanics. In its original version, performed by the English scientist Thomas Young sometime around 1805, the experiment consisted of letting light diffract through two slits producing fringes on a screen. The goal of the experiment was to resolve the question as to whether light is composed of particles or waves. The current versions of the experiment are performed with electrons instead of light (cf. Jonsson, 1974). Such experiments have shown that the probability (as assessed via the relative frequency) of some event, say B, when an event A always occurs is not equal to the conditional probability of B given A found from an experiment in which A occurs in some replications and the complement of A occurs in other replications. This tantamounts to a negation of the principle of conditionalization.

2.2. *Using Bayes' Law, directly*

The clearest, and perhaps the most natural way to address the question posed is via a use of Bayes' Law. But to better articulate the workings of this law in the present context, we introduce the convention (see Singpurwalla, 2006b) that for two events A and B, $P(A|B)$ denotes the conditioning (or supposition) that B is true, whereas $P(A;B)$ denotes the fact that B is actually true. With the above convention in place, our problem boils down to assessing $P(X>x+u;X>x)$. The answer is given by Eq. (2.2). But the arguments leading to this equation entail a transition from purely probabilistic considerations to the statistical ones, and these may be helpful to re-iterate.

To assess $P(X>x+u;X>x)$, one way to start is by considering the proposition $P(X>x+u|X>x)$, which by Bayes' Law leads us to the inverse relationship

$$P(X>x+u|X>x) \propto P(X>x|X>x+u)P(X>x+u), \tag{2.1}$$

where "$\propto$" denotes proportional to. Eq. (2.1) is an honest-to-goodness probability statement.

However, since $(X>x)$ has been observed as data, the middle term of Eq. (2.1) does not make sense as a probability. Instead, it is the *likelihood* of the event $X>x+u$ with $X>x$ fixed. We denote this likelihood by $\mathscr{L}(X>x+u;X>x)$. Similarly, $P(X>x+u|X>x)$ must now be written as $P(X>x+u;X>x)$. In writing $\mathscr{L}(X>x+u;X>x)$ we interpret $X>x+u$, $u>0$, as a hypothesis and $X>x$ as data. This interpretation is not conventional in the sense that in statistical inference likelihoods are generally functions of unknown

parameters, not unknown events. However, as stated by Edwards (1992, p. 12), the likelihood can be regarded as a function of the hypotheses or of the parameters. A treatment of the question posed involving the use of a parametric model which results in the likelihood being a function of the parameter will be discussed in Section 2.3.

With the above in place, Eq. (2.1) now becomes

$$P(X>x+u;X>x) \propto \mathscr{L}(X>x+u;X>x)P(X>x+u). \tag{2.2}$$

The last term of the above expression, being an unknown quantity, is $\overline{G}(x+u)$.

According to Basu (1975, 1982), when Fisher (1912) rediscovered the Gaussian notion of likelihood, he looked upon it as "a scale of comparative support lent by the data to various possible values of θ [an unknown parameter]"; also see Edwards (1992, p. 221). This interpretation of likelihood is (symmetrically) different from the conventional interpretation in which the likelihood tells us which hypothesis better supports the data (cf. Edwards, 1992, p. 9). The point of view that we adopt here is the former. Having done so, we are—in principle—free to choose the functional form of the likelihood function as we see fit. Suppose then, that the likelihood is taken to be a constant, say 1, over all values of $x+u$, with x fixed; see Fig. 1. Note that this choice will also be in keeping with the conventional use of the likelihood. Then Eq. (2.2) would become

$$P(X>x+u;X>x) \propto 1 \cdot P(X>x+u),$$

which when normalized yields $P(X>x+u)/P(X>x) = \overline{G}(x+u)/\overline{G}(x)$ as an answer. Thus, implicit to the answer given by those who subscribe to the principle of conditionalization (i.e. those who mimic the steps to assess conditional probability) is the assumption of a constant likelihood!

Since one is free to choose the functional form of the likelihood, what if the likelihood was chosen by us, see Fig. 1, to be some other function of u, say $\exp(-u)$, for $u>0$? Our assessment of $P(X>x+u;X>x)$ would be different; namely, it would be $\exp(-u)\overline{G}(x+u)/\overline{G}(x)$. This means that it is the form of the likelihood that dictates how we would bet on residual life. The standard answer $\overline{G}(x+u)/\overline{G}(x)$ arises only under the special case of a constant likelihood.

The constant likelihood encapsulates a user's disposition of indifference with respect to the observed $X>x$. A decreasing likelihood one of conservatism. The form of likelihood can therefore be given a behaviouristic justification.

2.3. *Using Bayes' Law, conventionally*

By a conventional use of the Bayes Law we mean the introduction of a parametric model into the analysis followed by a prior to posterior transformation of our uncertainty about the parameters. When we do so, an argument similar to the one of Section 2.2 can be made, and possibly with more transparency, because of the concrete nature of the set-up. Suppose then, that $P(X \leqslant x|\theta) = G(x|\theta)$, where $\theta>0$ is some unknown

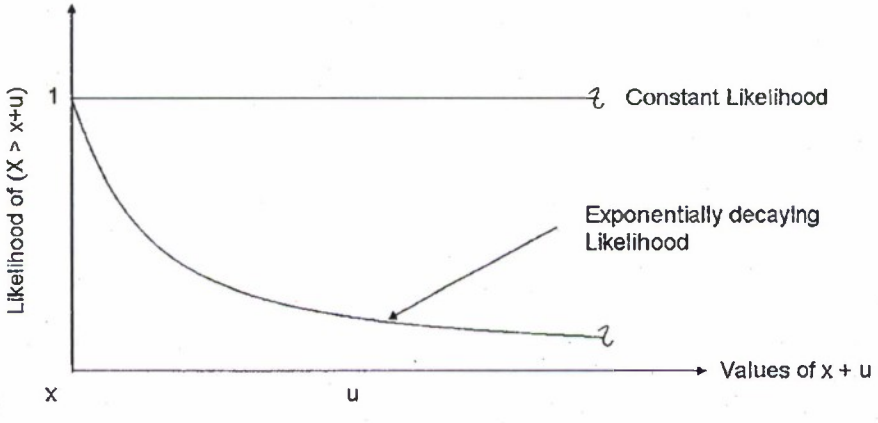

Fig. 1. Likelihood of event $(X>x+u)$ with $(X>x)$ fixed.

parameter. Using standard arguments involving the law of total probability, we may write

$$P(X \geqslant x+u|X \geqslant x) = \int_\theta P(X \geqslant x+u|X \geqslant x, \theta)\pi(\theta|X \geqslant x)\,\mathrm{d}\theta,$$

where by Bayes' Law

$$\pi(\theta|X \geqslant x) \propto P(X \geqslant x|\theta)\pi(\theta);$$

$\pi(\theta)$ is our prior distribution of $\theta > 0$.

With the event $(X \geqslant x)$ as data, the above relationship can be written as

$$P(X \geqslant x+u; X \geqslant x) = \int_\theta P(X \geqslant x+u|\theta; X \geqslant x)\pi(\theta; X \geqslant x)\,\mathrm{d}\theta \tag{2.3}$$

with

$$\pi(\theta; X \geqslant x) \propto \mathscr{L}(\theta; X \geqslant x)\pi(\theta); \tag{2.4}$$

$\mathscr{L}(\theta; X \geqslant x)$ is the likelihood of θ, with $X \geqslant x$ taken to be fixed, known, and also assumed to be credible.

Were we to subscribe to the principle of conditionalization, then $\mathscr{L}(\theta; X \geqslant x)$ will be prescribed by our chosen model $G(x|\theta)$. If otherwise, we are free to choose any other meaningful form for $\mathscr{L}(\theta; X \geqslant x)$, and thus our answers to $P(X \geqslant x+u; X \geqslant x)$ could be different. The example below illustrates this point.

Let $G(x|\theta) = 1 - \exp(-\theta x)$, an exponential distribution with mean $1/\theta$, $\theta > 0$, and let our prior on θ be a gamma distribution with scale (shape) parameter 1 (k). This is a natural conjugate prior for θ, though any other prior will also do. Then

$$\begin{aligned} P(X \geqslant x+u; X \geqslant x) &= \int_0^\infty P(X \geqslant x+u|\theta; X \geqslant x)\pi(\theta; X \geqslant x)\,\mathrm{d}\theta \\ &= \int_0^\infty \mathrm{e}^{-u\theta}\pi(\theta; X \geqslant x)\,\mathrm{d}\theta, \end{aligned}$$

and

$$\pi(\theta; X \geqslant x) \propto \mathscr{L}(\theta; X \geqslant x)\mathrm{e}^{-\theta}\theta^{k-1}/\Gamma(k).$$

When $\mathscr{L}(\theta; X \geqslant x) = \mathrm{e}^{-\theta x}$—which is what the principle of conditionality would mandate, and which is what is conventionally done—then it can be verified that the posterior distribution of θ is also a gamma with scale [shape] $(x+1)[k]$; i.e.

$$\pi(\theta; X \geqslant x) = \mathrm{e}^{-\theta(x+1)}\theta^{k-1}(x+1)^k/\Gamma(k).$$

It now follows that

$$\begin{aligned} P(X \geqslant x+u; X \geqslant x) &= \int_0^\infty \mathrm{e}^{-u\theta}\mathrm{e}^{-\theta(x+1)}\frac{\theta^{k-1}}{\Gamma(k)}(x+1)^k\,\mathrm{d}\theta \\ &= \left(\frac{x+1}{x+u+1}\right)^k. \end{aligned} \tag{2.5}$$

As an aside if the prior on θ were taken to be an *improper prior*, $\pi(\theta) = 1$, $\theta > 0$, then $P(X \geqslant x+u; X \geqslant x) = (x/(x+u))$. This assessment of residual life is similar, but not identical, to that of Eq. (2.5) with $k = 1$.

Suppose now that one were to not subscribe to the principle of conditionality and chose $\mathscr{L}(\theta; X \geqslant x) = c$; i.e. the likelihood is a constant $c > 0$. Then the posterior of θ would equal its prior, and Eq. (2.5) would become $(u+1)^{-k}$. Here the effect of x vanishes, because in choosing a flat likelihood one essentially says that irrespective of what x is, an equal weight is given to all values of θ. Clearly, this choice for a likelihood is not appealing. However, the following choice for $\mathscr{L}(\theta; X \geqslant x)$ appears to be a more sensible alternative.

Suppose that instead of choosing $\mathscr{L}(\theta; X \geqslant x) = \exp(-\theta x)$—a decreasing function of θ—one were to choose $\mathscr{L}(\theta; X \geqslant x) = \exp(-\theta\beta x)$, for some $\beta > 0$. The likelihood would still be a decreasing function of θ, but the rate of decrease would vary, depending on the value of β; see Fig. 2.

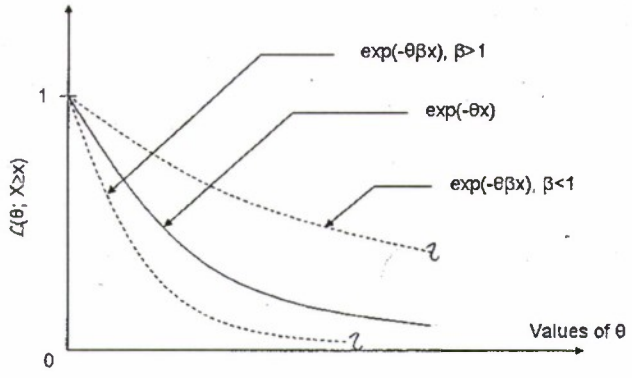

Fig. 2. Likelihood of θ with $X>x$ fixed.

For $\beta>0$, Eq. (2.5) would become

$$P(X\geqslant x+u; X\geqslant x)=\left(\frac{\beta x+1}{\beta x+u+1}\right)^{k}, \tag{2.6}$$

so that the introduction of a β in the likelihood tantamounts to assigning a weight β to the observed value of x. This in some scenarios could be a desirable feature to have, say when the accuracy (i.e. the credibility) of the observed x is suspect. The choice $\beta>(<)1$ would inflate (deflate) x, and this in turn would cause the likelihood to decay faster (slower) than the conventional $\exp(-\theta x)$. Since θ is the reciprocal of the mean time to failure, accentuating large values of θ, as the choice $\beta<1$ would tend to do, boils down to accentuating small values of the mean time to failure and thence small values of the residual life. Similarly with $\beta>1$. The choice $\beta=1$ encapsulates full faith in the observed x and also an adherence to the principle of conditionality. Eqs. (2.5) and (2.6) support our claim that the introduction of a parametric model increases the transparency of the point we are trying to make.

2.3.1. Discussion: the advantage of parametric models

Parametric models are used because they facilitate a coherent updating of the assessed uncertainties via a mechanistic application of Bayes' formula. The example of Section 2.3.2 underscores this point. By contrast, the direct approach of Section 2.2 requires of the user a fresh specification of the likelihood every time new evidence becomes available. This process, besides being cumbersome, has the danger of leading one to incoherence should one not be thoughtful about one's specifications. The disadvantage of parametric models is that the chosen model may not be an accurate reflection of reality. All the same the computational advantage offered by parametric models outweighs the disadvantage of misspecification, and thus their common use.

2.3.2. Application to survival time data on winding life

To illustrate the workings of the material of this section we consider here some service life data on "field windings" of generators given by Nelson (2000). The data below, abstracted from Nelson (2000, Table 1), consists of months in service of failed and unfailed windings. The 16 ranked failures and survival times—with the former tagged by an asterisk—in months are

$$31.7^{*}, 39.2^{*}, 57.5^{*}, 65.0, 65.8^{*}, 70.0^{*}, 75.0, 75.0, 87.5, 88.3, 94.2, 101.7, 105.8^{*}, 109.2, 110.0^{*}, \text{ and } 130.0.$$

Observe that seven out of the 16 field windings have experienced failures and of the nine that have not the largest (smallest) service life is 130 (65) months. Suppose that for the purposes of planning for maintenance, we are interested in the probability of any one of the surviving units not failing for an additional $u>0$ months. For the sake of discussion let us pick the unit with the largest accumulated life. That is, we need to assess $P(X\geqslant 130+u;\mathbf{d})$, where $\mathbf{d}$ denotes the life history data given above.

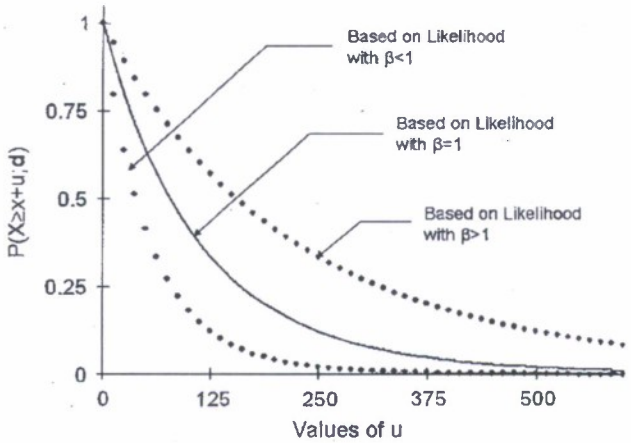

Fig. 3. Probability of the longest surviving unit surviving an additional u months.

Assuming that $P(X \geqslant x|\theta) = \exp(-\theta x)$, with a gamma prior for θ with scale (shape) parameter $1(k)$, it can be verified that under an adherence to the principle of conditionality, the posterior distribution of θ is also a gamma with scale $(\sum_1^m x_i + \sum_1^n t_i + 1)$, and shape $k+n$, where $\sum_1^m x_i$ is the sum of the m survival times and $\sum_1^n t_i$ is the sum of the n failure times. When such is the case, we have—as an analogue to Eq. (2.5)—the result that for any unfailed unit, that has experienced a service life of x,

$$P(X \geqslant x+u;\bullet) = \left(\frac{\sum_1^m x_i + \sum_1^n t_i + 1}{\sum_1^m x_i + \sum_1^n t_i + u + 1}\right)^{k+n}. \tag{2.7}$$

Eq. (2.7) when invoked—for $k=5$—in the context of the surviving unit with an accumulated service life of 130 months and the life history data given above yields, for $u \geqslant 0$,

$$P(X \geqslant 130+u;\mathbf{d}) = \left(\frac{1306.9}{1306.9+u}\right)^{12}. \tag{2.8}$$

A plot of $P(X \geqslant 130+u;\mathbf{d})$ versus u, for $u \geqslant 0$, is shown as the bold faced curve of Fig. 3.

Were the principle of conditionality not adhered to and the likelihood function be modulated by the constant $\beta > 0$, then our analogue to Eq. (2.6) would be

$$P(X \geqslant x+u;\bullet) = \left(\frac{\beta(\sum_1^m x_i + \sum_1^n t_i) + 1}{\beta(\sum_1^m x_i + \sum_1^n t_i) + u + 1}\right)^{k+n}. \tag{2.9}$$

Eq. (2.9) when invoked in the context of the scenario leading up to Eq. (2.8) for $\beta = \frac{1}{2}$ and 2 would result in the dotted curves of Fig. 3. Our assessed survival probability depends on the form chosen for the likelihood. In principle, likelihood plays a more crucial role than the prior, because whereas the prior gets updated with new evidence, the likelihood stays put from the start.

3. Conclusion

The innocuously simple problem of assessing conditional probabilities can get riddled with issues, both philosophical and technical, when the conditioning event becomes a reality. The cleanest way to approach it is through Bayes' Law. When this is done it can be seen that the standard answer arises as a special case under the assumption of a constant likelihood. Other forms of the likelihood will lead to other answers. Since the choice of a likelihood is an assessors prerogative—just like the choice of a probability model—there is no unique and correct way to bet on residual life. However, the traditional answer (presumably the one that will be subscribed to by *card carrying probabilists*) will be the correct and unique answer, but only when its argument is sheltered under the philosophical (or behaviouristic) principle of conditionalization.

Acknowledgements

The comments of a referee are acknowledged. The referee drew attention to Joyce (1999) and made us aware that matters discussed here have been considered by philosophers of science. Supported by Grants N00014-06-1-037 by Office of Naval Research and W911NF-05-1-2009 by the US Army Research Office.

References

Basu, D., 1975. Statistical information and likelihood. Sankhya Ser. A 37A (Pt. 1), 1–71.
Basu, D., 1982. A note on likelihood. In: Symposio Nacional de Probabilidadc e Estatisticia, Universidade De Sao Paolo, pp. 1–5.
de Finetti, 1937. La Prevision: ses Lois Logiques, ses Sources Subjectives. Ann. Inst. H. Poincare (Paris) 7, 1–68.
de Finetti, B., 1972. Probability, Induction and Statistics. Wiley, New York.
de Finetti, B., 1975. Theory of Probability, vol. 2. Wiley, New York.
Diaconis, P., Zabell, S.L., 1982. Updating subjective probabilities. J. Amer. Statist. Assoc. 77 (380), 822–830.
Edwards, A.W.F., 1992. Likelihood. The Johns Hopkins Press, Baltimore, MD.
Fisher, R.A., 1912. On an absolute criterion for fitting frequency curves. Mess. Math. 41, 155–160.
Freedman, D., Purves, R., 1969. Bayes method for bookies. Ann. Math. Statist. 40, 1177–1186.
Hacking, I., 1967. Slightly more realistic personal probability. Philos. Sci. 34, 311–325.
Howson, C., Urbach, P., 1989. Scientific Reasoning: The Bayesian Approach. Open Co., La Salle, IL.
Joyce, J., 1999. The Foundations of Causal Decision Theory. Cambridge University Press, Cambridge.
Jonsson, C., 1974. Electron diffraction at multiple slits. Amer. J. Phys. 42, 4–11.
Nelson, W., 2000. Theory and applications of hazard plotting for censored failure data. Technometrics 42 (1), 12–25.
Ramsey, F.P., 1931. Truth and probability. In: Braithwaite, R.G. (Ed.), The Foundations of Mathematics and Other Logical Essays. Routledge and Kegan Paul, London, pp. 156–198.
Singpurwalla, N.D., 2006a. On competing risk and degradation processes. In: Rojo, J. (Ed.), Optimality—The Second Erich Lehmann Symposium. IMS Lecture Notes—Monograph Series, vol. 49, pp. 289–304.
Singpurwalla, N.D., 2006b. Reliability and Risk: A Bayesian Perspective. Wiley, New York.
Teller, P., 1976. Conditionalization, observation, and change of preference. In: Harper, W.L., Hooker, C.A. (Eds.), Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, vol. 1. D. Deidel, Dordrecht, pp. 205–253.

26

A Bayesian Ponders "The Quality of Life"

Mounir Mesbah[1] and Nozer D. Singpurwalla[2]

[1]*Laboratoire de Statistique Théorique et Appliquée, Université Paris 6, Paris, France*

[2]*Department of Statistics, The George Washington University, Washington, DC, USA*

Abstract: The notion of quality of life (QoL) has recently received a high profile in the biomedical, the bioeconomic, and the biostatistical literature. This is despite the fact that the notion lacks a formal definition. The literature on QoL is fragmented and diverse because each of its constituents emphasizes its own point of view. Discussions have centered around ways of defining QoL, ways of making it operational, and ways of making it relevant to medical decision making. An integrated picture showing how all of the above can be brought together is desirable. The purpose of this chapter is to propose a framework that does the above. This we do via a Bayesian hierarchical model. Our framework includes linkages with item response theory, survival analysis, and accelerated testing. More important, it paves the way for proposing a definition of QoL.

This is an expository chapter. Our aim is to provide an architecture for conceptualizing the notion of QoL and its role in health care planning. Our approach could be of relevance to other scenarios such as educational, psychometric, and sociometric testing, marketing, sports science, and quality assessment.

Keywords and Phrases: Health care planning, hierarchical modeling, information integration, survival analysis, quality control, utility theory

26.1 Introduction and Overview

A general perspective on the various aspects of the QoL problem can be gained from the three-part paper of Fitzpatrick *et al.* (1992). For an appreciation of the statistical issues underlying QoL, the recent book by Mesbah, *et al.* (2002) is a good starting point. In the same vein is the paper of Cox *et al.* (1992) with the striking title, "Quality of Life Assessment: Can We Keep It Simple?" Reviewing the above and other related references on this topic, it is our position that QoL assessment can possibly be kept simple, but not too simple! To get a sense as to why we come upon this view, we start by selectively quoting phrases from

(*i*) "Many instruments reflect the multidimensionality of QoL," Fitzpatrick *et al.* (1992).

(*j*) "Summing disparate dimensions is not recommended, because contrary trends for different aspects of QoL are missed," Fitzpatrick *et al.* (1992).

(*k*) "In health economics QoL measures have ... more controversially (become) the means of prioritizing funding," Fitzpatrick *et al.* (1992).

(*l*) "The best understood application of QoL measures is in clinical trials, where they provide evidence of the effects of interventions," Fitzpatrick *et al.* (1992).

There is a variant of the notion of QoL, namely, the quality adjusted life (QAL). This variant is designed to incorporate the QoL notion into an analysis of survival data and history. A motivation for introducing QAL has been the often expressed view that medical interventions may prolong life, but that the discomfort that these may cause could offset any increase in longevity. The following four quotes provide some sense of the meaning of QAL.

(*m*) "QAL is an index combining survival and QoL...," Fitzpatrick *et al.* (1992).

(*n*) "QAL is a measure of the medical and psychological adjustments needed to induce an affordable QoL for patients undergoing problems," Sen (2002).

(*o*) "QAL is a patients' survival time weighted by QoL experience where the weights are based on utility values – measured on the unit interval," Cole and Kilbridge (2002).

(*p*) "QAL has emerged as an important yardstick in many clinical studies; this typically involves the lifetime as the primary endpoint with the incorporation of QAL or QoL measures through appropriate utility scores that are obtained through appropriate item analysis schemes," cf. Zhao and Tsiatis (2000).

26.1.2 Overview of this chapter

The above quotes encapsulate the essence of the QoL and its variant, the QAL. They indicate the diverse constituencies that are attracted to a QoL metric and the controversies that each constituency raises. For our objectives, the quotes provide ingredients for proposing a definition of QoL and developing a metric for measuring it. As a first step, it appears to us that any satisfactory discourse on QoL should encompass the involvement of three interest groups, the clinicians, the patients (or their advocates), and an economic entity, such as managers of

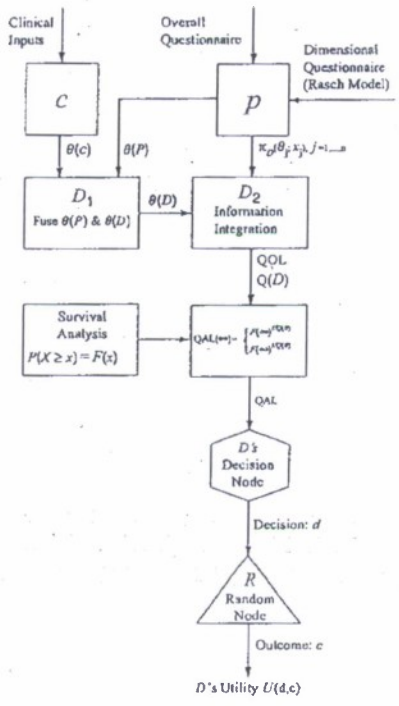

Figure 26.1. $\mathcal{D}$'s decision tree using QAL consideration (the unifying perspective of QAL).

The quantities $\theta(\mathcal{P})$, $\theta(\mathcal{C})$, and $\theta(\mathcal{D})$ are explained later in Sections 26.3 through 26.5. The hexagon denotes $\mathcal{D}$'s decision node and the triangle is a random node $\mathcal{R}$. At the decision node $\mathcal{D}$ takes one of several possible actions available to $\mathcal{D}$; let these actions be denoted by a generic d. At $\mathcal{R}$, we would see the possible outcomes of decision d. The quantity $U(d, \mathrm{c})$ at the terminus of the tree represents to $\mathcal{D}$ the utility of a decision d when the outcome is c. With medical decisions it is often the case that d influences c.

The quantity $Q(\mathcal{D})$ is $\mathcal{P}$'s QoL assessed by $\mathcal{D}$ subsequent to fusing the inputs of $\mathcal{P}$ and $\mathcal{C}$; $Q(\mathcal{D}) \in [0, 1]$. Let $P(X \geq x)$ denote $\mathcal{P}$'s *survival function*; this is assessed via survival data history on individuals judged exchangeable with $\mathcal{P}$, plus other covariate information that is specific to $\mathcal{P}$. Together with $P(X \geq x)$ and $\theta(\mathcal{D})$, $\mathcal{D}$ is able to assess $\mathcal{P}$'s QAL. There are two strategies for doing this. One is through the accelerated life model whereby $\mathrm{QAL}(x) = P(XQ(\mathcal{D}) \geq x)$. The other is via a proportional life model whereby $\mathrm{QAL}(x) = (P(X \geq x))^{1/Q(\mathcal{D})}$. Note that the QAL metric is, like the survival function, indexed by x. The effect of both of the above is to dampen the survival function of the

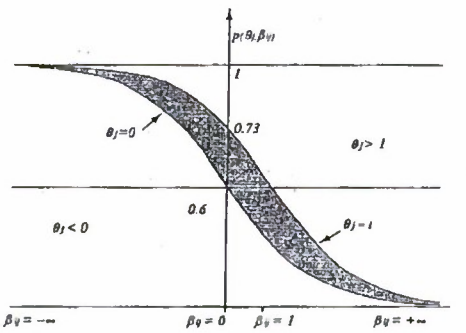

Figure 26.2. Envelope showing the range of values for $p(\theta_j, \beta_{ij})$.

such omnibus questions generate a response on a multinomial scale, but here we assume that $\mathcal{P}$'s response takes values in the continuum $[0, 1]$, with 1 denoting excellent. Let $\theta(\mathcal{P})$ denote $\mathcal{P}$'s response to an omnibus question.

26.3.1 The case of a single dimension: $\mathcal{D}$'s assessment of θ_j

Given the responses $\underset{\sim}{x}_j = (x_{1j}, \ldots, x_{kj})$ to a set of k questions pertaining to dimension j, the likelihood of θ_j and $\underset{\sim}{\beta}_j = (\beta_{1_j}, \ldots, \beta_{k_j})$ under the Rasch model is

$$\mathcal{L}(\theta_j, \underset{\sim}{\beta}_j; \underset{\sim}{x}_j) = \prod_{i=1}^{k} \frac{e^{x_{ij}(\theta_j - \beta_{ij})}}{1 + e^{\theta_j - \beta_{ij}}}, \tag{26.1}$$

for $\theta_j \in [0, 1]$ and $-\infty < \beta_{1j} < \cdots < \beta_{kj} < +\infty$.

If we suppose, as is reasonable to do so, that θ_j and $\underset{\sim}{\beta}_j$ are a priori independent with $\pi(\theta_j)$ and $\pi(\underset{\sim}{\beta}_j)$ denoting their respective prior densities, then by treating $\underset{\sim}{\beta}_j$ as a nuisance parameter and integrating it out, the posterior distribution of θ_j is

$$\pi(\theta_j; \underset{\sim}{x}_j) \propto \int_{\underset{\sim}{\beta}_j} \mathcal{L}(\theta_j, \underset{\sim}{\beta}_j; \underset{\sim}{x}_j) \pi(\theta_j) \pi(\underset{\sim}{\beta}_j) d\underset{\sim}{\beta}_j. \tag{26.2}$$

The question now arises as to what should $\pi(\theta_j)$ and $\pi(\underset{\sim}{\beta}_j)$ be? In order to answer this question we first need to ask who specifies these priors, $\mathcal{P}$, $\mathcal{C}$, or $\mathcal{D}$? The answer has to be either $\mathcal{C}$ or $\mathcal{D}$, because $\mathcal{P}$ cannot satisfy a prior and then respond to a questionnaire. Furthermore, in principle, these priors have to be $\mathcal{D}$'s priors because it is $\mathcal{D}$'s decision process that we are describing. Thus,

The quantity $\mathcal{L}(\theta(\mathcal{D});\theta(\mathcal{P}),\theta(\mathcal{C}))$ denotes $\mathcal{D}$'s likelihood that $\mathcal{P}$ will declare a $\theta(\mathcal{P})$, and $\mathcal{C}$ will declare a $\theta(\mathcal{C})$, were $\theta(\mathcal{D})$ to be a measure of $\mathcal{P}$'s overall quality of life. This likelihood will encapsulate any biases that $\mathcal{P}$ and $\mathcal{C}$ may have in declaring their $\theta(\mathcal{P})$ and $\theta(\mathcal{C})$, respectively, as perceived by $\mathcal{D}$, and also any correlations between the declared values by $\mathcal{P}$ and $\mathcal{C}$. The nature of this likelihood remains to be investigated. The quantity $\pi_{\mathcal{D}}(\theta(\mathcal{D}))$ is $\mathcal{D}$'s prior for $\theta(\mathcal{D})$, and following our previous convention, we assume that it is uniform on $[0,1]$. This completes our discussion on $\mathcal{D}$'s assessment of $\theta(\mathcal{D})$. It involves a $\theta(\mathcal{P})$, $\theta(\mathcal{C})$ and connotes information integration by $\mathcal{D}$ at one level.

26.4.2 Encoding the positive dependence between the θ_js

One way to capture the positive dependence between the θ_js is through mixtures of independent sequences. Specifically, we suppose, as if is reasonable to do so, that given $\theta(\mathcal{D})$ the θ_js are independent, with θ_j having a probability density function of the form $f_{\mathcal{D}}(\theta_j|\theta(\mathcal{D}))$, $j=1,\ldots,m$. The subscript $\mathcal{D}$ associated with f denotes the fact that the probability density in question is that of $\mathcal{D}$. A strategy for obtaining $f_{\mathcal{D}}(\theta_j|\theta(\mathcal{D}))$ is described later, subsequent to Equation (26.5).

With $\pi_{\mathcal{D}}(\theta_j;\underset{\sim}{x}_j)$, $j=1,\ldots,m$, and $\widehat{\pi}_{\mathcal{D}}(\theta(\mathcal{D}))$ at hand, $\mathcal{D}$ may extend the conversation to $\theta(\mathcal{D})$ and obtain the joint distribution of $\theta_1,\ldots,\theta_m$ as

$$\begin{aligned} &P_{\mathcal{D}}(\theta_1,\ldots,\theta_m;\underset{\sim}{x}_{j1},\ldots,\underset{\sim}{x}_{jm},\theta(\mathcal{P}),\theta(\mathcal{C})) \\ &\quad = \int_{\theta(\mathcal{D})} P(\theta_1,\ldots,\theta_m|\theta(\mathcal{D});\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m)\widehat{\pi}_{\mathcal{D}}(\theta(\mathcal{D}))d\theta(\mathcal{D}); \end{aligned} \tag{26.4}$$

in writing out the above, we have assumed that the $\underset{\sim}{x}_j$s, $j=1,\ldots,m$, have no bearing on $\theta(\mathcal{D})$, once $\theta(\mathcal{P})$ and $\theta(\mathcal{C})$ have been declared by $\mathcal{P}$ and $\mathcal{C}$, respectively. Applying the multiplication rule, and supposing that the $\underset{\sim}{x}_i$s, $i\neq j$ have no bearing on θ_j, $j=1,\ldots,m$, the right-hand side of the above equation becomes

$$\int_{\theta(\mathcal{D})} \prod_{j=1}^{m} f_{\mathcal{D}}(\theta_j|\theta(\mathcal{D});\underset{\sim}{x}_j)\widehat{\pi}_{\mathcal{D}}(\theta(\mathcal{D}))d\theta(\mathcal{D}). \tag{26.5}$$

We now invoke Bayes' law to write

$$f_{\mathcal{D}}(\theta_j|\theta(\mathcal{D});\underset{\sim}{x}_j) \propto f_{\mathcal{D}}(\theta(\mathcal{D})|\theta_j;\underset{\sim}{x}_j)\pi_{\mathcal{D}}(\theta_j;\underset{\sim}{x}_j),$$

where $f_{\mathcal{D}}(\theta(\mathcal{D})|\theta_j;\underset{\sim}{x}_j)$ is $\mathcal{D}$'s probability density of $\theta(\mathcal{D})$ were $\mathcal{D}$ to know θ_j, and in the light of $\underset{\sim}{x}_j$. A strategy for specifying this probability density is to suppose that $\theta(\mathcal{D})$ is uniform and symmetric around θ_j, with endpoints $\theta_j \pm \epsilon$,

There could be other possible ways for defining QoL. A few of these would be to consider $\min_j(\theta_j)$, $\max_j(\theta_j)$, or $mean_j(\theta_j)$, and to let QoL be a quantity such as

$$\text{QoL} = \mathcal{P}_{\mathcal{D}}(\min_j(\theta_j) \geq a)$$

for some $a \in [0, 1]$. Whereas the proposed definition(s) are appropriate in all situations, it is not clear whether a unique definition of QoL is palatable to all constituents. We see some merits to having a unique yardstick.

26.6 Summary and Conclusions

In this chapter we have proposed an approach for addressing a contemporary problem that can arise in many scenarios, the one of interest to us coming from the health sciences vis-a-vis the notion of "quality of life." What seems to be common to these scenarios is information from diverse sources that needs to be integrated, considerations of multidimensionality, and the need to make decisions whose consequences are of concern. Previous work on problems of this type has been piecemeal with statisticians mainly focusing on the frequentist aspects of item response models. Whereas such approaches have the advantages of "objectivity", they do not pave the path of integrating information from multiple sources. The approach of this chapter is based on a hierarchical Bayesian architecture. In principle, our architecture is able to do much, if not all, that is required by the users of QoL indices. The architecture also leads to a strategy by which QoL can be defined and measured in a formal manner. The current literature on this topic does not address the matter of definition. This chapter is expository in the sense that it outlines an encompassing and unifying approach for addressing the QoL and QAL problem. The normative development of this chapter has the advantage of coherence. However, this coherence is gained at the cost of simplicity. Some multidimensional priors with a restricted sample space are involved, and these remain to be articulated. So do some likelihoods. Finally, there is the matter of computations. However, all these limitations are only of a technical nature and these can eventually be addressed. We are continuing our work on such matters, including an application involving real data and real scenarios. The purpose of this chapter was to show how a Bayesian approach can address a contemporary problem, and the overall strategy that can be used to develop such an approach. The novel aspects of this chapter are: the conceptualization of the QoL problem as a scenario involving three groups of individuals, a structure whereby information from several sources can be integrated, and a definition of the notion of QoL.

7. Sen, P. K. (2002). Measures of quality adjusted life and quality of life deficiency: Statistical perspectives, In *Statistical Methods for Quality of Life Studies* (Eds., M. Mesbah, B. Cole, and M. T. Lee), pp. 255–266, Kluwer, Boston.

8. Slevin, M., Plant H., Lynch D., Drinkwater I., and Gregory, W. M. (1988). Who should measure quality of life, the doctor or the patient?, *British Journal of Cancer*, 57, 109–112.

9. WHOQoL Group (1994). The development of the World Health Organization quality of life assessment instrument, In *Quality of Life Assessment: International Perspectives* (Eds., J. Orley and W. Kuyken), Springer-Verlag, Heidelberg, Germany.

10. Zhao, H. and Tsiatis, A. A. (2000). Estimating mean quality of lifetime with censored data, *Sankhyā, Series B*, 62, 175–188.

Statistical Practice

Choosing a Coverage Probability for Prediction Intervals

Joshua LANDON and Nozer D. SINGPURWALLA

Coverage probabilities for prediction intervals are germane to filtering, forecasting, previsions, regression, and time series analysis. It is a common practice to choose the coverage probabilities for such intervals by convention or by astute judgment. We argue here that coverage probabilities can be chosen by decision theoretic considerations. But to do so, we need to specify meaningful utility functions. Some stylized choices of such functions are given, and a prototype approach is presented.

KEY WORDS: Confidence intervals; Decision making; Filtering; Forecasting; Previsions; Time series; Utilities.

1. INTRODUCTION AND BACKGROUND

Prediction is perhaps one of the most commonly undertaken activities in the physical, the engineering, and the biological sciences. In the econometric and the social sciences, prediction generally goes under the name of *forecasting*, and in the actuarial and the assurance sciences under the label *life-length assessment*. Automatic process control, filtering, and quality control, are some of the engineering techniques that use prediction as a basis of their modus operandus.

Statistical techniques play a key role in prediction, with regression, time series analysis, and dynamic linear models (also known as state space models) being the predominant tools for producing forecasts. The importance of statistical methods in forecasting was underscored by Pearson (1920) who claimed that prediction is the "fundamental problem of practical statistics." Similarly, with de Finetti (1972, Chaps. 3 and 4), who labeled prediction as "prevision," and made it the centerpiece of his notion of "exchangeability" and a subjectivistic Bayesian development around it. In what follows, we find it convenient to think in terms of regression, time series analysis, and forecasting techniques as vehicles for discussing an important aspect of prediction.

Joshua Landon is Post Doc, and Nozer D. Singpurwalla is Professor, Department of Statistics and Department of Decision Sciences, The George Washington University, Washington, DC 20052 (E-mail: *nozer@gwu.edu*). Supported by ONR Contract N00014-06-1-0037 and the ARO Grant W911NF-05-1-0209. The student retention problem was brought to our attention by Dr. Donald Lehman. The detailed comments of three referees and an Associate Editor have broadened the scope of the article. Professor Fred Joust made us aware of the papers by Granger, and by Tay and Wallis.

We start by noting that inherent to the above techniques is an underlying distribution (or error) theory, whose net effect is to produce predictions with an uncertainty bound; the normal (Gaussian) distribution is typical. An exception is Gardner (1988), who used a Chebychev inequality in lieu of a specific distribution. The result was a *prediction interval* whose width depends on a coverage probability; see, for example, Box and Jenkins (1976, p. 254), or Chatfield (1993). It has been a common practice to specify coverage probabilities by convention, the 90%, the 95%, and the 99% being typical choices. Indeed Granger (1996) stated that academic writers concentrate almost exclusively on 95% intervals, whereas practical forecasters seem to prefer 50% intervals. The larger the coverage probability, the wider the prediction interval, and vice versa. But wide prediction intervals tend to be of little value [see Granger (1996), who claimed 95% prediction intervals to be "embarassingly wide"]. By contrast, narrow prediction intervals tend to be risky in the sense that the actual values, when they become available, could fall outside the prediction interval. Thus, the question of what coverage probability one should choose in any particular application is crucial.

1.1 Objective

The purpose of this article is to make the case that the choice of a coverage probability for a prediction interval should be based on decision theoretic considerations. This would boil down to a trade-off between the utility of a narrow interval versus the disutility of an interval that fails to cover an observed value. It is hoped that our approach endows some formality to a commonly occurring problem that seems to have been traditionally addressed by convention and judgment, possibly because utilities are sometimes hard to pin down.

1.2 Related Issues

Before proceeding, it is important to note that in the context of this article, a prediction interval is not to be viewed as a *confidence interval*. The former is an estimate of a future observable value; the latter an estimate of some fixed but unknown (and often unobservable) parameter. Prediction intervals are produced via frequentist or Bayesian methods, whereas confidence intervals can only be constructed via a frequentist argument. The discussion of this article revolves around prediction intervals produced by a Bayesian approach; thus we are concerned here with *Bayesian prediction intervals*. For an application of frequentist prediction intervals, the article by Lawless and Fredette (2005)

  DOI: 10.1198/000313008X304062

is noteworthy; also the book by Hahn and Meeker (1991, Sect. 2.3), or the article of Beran (1990).

A decision theoretic approach for specifying the confidence coefficient of a confidence interval is not explored here. All the same, it appears that some efforts in this direction were embarked upon by Lindley and Savage [see Savage (1962), p. 173, who also alluded to some work by Lehmann (1958)]. By contrast, a decision theoretic approach for generating prediction intervals has been alluded to by Tay and Wallis (2000) and developed by Winkler (1972). However, Winkler's aim was not the determination of optimal coverage probabilities, even though the two issues of coverage probability and interval size are isomorphic. Our focus on coverage probability is dictated by its common use in regression, time series analysis, and forecasting.

Finally, predictions and prediction intervals should not be seen as being specific to regression and time series based models. In general they will arise in the context of any probability models used to make previsions, such as the ones used in reliability and survival analysis [see Singpurwalla (2006), Chap. 5].

2. MOTIVATING EXAMPLE

Our interest in this problem was motivated by the following scenario. For purposes of exposition, we shall anchor on this scenario.

A university wishes to predict the number of freshman students that will be retained to their sophomore year. Suppose that N is the number of freshman students, and X is the number retained to the sophomore year; $X \leq N$. Knowing N, the university wishes to predict X. The prediction is to be accompanied by a prediction interval, and the focus of this article pertains to the width of the interval. The width of the interval determines the amount of funds the university needs to set aside for meeting the needs of the sophomore students. The wider the interval, the greater the reserves; however, large reserves strain the budget. By contrast, the narrower the interval the greater is the risk of the actual number of sophomores falling outside the interval. This would result in poor budgetary planning due to insufficient or excessive reserves. Thus, a trade-off between the risks of over-budgeting and under-budgeting is called for.

The student retention scenario is archetypal because it arises in several other contexts under different guises. A direct parallel arises in the case of national defense involving an all-volunteer fighting force. Meaningful predictions of the retention of trained personnel are a matter of national security. A more classical scenario is the problem of inventory control wherein a large volume of stored items ties up capital, whereas too little inventory may result in poor customer satisfaction or emergency actions; see, for example, Hadley and Whitin (1960, Chap. 4). Another (more contemporary) scenario comes from the Basel II accords of the banking industry. Bank regulators need to assess how much capital a bank needs to set aside to guard against financial risks that a bank may face; see Decamps, Rochet, and Roger (2004) for an appreciation. From the biomedical and the engineering sciences arises the problem of predicting survival times subsequent to a major medical intervention or a repair.

In all the above scenarios, the width of the prediction interval is determined by the nature of an underlying probability model and its coverage probability. This point is best illustrated by a specific assumption about the distribution of the unknown X; this is done next. But before doing so, it is necessary to remark that neither the literature on inventory control, nor that on Basel II accords, addresses the issue of optimal coverage probabilities. In the former case, a possible reason could be the difficulties associated with quantifying customer dissatisfaction.

2.1 Distributional Assumptions

Suppose that the (posterior) predictive distribution of X obtained via a regression or a time series model is a normal (Gaussian) with a mean μ and variance σ^2, where μ and σ^2 have been pinned down; the normal distribution is typical in these contexts. Then, it is well known [see De Groot (1970), p. 228] that under a squared error loss for prediction error, μ is the best predictor of X. For a coverage probability of $(1 - \alpha)$, a prediction interval for X may be of the form $\mu \pm z_{\alpha/2}\sigma$. Here $z_{\alpha/2}$ is such that for some random variable W having a standard normal distribution, $P(W \geq z_{\alpha/2}) = \alpha/2$.

The question that we wish to address in this article is, what should α be? A small α will widen the prediction interval diminishing its value to a user. Indeed, $\alpha = 0$ will yield the embarrassing $(-\infty, +\infty)$ as a prediction interval. By contrast, with large values of α, one runs the risk of the prediction interval not covering the actual value (when it materializes). Thus, we need to determine an optimum value of α to use. To address the question posed, we need to introduce utilities, one for the worth of a prediction interval, and the other, a disutility, for the failure of coverage.

3. CANDIDATE UTILITY FUNCTIONS

Utilities are a key ingredient of decision making, and the principle of maximization of expected utility prescribes the decision (action) to be taken; see, for example, Lindley (1985, p. 71). Utilities measure the worth of a consequence to a decision maker, and disutilities the penalty (or loss) imposed by a consequence. With disutilities, a decision maker's actions are prescribed by the principle of minimization of expected disutilities. The unit of measurement of utilities is a "utile." However, in practice utilities are measured in terms of monetary units, such as dollars, and this is what we shall assume.

In the context of prediction, we make the natural assumption that, in principle, one prefers a prediction interval of width zero over any other prediction interval. This makes the utility of any prediction interval of nonzero width a disutility. Similarly, the failure of any prediction interval to cover an observed value results in a disutility. Following Winkler (1972), the two disutilities mentioned above are assumed to be additive, though this need not be so. Thus, for the scenario considered here, one endeavors to choose that value of α for which the total expected disutility is a minimum.

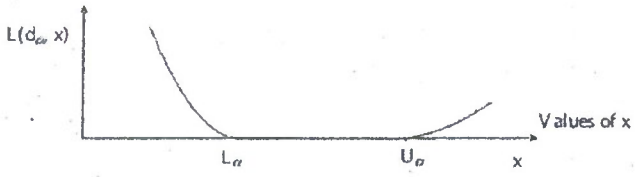

Figure 1. The disutility of noncoverage.

3.1 The Disutility of a Prediction Interval

The width d_α of a prediction interval of the type described in Section 2.1 is $d_\alpha = 2z_{\alpha/2}\sigma$; here the coverage probability is $(1-\alpha)$. Let $c(d_\alpha)$ be the disutility (or some kind of a dollar penalty) associated with a use of d_α. Clearly $c(d_\alpha)$ should be zero when $d_\alpha = 0$, and $c(d_\alpha)$ must increase with d_α, since there is a disadvantage to using wide intervals. A possible choice for $c(d_\alpha)$ could be

$$c(d_\alpha) = d_\alpha^\beta, \tag{1}$$

for $\beta > 0$. When $\beta < 1$, $c(d_\alpha)$ is a concave increasing function of d_α, and when $\beta > 1$, $c(d_\alpha)$ is convex and increasing in d_α. The choice of what β must be depends on the application. In certain applications, such as target tracking, $\beta < 1$ may be more desirable than $\beta > 1$; in others, such as econometric forecasting, a convex disutility function may be suitable. The choice of (3.1) for a disutility function is purely illustrative. The proposed approach is not restricted to any particular choice for $c(d_\alpha)$.

3.2 The Disutility of Noncoverage

A possible function for the disutility caused by a failure of the prediction interval to cover x, a realization of X, can be prescribed via the following line of reasoning.

Suppose that $U_\alpha = \mu + z_{\alpha/2}\sigma$ is the upper bound, and $L_\alpha = \mu - z_{\alpha/2}\sigma$, the lower bound of the $(1-\alpha)$ probability of coverage prediction interval. Let $L(d_\alpha, x)$ denote the disutility or penalty loss (in dollars) in using a prediction interval of width d_α when X reveals itself as x. Then $L(d_\alpha, x)$ could be of the form

$$L(d_\alpha, x) = \begin{cases} f_1(x - U_\alpha), & x > U_\alpha, \\ 0, & L_\alpha < x < U_\alpha, \\ f_2(L_\alpha - x), & x < L_\alpha, \end{cases} \tag{2}$$

where f_1 and f_2 are increasing functions of their arguments, which encapsulate the penalty of x overshooting and undershooting the prediction interval, respectively.

As illustrated in Figure 1, the said functions will generally be convex and increasing because a narrow miss by the interval will matter less than a large miss. Furthermore, these functions need not be symmetric. For example, as shown in Figure 1, the penalty for undershooting the interval is assumed to be more severe than that of overshooting.

3.3 The Expected Total Disutility

With $c(d_\alpha)$ and $L(d_\alpha, x)$ thus specified, there remains one caveat that needs to be addressed. When the α is chosen, the value of x is not known and thus $L(d_\alpha, x)$ needs to be averaged over the possible values that x can take. This is easy to do because the predictive distribution of X has to be specified. Accordingly, let

$$R(d_\alpha) = E_X[L(d_\alpha, x)], \tag{3}$$

be the expected value of $L(d_\alpha, x)$. In decision theory, $R(d_\alpha)$ is known as the *risk function*; it is free of X. $R(d_\alpha)$ encapsulates the risk of noncoverage by an interval of width d_α, with $R(d_\alpha)$ decreasing in d_α.

Since $c(d_\alpha)$ is devoid of unknown quantities—indeed d_α is a decision variable—the matter of taking an expectation of $c(d_\alpha)$ is moot. We may now combine $c(d_\alpha)$ and $R(d_\alpha)$ to obtain the *total expected disutility function* as

$$D(d_\alpha) = c(d_\alpha) + R(d_\alpha). \tag{4}$$

As mentioned before, the additive choice, albeit natural, is not binding. We choose that value of α for which $D(d_\alpha)$ is a minimum. This is described next.

4. CHOOSING AN OPTIMUM COVERAGE PROBABILITY

To make matters concrete, suppose that $c(d_\alpha) = \sqrt{d_\alpha}$, so that the β of Equation (1) is 1/2. Also, since $d_\alpha = 2z_{\alpha/2}\sigma$, $U_\alpha = \mu + z_{\alpha/2}\sigma$ can be written as $U_\alpha = \mu + d_\alpha/2$; similarly, $L_\alpha = \mu - d_\alpha/2$.

For the f_1 and f_2 of Equation (2), we let $f_1(x - U_\alpha) = (x - U_\alpha)^2/40$ and $f_2(L_\alpha - x) = (L_\alpha - x)^2/10$. These choices encapsulate a squared-error disutility, and make f_1 and f_2 asymmetric with respect to each other. Writing U_α and L_α in terms of d_α, we have $f_1(x - U_\alpha) = (x - \mu - d_\alpha/2)^2/40$, and $f_2(L_\alpha - x) = (\mu - d_\alpha/2 - x)^2/10$.

To compute the risk function of Equation (3) we need to specify μ and σ^2 of the normal distribution of X. Based on a Bayesian time series analysis of some student retention data, these were determined to be $\mu = 2140$ and $\sigma^2 = 396$. With the above in place, we may compute the total expected disutility as

$$D(d_\alpha) = \sqrt{d_\alpha} + R(d_\alpha),$$

where

$$R(d_\alpha) = \int_{\mu + d_\alpha/2}^{\infty} \frac{(x - \mu - d_\alpha/2)^2}{40} f(x)dx + \int_{-\infty}^{\mu - d_\alpha/2} \frac{(\mu - d_\alpha/2 - x)^2}{10} f(x)dx,$$

where $f(x)$ is the probability density at x of a normally distributed random variable with mean μ and variance σ^2.

The computation of $R(d_\alpha)$ has to be done numerically, and a plot of $D(d_\alpha)$ versus d_α, for $d_\alpha \geq 0$, is shown in Figure 2.

An examination of Figure 2 shows that $D(d_\alpha)$ attains its minimum at $d_\alpha = 62$. This suggests, via the relationship $d_\alpha =$

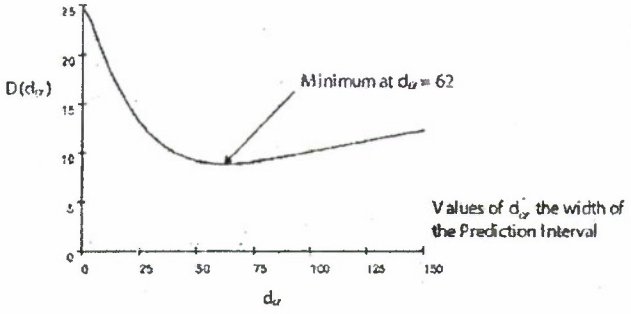

Figure 2. Total expected disutility versus d_α.

$2z_{\alpha/2}\sigma$ with $\sigma^2 = 396$, and a table look up in the standard normal distribution, that the optimal coverage probability for this scenario is 0.88. Using coverage probabilities other than $0.88 \approx 0.90$, say the conventional 0.95 or 0.99 would yield a wider interval but the utility of such intervals would be less than that provided by the 0.90 coverage probability.

5. GENERALITY OF THE APPROACH AND SOME CAVEATS

The proposed approach is general because it rests on the simple principle of minimizing $D(d_\alpha)$, the total expected disutility function—Equation (4). If $D(d_\alpha)$ attains a unique minimum, then a unique optimal coverage probability can be arrived upon. If the minimum is not unique, then several optimal coverage probabilities will result, and the user is free to choose any one of these. There could be circumstances under which $D(d_\alpha)$ will not attain a minimum, and the method will fail to produce an answer. The optimality conditions which ensure a minimum value of $D(d_\alpha)$ is a matter that needs to be formally addressed, but with $c(d_\alpha)$ monotonic and concave (or convex), and with $L(d_\alpha, x)$ U-shaped as shown in Figure 1, $D(d_\alpha)$ will indeed attain a minimum. The choice of $L(d_\alpha, x)$ prescribed in Equation (1) is quite general. It is easily adaptable to one-sided intervals, and also to the inventory and banking scenarios mentioned before. Furthermore, it is conventional in life-length prediction studies and in statistical inference wherein square error loss is a common assumption.

The assumed distribution of X with specified parameters plays two roles. One is to average out $L(d_\alpha, x)$ to produce the risk function $R(d_\alpha)$. In this role the choice of the distribution of X is not restrictive because its purpose here is to merely serve as a weighting function. Any well-known distribution can be used, especially when $R(d_\alpha)$ is obtained via numerical methods, as we have done with the normal. By contrast, frequentist prediction intervals that entail pivitol methods limit the choice of distributions. The second role played by the distribution of X, is to facilitate a relationship between d_α and α. In the case of the normal distribution with mean μ and variance σ^2, $d_\alpha = 2z_{\alpha/2}\sigma$; here μ does not matter. This type of relationship will arise with any symmetrical distribution, such as the Student's-t, the triangular, the uniform, the Laplace, etc. A relationship between d_α and α in the case of the exponential with scale λ turns out to be quite straightforward, indeed more direct than that encountered with the normal; specifically $d_\alpha = \frac{1}{\lambda}\log[(2-\alpha)/\alpha]$. By suitable transformations, the case of other skewed distributions such as the lognormal, the Weibull, and the chi-squared can be similarly treated. A difficult case in point is the Pareto distribution (popular in financial mathematics) wherein $P(X > x; \psi, \beta) = (\psi/(\psi + x))^\beta$. Here $d_\alpha = \psi[(1 + \alpha/2)^{-1/\beta} - (\alpha/2)^{-1/\beta}]$, and the relationship between d_α and α is involved for the method to be directly invoked.

Finally, besides the caveat of $D(d_\alpha)$ not having a minimum, the other caveat is the dependence of an optimal coverage probability on data. Specifically, the use of a posterior distribution of X to obtain $R(d_\alpha)$ makes this latter quantity depend on the observed data with the consequence that in the same problem one could conceivably end up using a different coverage probability from forecast to forecast. Unattractive as this may sound, it is the price that one must pay to ensure coherence. However, this dependence on the data becomes of less concern once the posterior distribution of X converges, so that the effect of the new data on the posterior diminishes. The same situation will also arise when the distribution of X is specified via a frequentist approach involving a plug-in rule.

6. SUMMARY AND CONCLUSIONS

The thesis of this article is to argue that choosing coverage probabilities for prediction intervals should be based on decision theoretic considerations. The current practice is to choose these by convention or astute judgment. Prediction intervals are one of the essentials of regression, time series, and state space models. They also occur in conjunction with previsions based on probability models entailing the judgment of exchangeability. Furthermore, the principles underlying the construction of prediction intervals share some commonality with those involving inventory planning and banking reserves.

The decision theoretic approach boils down to the minimization of total expected disutility. This disutility consists of two components. One is a disutility associated with the width of the interval and the other is associated with the failure of an interval to cover the observed value when it reveals itself. The proposed approach is illustrated via a consideration of stylized utility functions. It can be seen as a prototype for approaches based on other utility functions. The approach also entails a use of the normal distribution to describe the uncertainties. Again, this distributional assumption is not essential; other distributions will work equally well.

We emphasize that the material here pertains to prediction intervals, not confidence intervals. It would be interesting to develop a decision theoretic approach for choosing the confidence coefficient of a confidence interval. To the best of our knowledge, this remains to be satisfactorily done.

[Received June 2007. Revised December 2007.]

REFERENCES

Beran, R. (1990), "Calibrating Prediction Regions," *Journal of the American Statistical Association*, 85, 715–23.

Box, G. E. P., and Jenkins, G. M. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco, CA: Holden-Day.

Chatfield, C. (1993), "Calculating Interval Forecasts," *Journal of Business and Economic Statistics*, 11, 121–135.

Decamps, J. P., Rochet, J. C., and Roger, B. (2004), "The Three Pillars of Basel II: Optimizing the Mix," *Journal of Financial Intermediation*, 13, 132–155.

de Finetti, B. (1972), *Probability, Induction and Statistics*, New York: Wiley.

De Groot, M. H. (1970), *Optimal Statistical Decisions*, New York: McGraw-Hill.

Gardner, Jr., E. S. (1988), "A Simple Method of Computing Prediction Intervals for Time Series Forecasts," *Management Science*, 34, 541–546.

Granger, C. W. J. (1996), "Can We Improve the Perceived Quality of Economic Forecasts?" *Journal of Applied Econometrics*, 11, 455–473.

Hadley, G., and Whitin, T. M. (1963), *Analysis of Inventory Systems*, Englewood Cliffs, NJ: Prentice-Hall.

Hahn, G. J., and Meeker, W. Q. (1991), *Statistical Intervals: A Guide for Practitioners*, New York: Wiley.

Lawless, J. F., and Fredette, M. (2005), "Frequentist Prediction Intervals and Predictive Distributions," *Biometrika*, 92, 529–542.

Lehmann, E. L. (1958), "Significance Level and Power," *The Annals of Mathematical Statistics*, 29, 1167–1176.

Lindley, D. V. (1985), *Making Decisions* (2nd ed.), New York: Wiley.

Pearson, K. (1920), "The Fundamental Problem of Practical Statistics," *Biometrika*, 13, 1–16.

Savage, L. J. (1962), "Bayesian Statistics," in *Recent Developments in Information and Decision Processes*, eds. R. E. Machol and P. Gray, New York: The Macmillan Company, pp. 161–194.

Singpurwalla, N. D. (2006), *Reliability and Risk: A Bayesian Perspective*, England: Wiley.

Tay, A. S., and Wallis, K. F. (2000), "Density Forecasting: A Survey," *Journal of Forecasting*, 19, 235–254.

Winkler, R. L. (1972), "A Decision-Theoretic Approach to Interval Estimation," *Journal of the American Statistical Association*, 67, 187–191.

(5)

Encyclopedia of Statistics in Quality and Reliability

Damage Processes

Standard Article

Nozer D. Singpurwalla[1]
[1]George Washington University, Washington, DC, USA

DOI: 10.1002/9780470061572.eqr073
Article Online Posting Date: March 15, 2008

Abstract | Full Text: HTML PDF (79K)

Abstract

The point of view that we adopt here is that damage is an abstract notion that conveys an intuitive import. Specifically, an item fails when its damage exceeds a threshold. Whereas damage cannot be observed and measured, the surrogates that it spawns such as crack growth, CD4 cell counts, and wear, can be. These surrogates are known as *markers*. With the above in mind, we offer here a probabilistic architecture that enc to make the notion of damage precise and still retain its intuitive import. Our architecture looks at damage a: cumulative hazard and describes its evolution *via* a nondecreasing stochastic process. The observable mar associated with damage are also modeled by a stochastic process that is cross-correlated with the damage process. Thus a bivariate stochastic process with one component that is nondecreasing and the other that r fluctuate around some mean could be a suitable model for encapsulating the phenomenon of damage and i markers. The latent nondecreasing process leads to the fluctuating observable processes, and an item fails the former hits a threshold. The second feature of our architecture pertains to the threshold. We argue that t threshold needs to be random and has an exponential distribution with scale one; we call such a threshold t *hazard potential* of an item. To conclude, our perspective on what constitutes a damage process is starkly d from that which is prevalent in the reliability and the survival analysis literatures. Hopefully, it offers a platfor describing an ill-defined but much-discussed phenomenon.

Keywords: aging; crack growth; degradation; deterioration; fatigue; gamma processes; hazard potential; he status; latent variables; marker processes; quality of life; reliability; stochastic integral; stochastic processes surrogates; survival analysis; wear; Wiener process

Damage Processes

Introduction and Background

There is extensive and burgeoning material on the topic of damage and its associated factors like **aging**, cumulative damage, **degradation**, deterioration, fatigue, health status, and quality of life. This material appears in both the biostatistical and the engineering reliability literatures. However, these notions suffer from the feature that they lack a precise definition. Rather they convey an abstract but intuitive import in the sense of a decrease in residual (or remaining) life. This decrease in residual life is conceptualized *via* the feature that an item experiencing aging and degradation will fail when the damage hits some barrier or threshold. Alternatively, it is supposed that at inception, every item is endowed with a resource that gets depleted because of damage, and that the item fails when the resource gets exhausted. Thus, for example, to engineers like Bogdanoff and Kozin [1], "Degradation is the irreversible accumulation of damage throughout life that leads to failure." The term damage is not made precise; however it is claimed that damage reveals itself *via surrogates* or *markers*, such as cracks that grow in size, corrosion, measured wear (i.e., depletion of material), and so on. Similarly, Sobczyk [2] sees fatigue as "a phenomenon which takes place in units experiencing time-varying external actions which manifest in a deterioration of the unit's resistance to carry its intended loading". In the biostatistical literature, *aging* pertains to a unit's position in a state space wherein the probabilities of failure are greater than its former position. Aging manifests itself in terms of biomedical and physical difficulties experienced by individuals, and in certain scenarios, *via* things like low-CD4 cell counts; these serve as biomedical surrogates, or what are known as *biomarkers*.

The markers mentioned above are, in most cases, observable and measurable entities. Much of the recent work on what is known as *degradation modeling* centers around assessing lifetimes *via* an analysis of the observed markers and their hitting times to a threshold (*cf.* Doksum [3], Doksum and Normand [4], Lu and Meeker [5], Ebrahimi [6], and Lehmann [7]). However, treating the observable markers as substitutes for the unobservable degradation process that actually causes failure is tantamount to putting the cart before the horse. This is because the unobservable degradation process spawns the observable marker process, and is therefore its cause. An exception, however, is the work of Whitmore *et al.* [8] and of Lee *et al.* [9], who treat the degradation and the marker as separate but related processes. Also, see Nair [10], who makes the point that data on the observable surrogates of degradation help *sharpen* lifetime assessments. In this vein, a noteworthy contribution is by Cox [11] who systematically articulates the roles that the observable and the unobservable play in lifetime assessments. The premise upon which our bivariate stochastic process model with a random threshold is based has been inspired by the papers of Whitmore *et al.* [8] and Cox [11], and our work on hazard potential (see Singpurwalla [12; 13, p. 79]).

Preliminaries: The Hazard Potential

For an appreciation of the bivariate stochastic process model as a description of the damage process, some preliminaries on the notion of a hazard potential would be helpful. Accordingly, let T denote the lifetime of a unit and let $h(t)$ be the hazard rate of $P(T \geq t), t \geq 0$; let $H(t) = \int_0^t h(u)\, du$ be the cumulative hazard function at t. Then it is easy to see that

$$P(T \geq t;\ h(t), t \geq 0) = \exp(-H(t)) = P(X \geq H(t)) \qquad (1)$$

where X has an exponential distribution with scale one. The random variable X is called the *hazard potential* of the item, and it represents an unknown "resource" that the item is endowed with at inception. Furthermore $H(t)$ is a measure of the amount of resource consumed by time t, and $h(t)$, the rate at which the resource is consumed at t. The unit fails when $H(t)$ exceeds X; that is when $H(t)$ hits the random threshold X.

When the rate at which a unit's resource gets consumed is random, $h(t)$ is described by a stochastic process, making $\{H(t);\ t \geq 0\}$ a stochastic process as well. However, this latter process has to be nondecreasing. The unit fails when the process $\{H(t);\ t \geq 0\}$ hits a barrier X, where X is also random with a unit exponential distribution. Candidate stochastic processes for $\{H(t);\ t \geq 0\}$ are

also alluded to in Singpurwalla [12]. Since $H(t)$ is, in principle, nondecreasing in t, $t \geq 0$, the process $\{H(t);\ t \geq 0\}$ is a candidate for describing a damage process. Furthermore, since the conventional view claims that an item fails when the damage hits a threshold, the cumulative hazard and the damage reflect a parallel feature. This motivates us to view the (cumulative) damage as being isomorphic with the cumulative hazard. Doing so makes our perspective different from that which is currently being discussed in literature. Like (cumulative) damage, the cumulative hazard is not observable. However, the cumulative hazard does influence the time to failure. Consequently, the cumulative hazard and the (cumulative) damage are to be seen as *latent variables*, and for that matter, so is the hazard potential X.

A Stochastic Process Model for Damage and its Markers

Because markers are closely linked with damage, any suitable model for the damage process should be accompanied by some sort of description for the markers as well. The most general way to do this would be to assume that the markers are realizations of stochastic processes, just as the (cumulative) damage is a stochastic process. The simplest way to proceed would be to suppose that there is only one marker to focus attention upon, so that a bivariate stochastic process $\{H(t), Z(t);\ t \geq 0\}$ would be a suitable description of the damage and its marker. As stated in the section titled "Preliminaries: The Hazard Potential", the process $\{H(t);\ t \geq 0\}$ is nondecreasing in t. However, the process $\{Z(t);\ t \geq 0\}$ need not be restricted to being nondecreasing. Indeed, markers such as crack growth and CD4 cell counts fluctuate around some trend, and thus one is free to choose any suitable model for the process $\{Z(t);\ t \geq 0\}$. A Wiener process appears to be the model of choice, but this need not be so.

Thus to summarize, our proposed model for the (cumulative) damage and its associated marker is a bivariate stochastic process $\{H(t), Z(t);\ t \geq 0\}$ with $H(t)$ nondecreasing in t, and $Z(t)$ free to fluctuate around some constant or trend. We term such a process a *degradation process*. Since $H(t)$ spawns $Z(t)$, the two processes $\{H(t);\ t \geq 0\}$ and $\{Z(t);\ t \geq 0\}$ need to be linked; that is, they need to be *cross-correlated*. Without such linkage, the marker process cannot serve as predictor of failure, and the statistical exercise of degradation modeling is not meaningful. One way to achieve this linkage is to describe $\{Z(t);\ t \geq 0\}$ by a Wiener process, and the unobservable (cumulative) damage process by a *Wiener maximum process*, namely,

$$H(t) = \sup_{0<s\leq t} \{Z(s);\ s \geq 0\} \qquad (2)$$

This strategy has been proposed in Singpurwalla [14], wherein a Bayesian approach for inference about lifetimes, using data on the marker process, is also described. The item fails when $H(t)$ hits the (random) threshold X. Whereas the model of equation (2) could be a starting point, there is a *caveat* that needs to be addressed. Specifically, since $Z(t)$ is spawned by $H(t)$, the latter is the cause of the former. This means that $H(t)$ must lead $Z(t)$, and so any linkage between the two processes in question should incorporate a time lag. The model of equation (2) does not do this because here $H(t)$ is determined retrospective to $Z(t)$ and therefore lags $Z(t)$, instead of the other way around. Thus $H(t)$ and $Z(t)$ need to be connected, with the observable $Z(t)$ lagging the unobservable $H(t)$. This is a possible topic for future research.

In the section titled "Candidate Processes for Damage and Markers", we give an overview of some modeling strategies that have been proposed for the damage process $\{H(t);\ t \geq 0\}$, as well as for the marker process $\{Z(t);\ t \geq 0\}$ when each are treated separately; that is when no distinction is made between the damage (or degradation) process and the marker process. Supplementary material on the above can also be found in Chapters 7 and 8 of Singpurwalla [13].

Candidate Processes for Damage and Markers

The origins of the work on threshold crossing of cumulative damage as a basis for failure goes back to Epstein [15], Esary [16], and Gaver [17]. The idea of describing cumulative damage as a stochastic process can be traced, to the best of our knowledge, to Cox [18, p. 91], and to the Ph.D. thesis of Morey [19]. However, the granddaddy of all work on damage processes is the remarkable paper of Esary

et al. [20], who (without articulating what damage means) describe damage by a compound **Poisson process** with increments that are positive and have the Markov property. Failure occurs when the said process hits a random barrier whose distribution is exponential. The choice of an exponential distribution for the barrier is arbitrary, and Esary *et al.* [20] show that the hitting time has an exponential distribution. More recently Zacks [21, 22] has elaborated on this theme.

In the biostatistical arena there is a setup parallel to that of Esary *et al.* [20], which does not allude to damage, deterioration, or aging, but to the number of mice at some time t, that have typhoid organisms. The growth of such mice is described by a pure birth process, and the first passage time to a barrier is investigated. The specifics are in Cox and Miller [23, p. 160], and in Cox [11].

The Esary *et al.* [20] architecture is enhanced by Lemoine and Wenocur [24], who model wear (i.e., damage) by a suitable random process but who also allow for failure due to trauma. The latter is described by a Poisson process, the rate of which depends on the state of wear. An item fails when the wear reaches a threshold or when the item experiences fatal trauma. Thus in the model of Lemoine and Wenocur [24], the wear and the trauma processes compete with each other for an item's lifetime. The random process considered by the above authors is a diffusion process that is driven by a Wiener process. In a subsequent paper, Lemoine and Wenocur [25] describe wear by a shot-noise process. A disadvantage of the diffusion and the shot-noise process is that the wear (to us damage) is not monotonically nondecreasing. To rectify this deficiency, Wenocur [26] considers a gamma process for describing wear. His development of the gamma process proceeds along the following lines.

Partition the time interval into subintervals of length h, and let $X(n)$ denote the damage (or wear) at time nh, $n = 1, 2, \ldots$. Suppose that the damage at time $(n+1)h$ is prescribed *via* the relationship

$$X(n+1) - X(n) = \alpha(X(n))\varepsilon_n + \beta(X(n))h \quad (3)$$

where α, β are constants, and $\{\varepsilon_n\}$ is a sequence of independent and identically distributed random variables having a gamma distribution with shape parameter $h > 0$. Letting $h \downarrow 0$, we have

$$dX(t) = \alpha(X(t^-))\, d\gamma(t) + \beta(X(t^-)) \quad (4)$$

where $\{\gamma(t)\}$ is a *gamma process*.

In integral terms, equation (4) becomes the *stochastic integral*

$$X(t) = X(0) + \int_0^t \alpha(X(s^-))\, d\gamma(s) + \int_0^t \beta(X(s^-))\, ds \quad (5)$$

Since the gamma process has nonnegative increments, the wear (or damage) process is increasing. For an overview of the gamma process and their constructions, see Singpurwalla [27], or van der Weid [28]. Whereas a gamma process model may be attractive in scenarios wherein the damage causing shocks occur frequently, the models by Zacks [21, 22] for the compound Poisson process case and for the compound renewal process case, respectively, seem to be more appropriate when the shocks are infrequent.

Candidate Marker Processes

In engineering reliability, an archetypical marker process is crack growth, whereas in biostatistical studies, it appears that CD4 cell counts is a commonly mentioned biomarker. With archetypical markers come archetypical stochastic processes for $\{Z(t);\ t \geq 0\}$, and one such process is the Wiener process with a drift parameter η and a diffusion parameter $\sigma^2 > 0$; see Doksum [3] and Whitmore *et al.* [8]. As mentioned before, the marker is often viewed as a proxy for damage, and failure is said to occur when the marker process hits a threshold. As is well known, the hitting time to the threshold (assumed fixed and known) of a Wiener process has an *inverse Gaussian distribution* (see Singpurwalla [13, p. 68 and 136], for a discussion of this distribution).

The Wiener process has independent increments, so does a gamma process. This amounts to saying that the increments of crack size are independent of the existing crack length. This latter phenomenon is not always true. The bigger the crack, the bigger is its growth. This motivates one to consider transformations of the Wiener process. Furthermore, the crack growth phenomenon also exhibits abrupt growth. The Wiener process does not encapsulate such abruptness of growth. With the above in mind, Schabe [29] proposes the following as a model for $X(t)$, the size of a crack at time t, $t \geq 0$. Let $X(t) = (M(t))^a$, where

$$M(t) = bt + W(t) + \mu P(t) \quad (6)$$

Here $W(t)$ is a Wiener process with variance $\sigma^2 t$, and $P(t)$ is a Poisson process with intensity

λ. The constant $b \geq 0$ describes a trend, and the constant a is such that $a > (<)1$ encapsulates a progressive (regressive) velocity with which the crack grows. Under the model of equation (6), Schabe [29] obtains the hitting time of $X(t)$ to $h > 0$, a barrier. This distribution does not have a closed-form solution. However, the mean and the variance of this distribution are available; these are $h^{h/a}(b + \mu\lambda)$ and $h^{1/a}\left(\sigma^2 + \lambda\mu^2\right)/(b + \mu\lambda)^3$, respectively.

In Ebrahimi [6], a strategy that parallels those of Lemoine and Wenocur [24] and of Schabe [29] is taken, but instead of looking at the growth of a single crack, an ensemble of k cracks, each having its own growth rate is considered. Specifically, it is supposed that the growth of the ith crack, $i = 1, \ldots, k$ is governed by the stochastic differential equation

$$\mathrm{d}X_i(t) = \lambda_i(t)X_i(t) + \sigma X_i(t)\ \mathrm{d}W(t) \qquad (7)$$

where $\lambda_i(t)$ is the growth rate of the cracks, $\sigma > 0$ is a constant, and $\{W(t);\ t \geq 0\}$ is a standard Wiener process with mean 0 and variance t. Using standard results (i.e., the Ito formula) it can be seen that

$$X_i(t) = X(0)\exp\left[\Lambda_i(t) - \frac{\sigma^2}{2}t + \sigma W(t)\right] \qquad (8)$$

where $\Lambda_i(t) = \int_0^t \lambda_i(s)\ \mathrm{d}s$, and $X(0)$ is the initial crack size, assumed to be known, and is the same for all the k cracks. The item fails when the size of the largest crack hits some threshold, say a. If T denotes the passage time of the largest crack to the threshold a, then it can be seen that for $0 \leq u \leq t$

$$P(T \geq t) = P\left[W(u) < \left(\min_{1 \leq i \leq k}\left(\frac{\sigma}{2}u - \frac{1}{\sigma}\Lambda_i(u)\right) + c\right)\right] \qquad (9)$$

where $c = b/\sigma$ and $b = \log a - \log X(0)$.

Computation of the above follows from results on the the times taken by the Weiner process to hit a threshold.

Motivated by a model of Durham and Padgett [30], Park and Padgett [31] propose a model for cumulative damage, assuming that the damage is an observable measurable entity. This amounts to interpreting cumulative damage as a marker, like crack growth. The scheme proposed by Park and Padgett [31] is noteworthy, because it facilitates the introduction of both a **Brownian motion** process and a gamma process as the driving processes for crack growth. Here, for some functions $c(\cdot)$ and $h(\cdot)$, it is assumed that

$$\mathrm{d}c(X(t)) = h(X(t))\ \mathrm{d}D(t) \qquad (10)$$

where $D(t)$ is the damage at t, and $X(t)$ is the cumulative damage at t. As a consequence of the above

$$\int_0^t \frac{1}{h(X(u))}\ \mathrm{d}c(X(u)) = \int_0^t \mathrm{d}D(u) = D(t) - D(0) \qquad (11)$$

By choosing various forms for the function $c(\cdot)$ and $h(\cdot)$, and a stochastic process for $\{D(t);\ t \geq 0\}$, different models for $X(t)$ can be attained. For example, with $h(u) = 1, c(u) = \log u$, and a Brownian motion (or Wiener process) for $\{D(t);\ t \geq 0\}$, we obtain a geometric Brownian motion process for $X(t)$. With $h(u) = 1, c(u) = u$ and a Brownian motion process for $\{D(t);\ t \geq 0\}$ we obtain a Gaussian process for $X(t)$. With $h(u) = 1, c(u) = u$ and a gamma process for $\{D(t);\ t \geq 0\}$ we obtain a gamma process for $X(t)$. Whereas the Gaussian process is not always positive, the geometric Brownian motion process is always positive but not increasing. In contrast, the gamma process is both positive and increasing. Characteristics of the hitting times of the geometric Brownian motion and the gamma process to a fixed and known threshold are also obtained by Park and Padgett [31]. This completes our overview of stochastic process models for the damage process and the marker process – when viewed separately – save for the work of Desmond [32], who articulates on a two-parameter family of life distributions introduced by Birnbaum and Saunders [33]. This distribution is motivated *via* the notion that failure caused by fatigue is due to the initiation, growth, and extension of a dominant crack past some critical length.

The essence of Desmond's [32] idea is based on the notion that it is the environmental stresses called *impulses* that cause a crack to grow, so that if X_i is the size of the crack after the ith impulse, then

$$X_{i+1} = X_i + \Pi_{i+1}g(X_i), i = 0, 1, 2, \ldots \qquad (12)$$

here Π_i is taken to be the magnitude of the ith impulse; e.g. the stress caused by the ith impulse. The Π_i's are assumed to be random. If $\nabla X_i = X_{i+1} - X_i$

is taken to be sufficiently small, then

$$\sum_{1}^{n} \Pi_i \approx \int_{X_0}^{X_n} \frac{\mathrm{d}y}{g(y)} \tag{13}$$

is approximately normal; $g(y)$ is some function of y. The quantity X_0 is the initial size of the dominant crack.

With $g(y) = 1$, and assuming that the Π_i's have a common mean μ and variance σ^2

$$I(X(t) \stackrel{\text{def}}{=} \int_{X_0}^{X_n} \frac{\mathrm{d}y}{g(y)} \sim N(t\mu, t\sigma^2) \tag{14}$$

If X_c denotes the critical crack size, and T the time to failure of the unit experiencing the impulses, then

$$T = \inf\{t: X(t) > X_c\} \tag{15}$$

Simple manipulations show that

$$P(T \leq t) = \Phi\left(\frac{t\mu - I(X_c)}{\sigma\sqrt{t}}\right) \tag{16}$$

Where $\Phi(\cdot)$ is the unit **normal distribution** function. The distribution function given above is a member of the Birnbaum–Saunders [33] family of distributions.

Acknowledgment

Supported by the US Army Research Office Grant W911NF-05-01-2009 and the Office of Naval Research Contract N00014-06-1-037 with the George Washington University.

References

[1] Bogdanoff, J.L. & Kozin, F. (1985). *Probabilistic Models of Cumulative Damage*, John Wiley & Sons, New York.

[2] Sobczyk, K. (1987). Stochastic models for fatigue damage of materials, *Advances in Applied Probability* **19**, 652–673.

[3] Doksum, K.A. (1991). Degradation models for failure time and survival data, *CWI Quarterly, Amsterdam* **4**, 195–203.

[4] Doksum, K.A. & Normand, S.L.T. (1995). Gaussian models for degradation processes-part I: methods for the analysis of biomarker data, *Lifetime Data Analysis* **1**(2), 131–144.

[5] Lu, C.J. & Meeker, W.Q. (1993). Using degradation measures to estimate a time-to-failure distribution, *Technometrics* **35**(2), 161–174.

[6] Ebrahimi, N. (2004). System reliability based on diffusion models for fatigue crack growth, *Naval Research Logistics* **52**(1), 46–57.

[7] Lehmann, A. (2006). Degradation-threshold-shock models, in *Probability, Statistics and Modeling in Public Health*, M. Nikulin, D. Commenges & C. Huber, eds, Springer, pp. 286–298.

[8] Whitmore, G.A., Crowder, M.J. & Lawless, J.F. (1998). Failure inference from a marker process based on a bivariate Wiener process, *Lifetime Data Analysis* **4**, 229–251.

[9] Lee, M.L.T., Grutolla, V.D. & Schoenfeld, D. (2000). A model for markers and latent health status, *Journal of the Royal Statistical Society. Series B* **62**(4), 747–762.

[10] Nair, V.N. (1998). Discussion of estimation of reliability in field performance studies by J.D. Kalbfleisch and J.F. Lawless, *Technometrics* **30**, 379–383.

[11] Cox, D.R. (1999). Some remarks on failure-times, surrogate markers, degradation, wear, and the quality of life, *Lifetime Data Analysis* **5**, 307–314.

[12] Singpurwalla, N.D. (2006). The hazard potential: introduction and overview, *Journal of the American Statistical Association* **101**(476), 1705–1717.

[13] Singpurwalla, N.D. (2006). *Reliability and Risk: A Bayesian Perspective*, John Wiley & Sons.

[14] Singpurwalla, N.D. (2006). On competing risk and degradation processes, in *The Second Erich L. Lehman Symposium – Optimality, Institute of Mathematical Statistics – Monograph Series, Vol. 49*, J. Rojo, ed, The Institute of Mathematical Statistics, pp. 289–304.

[15] Epstein, B. (1958). The exponential distribution and its role in life testing, *Industrial Quality Control* **15**, 2–7.

[16] Esary, J.D. (1957). A stochastic theory of accident survival and fatality, Ph.D. dissertation, University of California, Berkeley.

[17] Gaver Jr, D.P. (1963). Random hazard in reliability problem, *Technometrics* **5**, 211–226.

[18] Cox, D.R. (1962). *Renewal Theory*, Methuen, London.

[19] Morey, R.C. (1965). Stochastic wear processes, Technical Report ORC 65-16, Operations Research Center, University of California, Berkeley.

[20] Esary, J.D., Marshall, A.W. & Proschan, F. (1973). Shock models and wear processes, *Annals of Probability* **1**, 627–649.

[21] Zacks, S. (2004). Distributions of failure times associated with non-homogeneous compound Poisson damage processes, *A Festschrift for Herman Rubin, Institute of Mathematical Statistics – Lecture Notes, Vol. 45*, pp. 396–407.

[22] Zacks, S. (2006). Failure distributions associated with general compound renewal damage processes, in *Probability, Statistics and Modeling in Public Health*, M. Nikulin, D. Commenges & C. Huber, eds, Springer, pp. 466–474.

[23] Cox, D.R. & Miller, H.D. (1965). *The Theory of Stochastic Processes*, Chapman and Hall, London.

[24] Lemoine, A.J. & Wenocur, M.L. (1985). On failure modeling, *Naval Research Logistics Quarterly* **32**, 497–508.

[25] Lemoine, A.J. & Wenocur, M.L. (1986). A note on shot-noise and reliability modeling, *Operations Research* **34**, 320–323.

[26] Wenocur, M.L. (1989). A reliability model based on gamma process and its analytic theory, *Advances in Applied Probability* **21**, 899–918.

[27] Singpurwalla, N.D. (1997). Gamma processes and their generalizations: an overview, in *Engineering Probabilistic Design and Maintenance for Flood Protection*, R. Cook, M. Mendel & H. Vrijling, eds, Kluwer Academic Publishers, pp. 67–73.

[28] van der Weid, H. (1997). Gamma processes, in *Engineering Probabilistic Design and Maintenance for Flood Protection*, R. Cook, M. Mendel & H. Vrijling, eds, Kluwer Academic Publishers, pp. 77–83.

[29] Schabe, H. (1990). A new stochastic model for crack propagation, *Quality and Reliability Engineering International* **6**, 341–344.

[30] Durham, S.D. & Padgett, W.J. (1997). A cumulative damage model for system failure with application to carbon fibers and composites, *Technometrics* **39**, 34–44.

[31] Park, C. & Padgett, W.J. (2005). Accelerated degradation models for failure based on geometric Brownian motion and gamma processes, *Lifetime Data Analysis* **11**, 511–527.

[32] Desmond, A. (1985). Stochastic models of failure in random environments, *The Canadian Journal of Statistics* **13**(2), 171–183.

[33] Birnbaum, Z.W. & Saunders, S.C. (1969). A new family of life distributions, *Journal of Applied Probability* **6**, 319–327.

Related Articles

Cumulative Damage Models Based on Gamma Processes; Degradation and Failure; Degradation Processes.

NOZER D. SINGPURWALLA

(6)

The Hazard Potential: Introduction and Overview

Nozer D. SINGPURWALLA

This is an expository article directed at reliability theorists, survival analysts, and others interested in looking at life history and event data. Here we introduce the notion of a hazard potential as an unknown resource that an item is endowed with at inception. The item fails when this resource becomes depleted. The cumulative hazard is a proxy for the amount of resource consumed, and the hazard function is a proxy for the rate at which this resource is consumed. With this conceptualization of the failure process, we are able to characterize accelerated, decelerated, and normal tests and are also able to provide a perspective on the cause of interdependent lifetimes. Specifically, we show that dependent life lengths are the result of dependent hazard potentials. Consequently, we are able to generate new families of multivariate life distributions using dependent hazard potentials as a seed. For an item that operates in a dynamic environment, we argue that its lifetime is the killing time of a continuously increasing stochastic process by a random barrier, and this barrier is the item's hazard potential. The killing time perspective enables us to see competing risks from a process standpoint and to propose a framework for the joint modeling of degradation or cumulative damage and its markers. The notion of the hazard potential generalizes to the multivariate case. This generalization enables us to replace a collection of dependent random variables by a collection of independent exponentially distributed random variables, each having a different time scale.

KEY WORDS: Competing-risk process; Degradation process; Dependence; Exchangeable lifetimes; Killing times; Lévy process; Marker; Multivariate failure models; Random killing; Reliability; Survival analysis.

1. INTRODUCTION AND OVERVIEW

1.1 Preliminaries: The Hazard Rate and the Hazard Potential

The mathematical theory of reliability, the statistical theory of life history or survival analysis, and the underlying premise of actuarial sciences are driven by a notion unique to them: the *hazard rate function* (see, e.g., Gjessing, Aalen, and Hjort 2003). The hazard rate function is both a theoretical and a descriptive tool that also plays a fundamental role in event history analysis. Specifically, there is a parallel between the hazard rate function and the *intensity function* of a nonhomogeneous Poisson process (see Grandell 1975), and also between the intensity function of a doubly stochastic Poisson process and the hazard rate function when the latter is viewed as a stochastic process (see Kebir 1991). There are two virtues of the hazard function: (a) an interpretive content, in the sense that the aging characteristics of single and one-of-a-kind items can be encapsulated by the shape of the hazard function, and (b) that under some regularity conditions (see Yashin and Arjas 1988; Singpurwalla and Wilson 1995), the hazard function uniquely determines a survival function. There are other scenarios in which (a) is also germane; these have been alluded to by Gjessing et al. (2003); some examples are an understanding of neuronal degeneration, the sleep–wake cycles of individuals, and the longevity of humans (see Gavrilov and Gavrilova 2001).

This is an expository article directed at reliability theorists, survival analysts, actuaries, and others interested in event history analysis. Our purpose here is to introduce a new notion, the *hazard potential* (HP) as a conceptual tool that provides a different way of looking at the stochastic behavior of lifetimes. The term "potential" refers to a feature parallel to that of *potential energy* in physics. The difference here is that we are alluding to an item's resistance to failure rather than its capacity for work. In Section 3 of this article we put forth the view that the HP can be interpreted as the (random) amount of an unknown resource with which an item is endowed at inception, and that the item fails when this resource is depleted. Looking at lifetime in terms of a depleting resource can be more satisfying than one based on conditional probabilities, which is what the hazard function represents.

Besides providing an alternative platform for conceptualizing the process that leads to failure, and for processes that compete for failure, the HP has the following attractive features:

- It is inherently robust, in the sense that the HP of any and all items has an exponential (1) distribution on a suitably chosen time scale.
- It provides a context-free means for characterizing accelerated, decelerated, normal, and partially accelerated life tests.
- In the language of probabilistic causality (see Suppes 1970), it can be seen as either a prima facie or a genuine cause of dependence between lifetimes.
- It provides a vehicle for developing new families of univariate and multivariate survival functions by looking at the killing times of continuously increasing stochastic processes to random barriers.
- It offers a natural platform from which the abstract phenomenon of degradation (or damage accumulation) and its markers can be stochastically described.

The HP generalizes to the multivariate case. This generalization, when used in conjunction with the notion of a *hazard gradient* due to Marshall (1975a), enables us to represent a collection of dependent lifetimes in terms of a collection of independent exponential (1) random variables, each on a different time scale.

In light of the foregoing features, we may liken the HP to the notion of a *hidden parameter* in physics. Hidden parameters per se do not have a physical reality, but nonetheless are valuable because they provide explanations for observable phenomena.

1.2 Overview

This article is organized as follows. In Section 2 we introduce our notation and review some basic relationships. In Section 3

Nozer D. Singpurwalla is Professor of Statistics and Decision Sciences, Department of Statistics, George Washington University, Washington, DC 20052 (E-mail: *nozer@gwu.edu*). The author thanks Hakon Gjessing for his help regarding the distribution of killing times of an integrated geometric Brownian motion process. The detailed comments by two referees are gratefully acknowledged. This research was supported in part by U.S. Army Research Office grant W911NF-05-1-0209 and Office of Naval Research grant N00014-06-1-0037.

Journal of the American Statistical Association
December 2006, Vol. 101, No. 476, Review Article
DOI 10.1198/016214506000001068

we define the HP and interpret its nature from both physical and probabilistic standpoints. We also provide a way to formally distinguish between accelerated, decelerated, normal, and partially accelerated life tests from a context-free standpoint. The state of the art in accelerated testing seems vague when it comes to being specific about what a normal life test means; it treats this matter as a given. We conclude Section 3 by generalizing the HP to the nonexponential case through the notion of a *G-hazard potential*. In Section 4 we present several qualitative results pertaining to the claim that dependent HPs are a prima facie cause of dependent lifetimes, whereas a common HP is a genuine cause of dependence. Dependent HPs are a manifestation of commonalities in manufacturing or, in the context of biological units, a shared genetic makeup. In Section 5 we put the material of Section 4 to work by generating new families of dependent lifetimes using dependent HPs as a seed. In Section 6 we develop new families of survival functions for items destined to operate in random environments. The material here revolves around obtaining the distribution of the killing time of a continuously increasing stochastic process by a random barrier that is an item's HP. Although the approach of Section 6 is general, attention focuses only on the following processes: the running maxima of a Brownian motion, a Markov process with nonnegative increments, a family of nonnegative Lévy processes, and the integral of a geometric Brownian motion. The material of Sections 5 and 6 is not purely conceptual; it has the attractiveness of having a practical import. This can be seen as an argument in favor of looking at the HP as a useful tool. In Section 7 we explore the role of the HP in articulating the notion of competing-risk processes and casting the phenomenon of degradation and its markers in a manner that accords with that described in the engineering and materials science literature. We devote Section 8 to the multivariate case, which entails a relationship between the hazard gradient and what we introduce as a *conditional* HP. This connection allows us to replace a collection of dependent lifetimes by a collection of independent exponential (1) lifetimes, each indexed by a different time scale. In Section 9 we close the article by reemphasizing the point of view that the HP offers an alternative perspective for appreciating the failure process and that it is a useful conceptual tool for understanding the cause of interdependent lifetimes in engineering and biological systems. We close Section 9 by expressing our hope that the role of the HP could turn out to be as useful to reliability and survival analysis as the failure rate and the intensity functions.

2. NOTATION, TERMINOLOGY, AND PRELIMINARY RELATIONSHIPS

Let T denote the (unknown) time to failure of a unit that is scheduled to operate in some environment, labeled $\mathcal{E}$. Based on the characteristics of the unit, and on knowledge of how the unit interacts with $\mathcal{E}$ (vis-à-vis T), one is able to subjectively specify $h(t)$, $t \geq 0$, the *hazard rate function* of $P(T > t)$, the *survival function* of T, assumed to be absolutely continuous. We interpret $h(t)$ through the relationship

$$h(t)\,dt \approx P(t \leq T \leq t + dt | T \geq t),$$

where the right side is a conditional probability. A formal definition of $h(t)$ can be found in the recent book of Aven and Jensen (1999). We claim that the hazard function is a theoretical (or abstract) notion because, unlike lifetimes that can be directly observed, conditional probabilities are either subjectively specified or inferred from data.

Let $H(t) = \int_0^t h(u)\,du$; $H(t)$ is known as the *cumulative hazard* at time t. Observe that $H(t)$ is nondecreasing in t. But what does $H(t)$ mean? Whereas $h(t)\,dt$ can be given an intuitive import, $H(t)$ cannot! It is not the sum of conditional probabilities—because the conditioning event changes with t—and there is no law of probability that leads us to $H(t)$. Thus $H(t)$ does not have a probabilistic connotation. Yet $H(t)$ plays a key role in reliability and survival analysis, because of the *exponentiation formula* (see Barlow and Proschan 1975, p. 53), which says that with $H(t)$ specified,

$$P(T \geq t; H(t), t \geq 0) = e^{-H(t)}. \tag{1}$$

In the foregoing equation, plus those that follow, we introduce the convention that all quantities to the right of the semicolon are viewed as being specified. In contrast, all quantities to the right of the vertical slash are conditional, that is, if they are known.

Equation (1) relates the survival function $P(T \geq t)$ to $H(t)$; however, $H(t)$ lacks an interpretive content. Our interest in this article is motivated by the desire to interpret $H(t)$ in a manner that provides insight into the relationship of (1).

In the case of a one-of-a-kind item, $h(t)\,dt$ encapsulates an assessor's judgment about the inherent quality of an item and the environment in which it operates. By quality, we mean a resistance to failure-causing agents, such as crack growth, weakening of the immune system, and so on. Consequently, the hazard rate of an item of poor quality that operates in a benign environment could be smaller than that of a high-quality item that operates in a harsh environment. In effect, the quantity $h(t)\,dt$ encapsulates an assessor's subjective view of the manner in which an item and its environment interact. Thus, in principle, $h(t)\,dt$ does not have a physical reality.

Turning attention to the right side of (1), we note that $e^{-H(t)}$ is the survival function of an exponentially distributed random variable, say X, if its scale parameter is 1, evaluated at $H(t)$, that is,

$$P(T \geq t; H(t), t \geq 0) = e^{-H(t)} = P(X \geq H(t) | 1). \tag{2}$$

3. INTERPRETATION: THE NOTION OF A HAZARD POTENTIAL

Thus far, we have introduced three quantities, X, H, and T. Given any two of these, we can find the third using (2). But what insight can (2) provide about $H(t)$ and X? We see two possibilities, one providing an indifference principle for reliability and survival analysis and the other having a physical connotation.

To appreciate the first, we see from (2) that, corresponding to every nonnegative random variable T having an absolutely continuous survival function $\overline{F}(t) = P(T \geq t)$, there exists a random variable X taking values $H(t)$, $0 \leq H(t) < \infty$, whose survival function is an exponential with a scale parameter of 1. The survival function of T is indexed by t, $t \geq 0$, whereas that of X is indexed by $H(t) = -\int_0^t d\overline{F}(u)/\overline{F}(u)$. We can summarize the foregoing in the following theorem.

Theorem 1. The lifetime of any and all items has an exponential (1) distribution on $H(t)$, the cumulative hazard, as the scale.

The essence of Theorem 1 has been noted by Cinlar and Ozekici (1987); it is stated here as a prelude to Theorem 5, which pertains to the multivariable case. In the context of point processes, Theorem 1 has a parallel with the result that any nonhomogeneous Poisson process can be transformed by a change in clock time to a homogeneous Poisson process with rate one (see Kingman 1964). This parallel leads us to make precise the notions of accelerated and normal life tests in Section 3.1.

3.1 The Physical Connotation

To appreciate the physical connotation implied by (2), we note that because

$$P(T \le t; H(t), t \ge 0) = P(X \le H(t) | 1),$$

we may claim that the time to failure T of an item coincides with the time at which the cumulative hazard $H(t)$ crosses a random threshold X, where X has an exponential (1) distribution (Fig. 1), that is, $T = H^{-1}(X)$.

The random threshold X, where $X = H(T)$, is defined as the HP of the item. Furthermore, because the exponential (1) distribution of X does not depend on $\mathcal{E}$, we may interpret X as an unknown resource with which the item is endowed at the time of its inception. With X considered a resource, $H(t)$ can be interpreted as the amount of resource consumed by time t. Consequently, the hazard rate, $h(t) = \frac{d}{dt}H(t)$, can be considered the rate at which the resource is consumed. With this alternative perspective on $H(t)$ and $h(t)$, we may view a *normal life-test* as one for which $H(t) = t$, a *uniformly accelerated (decelerated)* test as one for which $H(t) > (<)\ t$, and a *partially accelerated (decelerated)* test as one for which $H(t)$ crosses t from above (below). The qualifier accelerated (decelerated) signals a contraction (expansion) of the clock time from t to $H(t)$, and by shifting attention from the applied stress (which is what is normally done when discussing accelerated tests) to time, we achieve the context-free feature mentioned earlier. The concept of looking at failure as the depletion of a resource dates back to a Soviet physicist Sedyakin (1966), who enunciated this viewpoint without a formal architecture.

It is useful to note that the exponential (1) random variable X has an entropy of 1, and also the lack of memory property if and only if $H(t) = t$. A change in clock time from t to $H(t)$ changes the entropy and destroys the memoryless property.

3.2 The *G*-Hazard Potential

There is a generalization of Theorem 1 such that the HP can be made to have a distribution other than an exponential (1). Specifically, suppose that G is some absolutely continuous distribution function with support $[0, \infty)$; let $W = G^{-1}$. Then it can be seen (Bagdonavicius and Nikulin 1999) that $Y \stackrel{\text{def}}{=} W(\overline{F}(T))$ has the survival function G, irrespective of $\mathcal{E}$. Consequently;

$$P(T \ge t) = P\big(W(\overline{F}(T)) \ge W(\overline{F}(t))\big) = P\big(Y \ge W\big(e^{-H(t)}\big)\big),$$

so that

$$P(T \le t) = P\big(Y \le W\big(e^{-H(t)}\big)\big). \tag{3}$$

Equation (3) implies that the item fails when $W(e^{-H(t)})$, exceeds a threshold Y, where Y has the distribution G. We refer to Y as the *G-hazard potential* and $W(e^{-H(t)})$ as the *G-resource used until time t*. Then we have, as a generalization of Theorem 1, the following result.

Theorem 2. The lifetime of any item can be made to have any absolutely continuous survival function G, provided that G is indexed by $G^{-1}(\exp(-H(t)))$.

As of now, Theorem 2 is mainly of an academic interest; it is given here for completeness.

4. HAZARD POTENTIALS AND DEPENDENT LIFETIMES

The aim of this section is to discuss the nature of dependence between lifetimes and offer a new perspective on the

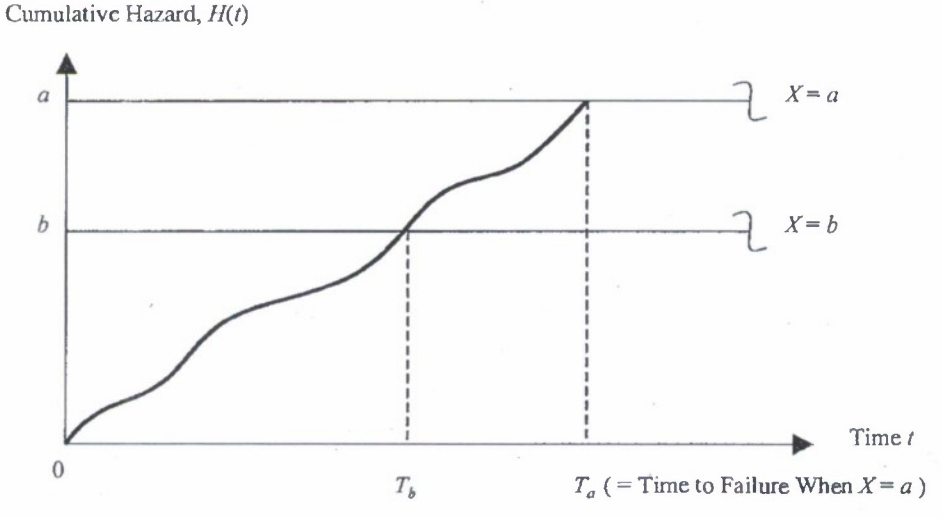

Figure 1. Relationship Between Cumulative Hazard, Threshold X, and Failure Time.

cause of interdependence. We argue that the HP offers a convenient platform for doing this. We view dependence and independence from a subjectivistic (de Finettian) viewpoint; that is, two events A and B are dependent if knowledge about B causes us to change our two-sided bets on A.

Because $H(t)$ encapsulates an assessor's view about the interaction between an item's quality and its environment, it is likely that two different items operating in a common environment will have different $H(t)$'s, say $H_1(t)$ and $H_2(t)$. Similarly, for a single item, changing its environment from $\mathcal{E}_1$ to $\mathcal{E}_2$ will change its cumulative hazard from $H_1(t)$ to $H_2(t)$ (Fig. 2).

Figure 2 suggests that the lifetimes T_1 and T_2 of two items having the same hazard potential will be dependent. Equivalently, the lifetimes T_1^* and T_2^* of a single item scheduled to operate in two environments, $\mathcal{E}_1$ and $\mathcal{E}_2$, will also be dependent. However, from a subjectivistic perspective, the dependence will come into play only when one is able to specify $H_1(t)$ and $H_2(t)$, or a relationship between the two, when only one of them is known. This is because knowledge of, say, T_1 together with $H_1(t)$ will tell us something about the unknown X_1, and if X_1 and X_2 are dependent, then knowledge of X_1 will enlighten us about X_2. Consequently, X_2 together with $H_2(t)$ will help change our assessment of T_2. To summarize, if the HPs X_1 and X_2 are dependent, then the lifetimes T_1 and T_2 will also be dependent, provided that $H_1(t)$ and $H_2(t)$ are known or a relationship between them is specified. In contrast, if X_1 and X_2 are independent, then so are T_1 and T_2, irrespective of whether or not $H_1(t)$ and $H_2(t)$ are known. These assertions are summarized in the remarks that follow.

Remark 1. When $H_1(t)$ and $H_2(t)$, $t \geq 0$, are known, lifetimes T_1 and T_2 are independent if and only if their hazard potentials, X_1 and X_2, are independent.

Proof. When X_1 and X_2 are independent,

$$P\big(X_1 \geq H_1(t_1), X_2 \geq H_2(t_2)\big) = P(X_1 \geq H_1(t_1)) \cdot P(X_2 \geq H_2(t_2)),$$

for any $H_1(t_1)$ and $H_2(t_2)$. Consequently,

$$\begin{aligned} &P\big(T_1 \geq t_1, T_2 \geq t_2; H_1(t), H_2(t), t \geq 0\big) \\ &\quad = P\big(X_1 \geq H_1(t_1), X_2 \geq H_2(t_2)\big) \\ &\quad = P(X_1 \geq H_1(t_1)) \cdot P(X_2 \geq H_2(t_2)) \\ &\quad = P(T_1 \geq t_1; H_1(t), t \geq 0) \cdot P(T_2 \geq t_2; H_2(t), t \geq 0). \end{aligned}$$

Thus, knowing $H_1(t)$ and $H_2(t)$, T_1 and T_2 are independent, and similarly for the converse.

When $H_i(t)$, $i = 1, 2$ or both $i = 1$ and 2, for $t \geq 0$ are not known, Remark 1 is weakened in the sense that only the "if" part holds. Specifically, T_1 and T_2 are independent even when X_1 and X_2 are dependent. The subjectivistic line of reasoning justifying this claim goes as follows.

Observing T_1 provides no insight about X_1, because $H_1(t)$ is not known. Consequently, there also is no insight into X_2 or T_2. Thus T_1 and T_2 are independent. Mathematically, without knowing $H_i(t)$, $i = 1, 2$, we are unable to relate $P(T_1 \geq t_1, T_2 \geq t_2)$ with the distribution of X_1 and X_2. We summarize the foregoing in the following remark.

Remark 2. Lifetimes T_1 and T_2 are independent whenever $H_1(t)$ and (or) $H_2(t)$, $t \geq 0$, are not known.

As a consequence of Remarks 1 and 2, we may state the following theorem.

Theorem 3. Lifetimes T_1 and T_2 are dependent if and only if their hazard potentials X_1 and X_2 are dependent and if $H_1(t)$ and $H_2(t)$ are known.

Theorem 3 puts aside the often expressed view that the lifetimes of items sharing a common environment are necessarily dependent (see Marshall 1975b; Lindley and Singpurwalla 1986); that is, it is a common environment that causes dependence among lifetimes. Theorem 3 asserts that it is the commonalities in the HPs or identical HPs, both of which result in dependent HPs, that cause of interdependent lifetimes. Dependent HPs are a manifestation of similarities in design, manufacture, or genetic makeup. In the language of probabilistic

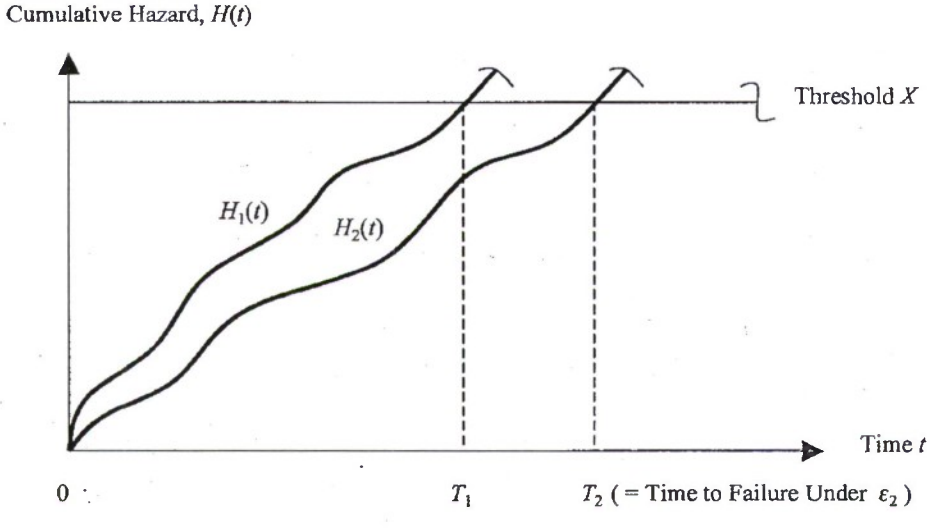

Figure 2. Effect of Changing Environment on Lifetimes.

causality of Suppes (1970), the common environment can be interpreted as a spurious cause of dependent lifetimes, whereas dependent (or identical) HPs are their prima facie (or genuine) cause.

The role of Theorem 3 is to generate new families of dependent lifetimes using multivariate distributions with exponential marginals as a seed; see Section 5. Remarks 1 and 2 pertain to the two extreme cases in which the $H_i(t)$'s, $i = 1, 2$, are either known or not. An intermediate case is one in which an $H_i(t)$, say $H_1(t)$, $t \geq 0$, is known and the other is not, except for the fact that $H_1(t) > H_2(t)$. For such scenarios, we have the following.

Remark 3. Suppose that $H_1(t) > (\leq)\ H_2(t)$ and that either $H_1(t)$ or $H_2(t)$, $t \geq 0$, is known; then X_1 and X_2 dependent implies that T_1 and T_2 are also dependent.

Proof. The proof is by contradiction. For this, suppose that X_1 and X_2 have the Bivariate Exponential Distribution (BVE) of Marshall and Olkin (1967); specifically, for λ_1, λ_2, and $\lambda_{12} > 0$,

$$P(X_1 \geq x, X_2 \geq y) = \exp\big(-\lambda_1 x - \lambda_2 y - \lambda_{12}\max(x, y)\big),$$
$$= \exp(-(\lambda_1 + \lambda_{12})x - \lambda_2 y), \quad \text{if } x > y.$$

The marginal distribution of X_i, $P(X_i \geq x) = \exp(-(\lambda_i + \lambda_{12})x)$, $i = 1, 2$. For the X_i's to be dependent HPs, we need to have $(\lambda_i + \lambda_{12}) = 1$, for $i = 1, 2$, and $\lambda_{12} > 0$; this would imply that $\lambda_1 = \lambda_2 = \lambda$. Thus

$$P(X_1 \geq x, X_2 \geq y) = \exp(-(x + \lambda_2 y)).$$

If we set $x = H_1(t_1)$ and $y = H_2(t_2)$, for some $t_1, t_2 \geq 0$, then $x > y$ would imply that $H_2(t_2) = H_1(t_2) - \delta$, for some unknown $\delta > 0$. Consequently,

$$P(X_1 \geq x, X_2 \geq y) = P\big(X_1 \geq H_1(t_1), X_2 \geq H_1(t_2) - \delta\big)$$
$$= \exp\big(-\big(H_1(t_1) + \lambda_2(H_1(t_2) - \delta)\big)\big). \quad (4)$$

Given the foregoing, we need to show that T_1 and T_2 are dependent. Suppose that they are not; then

$$P\big(T_1 \geq t_1, T_2 \geq t_2; H_1(t_1), H_2(t_2), t_1, t_2 \geq 0\big)$$
$$= P(T_1 \geq t_1; H_1(t_1), t_1 \geq 0)P(T_2 \geq t_2; H_2(t_2), t_2 \geq 0)$$
$$= P(X_1 \geq H_1(t_1))P(X_2 \geq H_2(t_2))$$
$$= \exp(-H_1(t_1))\exp(-(H_1(t_2) - \delta))$$
$$= P\big(X_1 \geq H_1(t_1), X_2 \geq H_1(t_2) - \delta\big), \quad (5)$$

because the first term of (5) does not entail elements of the second term. Thus we have

$$P\big(X_1 \geq H_1(t_1), X_2 \geq H_1(t_2) - \delta\big)$$
$$= \exp\big(-(H_1(t_1) + H_1(t_2) - \delta)\big). \quad (6)$$

Equation (6) agrees with (4) if $\lambda_2 = 1$. However, $\lambda_2 = 1$ implies that $\lambda_{12} = 0$, which contradicts the hypothesis that X_1 and X_2 are dependent. The proof when $H_1(t) \leq H_2(t)$ follows along similar lines.

A broader, but weaker version of Remark 1 pertains to the case where X_1 and X_2 are *exchangeable*. Here again, we require that $H_i(t)$, $i = 1, 2$, $t \geq 0$, be specified. We then have the following result.

Remark 4. If the hazard potentials X_1 and X_2 are exchangeable and if $H_1(t), H_2(t), t \geq 0$, are known, then the lifetimes T_1 and T_2 are also exchangeable.

Proof. Let $x = H_1(t)$ and $y = H_2(t)$ for any $t_1, t_2 \geq 0$; then

$$P(X_1 \geq x, X_2 \geq y) = P\big(T_1 \geq t_1, T_2 \geq t_2; H_1(t_1), H_2(t_2)\big).$$

Similarly,

$$P(X_1 \geq y, X_2 \geq x) = P\big(T_1 \geq t_2, T_2 \geq t_1; H_1(t_1), H_2(t_2)\big).$$

Because the exchangeability of X_1 and X_2 implies that

$$P(X_1 \geq x, X_2 \geq y) = P(X_1 \geq y, X_2 \geq x),$$

the statement of the remark now follows.

5. GENERATING NEW FAMILIES OF DEPENDENT LIFETIMES

The aim of this section is to put Theorem 3 to work. Here we show how dependent HPs can be used to generate new families of multivariate distributions through multivariate distributions with unit exponentials as a seed. Of course, this is by no means the only way to generate multivariate distributions. For the purpose of illustration, we limit attention to the bivariate case and consider as seeds the bivariate exponentials of Marshall and Olkin (1967), Gumbel (1960), and Singpurwalla and Youngren (1993; henceforth S–Y), and a bivariate exponential induced by the copula of a bivariate Pareto distribution.

5.1 The Bivariate Exponential of Marshall and Olkin

Suppose that the HPs X_1 and X_2 have the BVE of Marshall and Olkin (1967), with λ_1, λ_2, and λ_{12} as parameters. To ensure that the marginal distributions are unit exponentials, we need to have $\lambda_1 = \lambda_2 = \lambda$ and $\lambda + \lambda_{12} = 1$, with $\lambda_{12} > 0$; the latter inequality ensures dependence between X_1 and X_2.

Let T_1 and T_2 be the lifetimes corresponding to X_1 and X_2 and the cumulative hazard functions $H_1(t_1)$ and $H_2(t_2)$. Then, because

$$P(T_1 \geq t_1, T_2 \geq t_2; \cdot)$$
$$= P\big(X_1 \geq H_1(t_1), X_2 \geq H_2(t_2); \lambda, \lambda_{12}\big)$$
$$= \exp\big[-\lambda(H_1(t_1) + H_2(t_2)) - \lambda_{12}\max(H_1(t_1), H_2(t_2))\big],$$

we can generate families of bivariate distributions for T_1 and T_2, by assuming specific forms for $H_i(t)$, for $i = 1, 2$. In particular, if $H_i(t_i) = (\alpha_i t_i)^{\beta_i}$, $i = 1, 2$, then

$$P(T_1 \geq t_1, T_2 \geq t_2; \cdot) = \exp\big[-\big\{\lambda[(\alpha_1 t_1)^{\beta_1} + (\alpha_2 t_2)^{\beta_2}]$$
$$+ \lambda_{12}\max[(\alpha_1 t_1)^{\beta_1}, (\alpha_2 t_2)^{\beta_2}]\big\}\big],$$

which is a bivariate Weibull of the Marshall–Olkin type.

If $H_i(t_i) = \alpha_i \ln(1 + \beta_i t_i)$, $i = 1, 2$, then

$$P(T_1 \geq t_1, T_2 \geq t_2; \cdot)$$
$$= \left(\frac{1}{1 + \beta_1 t_1}\right)^{\lambda\alpha_1}\left(\frac{1}{1 + \beta_2 t_2}\right)^{\lambda\alpha_2}$$
$$\times \min\left[\left(\frac{1}{1 + \beta_1 t_1}\right)^{\lambda_{12}\alpha_1}, \left(\frac{1}{1 + \beta_2 t_2}\right)^{\lambda_{12}\alpha_2}\right],$$

which resembles the bivariate distribution of Muliere and Scarsini (see Kotz, Balakrishnan, and Johnson 2000, henceforth KBJ, pp. 408 and 595). This distribution is also known as the Marshall–Olkin–type Pareto distribution (see KBJ 2000, p. 612). Note that $H_i(t_i) = (\alpha_i t_i)^{\beta_i}$ $[\alpha_i \ln(1+\beta_i t_i)]$ corresponds to an inereasing (decreasing) rate of consumption of the HP.

Continuing in the foregoing vein, if $H_i(t_i) = \alpha_i(e^{\beta_i t_i} - 1)$, $i = 1, 2$, then the induced distribution of T_1 and T_2 is given as

$$P(T_1 \geq t_1, T_2 \geq t_2; \cdot) = e^{(\alpha_1+\alpha_2)\lambda} \cdot \exp\Big[-\big\{\alpha_1\lambda e^{\beta_1 t_1} + \alpha_2\lambda e^{\beta_2 t_2} + \lambda_{12}\max\big(\alpha_1(e^{\beta_1 t_1} - 1), \alpha_2(e^{\beta_2 t_2} - 1)\big)\big\}\Big],$$

and if $H_i(t_i) = (1 - e^{-t_i})/(1 + e^{-t_i})$, $i = 1, 2$, the logistic function, then

$$P(T_1 \geq t_1, T_2 \geq t_2; \cdot) = \exp\left[-\left\{\lambda\left(\frac{1-e^{-t_1}}{1+e^{-t_1}} + \frac{1-e^{-t_2}}{1+e^{-t_2}}\right) + \lambda_{12}\max\left(\frac{1-e^{-t_1}}{1+e^{-t_1}}, \frac{1-e^{-t_2}}{1+e^{-t_2}}\right)\right\}\right].$$

Neither of these distributions is of a recognized form. The first form of $H_i(t_i)$ corresponds to an exponential rate of consumption of the HP, whereas the second corresponds to a rate of that which starts at $\frac{1}{2}$ at $t = 0$ and asymptotes to 1 as t becomes infinite.

5.2 The Bivariate Exponential of Gumbel

Following the notation of Scction 5.1, suppose that for some parameter $0 \leq \theta \leq 1$,

$$P\big(X_1 \geq H_1(t_1), X_2 \geq H_2(t_2); \theta\big) = \exp[-H_1(t_1) - H_2(t_2) - \theta H_1(t_1)H_2(t_2)].$$

This is the bivariate exponential of Gumbel (1960), with marginals that are always unit exponentials. If $H_i(t_i) = (\alpha_i t_i)^{\beta_i}$, $i = 1, 2$, then the induced distribution of T_1 and T_2 is

$$P(T_1 \geq t_1, T_2 \geq t_2; \cdot) = \exp\Big[-\big\{(\alpha_1 t_1)^{\beta_1} + (\alpha_2 t_2)^{\beta_2} + \theta(\alpha_1 t_1)^{\beta_1}(\alpha_2 t_2)^{\beta_2}\big\}\Big];$$

we call this distribution the bivariate Weibull of the Gumbel type.

If $H_i(t_i) = \alpha_i \ln(1 + \beta_i t_i)$, $i = 1, 2$, then

$$P(T_1 \geq t_1, T_2 \geq t_2; \cdot) = \left(\frac{1}{1+\beta_1 t_1}\right)^{\alpha_1}\left(\frac{1}{1+\beta_2 t_2}\right)^{\alpha_2} \times \exp\big(-\theta\alpha_1\alpha_2 \ln(1+\beta_1 t_1)\ln(1+\beta_2 t_2)\big),$$

which is a multivariate distribution with marginals that are a Pareto; we call this distribution a bivariate Pareto of the Gumbel type.

5.3 The Bivariate Exponential of S–Y

Here again, we follow the notation of Section 5.1 and suppose that for some parameter m,

$$P\big(X_1 \geq H_1(t_1), X_2 \geq H_2(t_2); m\big) = \sqrt{\frac{1 - m\cdot\min(H_1(t_1), H_2(t_2)) + m\cdot\max(H_1(t_1), H_2(t_2))}{1 + m(H_1(t_1) + H_2(t_2))}} \times \sqrt{e^{-m\cdot\max(H_1(t_1), H_2(t_2))}}.$$

This distribution has unit exponential marginals if $m = 2$.

If we set $H_1(t_1) \geq H_2(t_2)$, then

$$P\big(X_1 \geq H_1(t_1), X_2 \geq H_2(t_2)\big) = \sqrt{\frac{1 - 2H_2(t_2) + 2H_1(t_1)}{1 + 2(H_1(t_1) + H_2(t_2))}\exp(-2H_1(t_1))}.$$

The multivariate distributions for T_1 and T_2, when derived assuming that the $H_i(t_i)$ take any of the forms given in Section 5.1, are not of any recognizable type; they appear to be new. This is not surprising, because the bivariate exponential given earlier is also not of a well-recognized form.

5.4 Unit Exponentials Induced by Copulas

New families of multivariate distributions with unit exponentials can be created by the method of copulas and by invoking Sklar's theorem in reverse (see, e.g., Nelson 1995). We can then use these multivariate exponentials as a seed for generating other families of multivariate distributions.

As an example of the foregoing, consider a bivariate Pareto distribution of the form

$$P(X_1 \geq x_1, X_2 \geq x_2; \cdot) = \left(\frac{b}{b + x_1 + x_2}\right)^{a+1};$$

its copula, for $u \geq 0$ and $v \leq 1$, is

$$C_a(u, v) = u + v - 1 + \big((1-u)^{-(a+1)} + (1-v)^{-(a+1)} - 1\big)^{-(a+1)}.$$

If we set $u = 1 - \exp(-H_1(t_1))$ and $v = 1 - \exp(-H_2(t_2))$, then it can be seen (see Singpurwalla and Kong 2004) that

$$P\big(X_1 \geq H_1(t_1), X_2 \geq H_2(t_2); a\big) = \left(\exp\left(\frac{H_1(t_1)}{a+1}\right) + \exp\left(\frac{H_2(t_2)}{a+1}\right) - 1\right)^{a+1},$$

which is a bivariate distribution with unit exponentials as marginals. We may now choose any desired form for the $H_i(t_i)$, $i = 1, 2$, to produce new families of bivariate distributions of the form $P(T_1 \geq t_1, T_2 \geq t_2; \cdot)$.

6. CUMULATIVE HAZARD PROCESSES AND RANDOM KILLING

Our discussion thus far has been based on the premise that $H(t)$ is a deterministic function of t. This may be a reasonable first step. A more meaningful strategy is to assume that $H(t)$ is described by some nondecreasing and nonnegative stochastic process $\{H(t); t \geq 0\}$. There is some precedence for doing so in both the biostatistical and the reliability literature (see Singpurwalla 1995), although the motivation there is different

from what we give here. This is because we see $H(t)$ as a proxy for usage until time t, and conceptualizing usage as a random process is more natural than simply declaring that the cumulative hazard is a stochastic process. With $H(t)$ described as a stochastic process, the time to failure T will be the hitting time of $\{H(t); t \geq 0\}$ to a random barrier X, which is the HP of the item; see Figure 1. Put alternatively, the lifetime of an item corresponds to the killing time of $\{H(t); t \geq 0\}$ by a random threshold X. The notion that lifetimes correspond to hitting times of stochastic processes to some barrier was also explored in the pioneering work of Esary, Marshall, and Proschan (1973; henceforth EMP) and in the more recent works of Durham and Padgett (1997), Pettit and Young (1999), Yang and Klutke (2000), and Duchesne and Rosenthal (2003), the difference being that to these authors, the underlying stochastic process is an observable phenomenon such as degradation, aging, or cumulative damage. A consequence of the foregoing is that the results thus obtained pertain to specific scenarios. In contrast, the approach of considering any failure time as the hitting time of a process $\{H(t); t \geq 0\}$ to a random threshold X whose distribution is an exponential (1) provides a common architecture for developing classes of survival functions, with each class determined by the nature of the process. For example, when $\{H(t); t \geq 0\}$ is a positive nondecreasing Lévy process (special cases of which are the compound Poisson, the gamma, and the stable), a general result for the survival function is obtained. We discuss this and related matters in what follows.

6.1 The Hazard Rate and Cumulative Hazard Process

The purpose of this section is to obtain a result analogous to that of (2) when $H(t)$ is a stochastic process. To obtain an analog to the left side of (2), we proceed formally by considering a probability measure space $(\Omega, \mathcal{F}, P)$ on which all random variables and processes are defined.

Let $\{h(s); s \geq 0\}$ be a nonnegative and right-continuous stochastic process, and let T be a real-valued random variable denoting the lifetime of an item. For $t \geq 0$, we define the σ-algebras $\mathcal{F}_t$ and $\mathcal{F}$ as

$$\mathcal{F}_t = \sigma\{h(s); s \leq t\} \qquad \text{and} \qquad \mathcal{F} = \sigma\{h(s); s \geq 0\}.$$

Then $\{h(s); s \geq 0\}$ is defined as the *hazard rate process* of T, if, for $t > 0$,

$$P(T > t|\mathcal{F}) = \exp\left(-\int_0^t h(s)\,ds\right).$$

It now follows, from a result of Pitman and Speed (1973), that T is a *randomized stopping time*, so that

$$P(T > t|\mathcal{F}_t) = \exp\left(-\int_0^t h(s)\,ds\right), \qquad t \geq 0.$$

Consequently,

$$P(T > t) = E\left[\exp\left(-\int_0^t h(s)\,ds\right)\right]$$

or

$$P(T > t) = E\big[\exp(-H(t))\big], \tag{7}$$

where $\{H(t); t \geq 0\}$ is the cumulative hazard process. Equation (7) is our analog of the left side of (2).

For an analog of the right side of (2), we assume that $\{H(t); t \geq 0\}$ is a nonnegative, nondecreasing stochastic process and consider the hitting time of this process to a random threshold X whose distribution is an exponential (1). Then, assuming independence of $H(t)$ and X,

$$P(T > t) = P(X > H(t)) = \int_0^\infty \exp(-y) H_t(dy) = E\big[\exp(-H(t))\big], \tag{8}$$

where $H_t(\cdot)$ is the density of the distribution of $H(t)$. Thus an analog to the right side of (2), with $\{H(t); t \geq 0\}$ a stochastic process, is

$$P(T > t) = E\big[\exp(-H(t))\big]. \tag{9}$$

The right side of (9) is the Laplace transform of the process $\{H(t); t \geq 0\}$, which for the Lévy process has an explicit form, namely

$$E\big[\exp(-H(t))\big] = \exp\left[-t\int_0^\infty [1 - \exp(-y)]\nu(dy)\right], \tag{10}$$

where $\nu(dy)$ is the Lévy measure of the process and the integral term is the Laplace exponent of the Lévy process; complete the Lévy–Khinchin formula of Protter (1990). An attractive feature of the argument that leads to (8) is the straightforward manner in which it is developed. In contrast, the argument of (7) calls for some appreciation of randomized stopping rules associated with stochastic processes.

In what follows we consider several possible candidates for the process $\{H(t); t \geq 0\}$, starting with the simplest and moving to the more general. In most cases, explicit expressions for $P(T > t)$ are obtained; in others, computations and approximations may be needed.

The choice of which of the following processes to use depends on the application. Presumably, because $H(t)$ encapsulates the resource used until time t, the selection of a suitable process for $\{H(t); t \geq 0\}$ would depend on the pattern of use of the item.

6.2 Cumulative Hazard Processes and Their Survival Functions

The process $\{H(t); t \geq 0\}$ is required to be nonnegative, nondecreasing, and right-continuous. Thus our choice of candidate processes is limited. Clearly, the Brownian motion process, which has often been used to describe degradation and wear, must be eliminated. However, certain functionals of the Brownian motion, such as the running maxima, are viable candidates, and this is the first process considered.

6.2.1 The Maxima of Brownian Motion. Suppose that $\{W(t); t \geq 0\}$ is a standard Brownian motion process [i.e., $W(0) = 0$]; for any $t > 0$, $W(t)$ has a Gaussian distribution with mean 0 and variance t, and $\{W(t); t \geq 0\}$ has stationary independent increments. If we set

$$H(t) = \sup_{0<s\leq t} \{W(s)\}, \qquad t \geq 0,$$

then the process $\{H(t); t \geq 0\}$ will be continuous, nonnegative, and nondecreasing; this is called a *Brownian maximum process*.

It is well known that $T_x \stackrel{\text{def}}{=} \inf\{t \geq 0; W(t) \geq x\} = \inf\{t \geq 0; H(t) \geq x\}$, the time at which the process $\{W(t); t \geq 0\}$ first hits a barrier x, $x > 0$, has an inverse-Gaussian distribution (see Pettit and Young 1999). Consequently, the hitting time of the process $\{H(t); t \geq 0\}$ to x also has an inverse-Gaussian distribution, specifically,

$$P(T_x \leq t) = 2(1 - \Phi(x/\sqrt{t})),$$

where $\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-s^2/2}\, ds$.

Because the time to failure of an item is the time at which the process $\{H(t); t \geq 0\}$ first hits the (HP) X, where X is exponential (1),

$$P(T \leq t) = \int_0^\infty 2(1 - \Phi(x/\sqrt{t}))e^{-x}\, dx = 1 - 2e^{t/2}\Phi(-\sqrt{t}),$$

so that

$$P(T > t) = 2e^{t/2}\Phi(-\sqrt{t}), \tag{11}$$

an expression that is easily evaluated.

6.2.2 The Compound Poisson Process. The compound Poisson process with an (arrival) rate λ and iid jumps J_i, $i = 1, 2, \ldots$, with $P(J_i \leq \omega) = G(\omega)$ is another possible candidate for describing the process $\{H(t); t \geq 0\}$. This process increases only by jumps of size J_i, $i = 1, 2, \ldots$. If we assume that the J_i's are also independent of the HP X, then, given λ,

$$P(T > t|\lambda) = \sum_{k=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^k}{k!} \int_0^\infty G^{(k)}(x)e^{-x}\, dx,$$

where $G^{(k)}(\cdot)$ is the k-fold convolution of $G(\cdot)$ with itself. The foregoing simplifies (see EMP 1973) as

$$P(T > t|\lambda) = \exp(-\lambda t), \tag{12}$$

for all $G(\cdot)$ with $G(x) = 0$, $x \leq 0$.

When $\lambda > 0$ is unknown, we may average out $P(T > t|\lambda)$ with respect to any distribution of λ. This would lead us to conclude that $P(T > t)$ has a hazard rate function $h(t)$ that is a decreasing function of $t \geq 0$. Thus items experiencing use of a resource described by a compound Poisson process will necessarily have lifetimes with a heavy-tailed distribution function. For example, if λ has a gamma distribution, then $P(T > t)$ will have a Pareto distribution, that is heavy-tailed.

6.2.3 A Special Markov Process. EMP (1973) considered a Markov process for $\{H(t); t \geq 0\}$ with a special feature that describes proneness to wear. Whereas their interpretation of $H(t)$ is unlike ours, their special feature is appropriate to our setup, specifically (a) $H(0) = 0$; (b) $H(t + \Delta) - H(t) \geq 0$, $\forall t, \Delta \geq 0$; and (c) $P(H(t + \Delta) - H(t) \leq u|H(t) = z) \downarrow z$, t. The practical import of (c) is that proneness to wear increases with usage. With the foregoing in place, EMP (1973) showed that for any barrier x, the hitting time of the process $\{H(t); t \geq 0\}$ has a distribution with a failure rate function $h(u)$ such that $\widetilde{h}(t)$ increases in t, where

$$\widetilde{h}(t) = \frac{1}{t}\int_0^t h(u)\, du. \tag{13}$$

Such distributions are said to have an increasing hazard rate average property. Because the barrier in our case is the HP X, where X has an exponential (1) distribution, we note that for the special Markov process for $\{H(t); t \geq 0\}$, the survival function $P(T > t)$ can be written as an exponential (1) mixture of distributions with the increasing hazard rate average property.

6.2.4 A Nonnegative Lévy Process. An omnibus way of describing $\{H(t); t \geq 0\}$ is through a nonnegative Lévy process, that is, a continuous process with stationary independent increments. Such processes are examples of Markov processes and include the compound Poisson, the gamma, and all stable processes as special cases. Furthermore, a Lévy process renews itself at stopping times and has a strong Markov property, and all of the nonnegative Lévy processes are limits of compound Poisson processes (see Protter 1990). Thus the process provides a convenient general platform for describing $\{H(t); t \geq 0\}$ and makes the result of (12) based on the compound Poisson process central. Besides the foregoing generalities, the main attraction of considering a Lévy process stems from the fact that its Laplace transform (given by the Lévy–Khinchin formula) takes a form identical to that of (10), namely

$$P(T > t) = \exp\left[-t\int_0^\infty (1 - \exp(-y))\nu(dy)\right], \tag{14}$$

where $\nu(dy)$, the Lévy measure, characterizes both the expected frequency and the size of the jumps (nonnegative in our case) in a Lévy process.

For the compound Poisson process of Section 6.2.2, $\nu(dy) = \lambda G(dy)$, and if G had a gamma distribution with scale $\alpha > 0$ and shape $\beta > 0$, then

$$\nu(dy) = \lambda\alpha^\beta y^{\beta-1}e^{-\alpha y}\, dy/\Gamma(\beta).$$

In the case of a gamma process [i.e., when for any $t \geq 0$, $H(t)$ has a gamma distribution with scale $\alpha > 0$, and shape βt],

$$\nu(dy) = (\beta e^{-\alpha y}/y)\, dy, \tag{15}$$

whereas when $\{H(t); t \geq 0\}$ is described by a stable process,

$$\nu(dy) = \frac{\alpha\beta}{\Gamma(1-\beta)} y^{-(1+\beta)}\, dy \tag{16}$$

for parameters $\alpha > 0$ and $\beta \in (0, 1)$. Plugging (15) and (16) into (14) will give $P(T > t)$ for the special cases of the gamma and the stable process; also see (18) in the next section.

6.2.5 Continuous and Increasing Strong Markov Processes. One of the more striking results in stochastic processes theory pertains to continuous and increasing processes that have the strong Markov property. It has been shown that such processes have deterministic paths up to random killing. Essentially, this means that a continuous increasing strong Markov process is essentially deterministic. This result dates back to work of Blumenthal, Getoor, and McKean (1962). Loosely speaking, if $\{H(t); t \geq 0\}$ is an increasing, continuous, strong Markov process with a state space of form $[a, b)$, then there exists a strictly increasing continuous function $k(\cdot)$ on the state space such that for all $t \geq 0$, $H(t) = k^{-1}[k(H(0)) + t]$; for specifics, see corollary 1 of Cinlar (1979). Thus the sample path of the $\{H(t); t \geq 0\}$ process is a deterministic function of the initial state of the process, namely $H(0) = 0$, and time t. Once the process $\{H(t); t \geq 0\}$ is considered (essentially) deterministic, obtaining the hitting time of $H(t)$ to a barrier is relatively

straightforward; it is also deterministic if the barrier is a known constant. Randomness of hitting times enters into the picture when the barrier is random, which is so in our case.

As an illustration of the foregoing, suppose that the process $\{H(t); t \geq 0\}$ is an increasing Lévy process. Recall that Lévy processes are continuous, have stationary independent increments, and thus are strong Markov. When this is the case, the function $k(\cdot)$ is such that for some $a \geq 0$,

$$H(t) = at + \int_0^\infty (1 - \exp(-ut))\nu(du), \tag{17}$$

where $\nu(du)$ is the Lévy measure of the process.

If we set $a = 0$ and assume that $\{H(t); t \geq 0\}$ is a gamma process, then $\nu(du)$ is given by (15), and the deterministic cumulative hazard function turns out to be

$$H(t) = \beta \log\left(\frac{\alpha + t}{\alpha}\right), \qquad t \geq 0$$

(see Kebir 1991 for more details).

The unit fails when $H(t)$ gets killed by a threshold x; that is, T_x, the time to failure for a fixed threshold x, is

$$T_x = \alpha\left(e^{x/\beta} - 1\right).$$

Averaging with respect to its exponential (1) distribution, we have

$$P(T \geq t) = \left(1 + \frac{t}{\alpha}\right)^{-\beta}, \tag{18}$$

which is a Pareto distribution. Note that the Pareto distribution also arises in the context of a compound Poisson process for $\{H(t); t \geq 0\}$ when the distribution of λ, the arrival rate, is assumed to be a gamma; see the discussion after (12).

To summarize, in practically all of the cases that we have considered so far, closed-form expressions for $P(T > t)$ are available. The sole exception is the special Markov process of Section 6.2.3, for which our result is merely qualitative. Our final case, considered next, pertains to an exponential functional of Brownian motion; here a closed-form result is not available. We chose this case because of its novelty and plausible applicability.

6.2.6 Integrated Geometric Brownian Motion Process. In Section 6.2.1 we considered the running maximum of a standard Brownian motion as a model for $\{H(t); t \geq 0\}$. Here we consider another functional. Specifically, let

$$H(t) = \int_0^t \exp(2W(s))\, ds, \tag{19}$$

where $W(s)$ is a standard Brownian motion. We choose the scalar 2 for convenience; its role will become clear in the sequel. Observe that $\exp(2W(s))$ is always positive and that $H(t)$ is continuous and strictly increasing in t. Recall that a Brownian motion has continuous sample paths. Whereas $\sup_{0<s\leq t}\{W(s); s \geq 0\}$ increases in t by steps, the $H(t)$ of (19) is a strictly increasing function of t. As stated earlier, Brownian motion has often been used to describe crack growth and degradation. The foregoing transformation of the process is necessitated by the requirement that $H(t)$ be nonnegative and nondecreasing. Our sense is that the $H(t)$ of (19) also could be a viable candidate for describing degradation and wear.

With the foregoing in place, we let

$$T_x = \inf\{t > 0 : H(t) = x\}, \tag{20}$$

for some barrier $x \geq 0$; that is, T_x is the hitting (killing) time of the process $\{H(t); t \geq 0\}$ to a threshold x. Because $H(t)$ is continuous and increasing, we have that

$$P(T_x > t) = P(H(t) < x). \tag{21}$$

To evaluate the right side of the foregoing, we need to know the density of $H(t)$ for a fixed value of t. For convenience, we denote $H(t)$ by H_t, and, following the notation of Yor (1992), note that

$$\frac{P(H_t \in dv)}{dv} = \sqrt{\frac{2\pi}{v}} \int_0^\infty \exp\left(-\frac{y^2}{2t} + \frac{v}{2}\cosh^2 y\right) \times \sinh y \sin\left(\frac{\pi y}{t}\right)\left(1 - \Phi(\sqrt{v}\cosh y)\right) dy,$$

where $\Phi(u)$ is as defined in Section 6.2.1. Consequently,

$$P(T_x > t) = \int_0^x P(H_t \in dv),$$

from which it follows that

$$P(T > t) = \int_0^\infty P(T_x > t)e^{-x}\, dx, \tag{22}$$

$$= \int_0^\infty \int_0^\infty P(H_t \in dv)e^{-x}\, dx, \tag{23}$$

with $P(H_t \in dv)$ as given earlier.

7. COMPETING-RISK AND DEGRADATION PROCESSES

7.1 Competing Risks and Competing-Risk Processes

Loosely speaking, the term "competing risks" connotes competing causes of failure, and interest centers on the cause of failure and/or the time to failure given that there are several agents competing for an item's lifetime. The issue can be quite complex because the causes do not operate in isolation of one another, it often being the case that one cause acerbates the effect of the other. Traditionally, the model used for encapsulating the scenario of failure under competing risks is the reliability of a series system with independent (or dependent) component lifetimes, the latter representing the causes of system failure. In what follows, we shift focus from independent or dependent lifetimes to independent or dependent HPs to develop a framework that could provide a more realistic description of the competing-risk phenomenon. Accordingly, let T_i denote the time to failure of the ith component of a series system of k components, $i = 1, \ldots, k$, and T the time to failure of a system. Then

$$P(T \geq t) = P\big(H_1(t) \leq X_1, \ldots, H_k(t) \leq X_k\big),$$

where $H_i(t)$ is the cumulative hazard (or risk) experienced by the ith component and X_i is its HP. If the HPs are assumed to be independent, then

$$P(T \geq t) = \exp\big[-\big(H_1(t) + \cdots + H_k(t)\big)\big], \tag{24}$$

suggesting an additivity of the cumulative hazards (or risks). If the HPs are assumed to be dependent, then the nature of dependence would dictate the form taken by $P(T \geq t)$; see Section 5.

In either case, our expression for $P(T \geq t)$ would be the same as what we would obtain assuming the dependence or independence of the lifetimes T_i. Thus it would appear that little, if any, gain has been achieved by shifting focus from the T_i's to the X_i's. But there is another way to look at (24), a way with paves the path for obtaining another expression for the survival function of an item experiencing multiple risks.

Observe that (24) is also the survival function of a single item that has a cumulative hazard of $H(t) \stackrel{\text{def}}{=} \sum_{i=1}^{k} H_i(t)$, at time t. But when this is the case, how can we interpret each $H_i(t)$? More generally, in the case of a single item with a cumulative hazard of $H(t)$, can there be a meaningful decomposition of $H(t)$, and, if so, can it be additive? Moreover, which of the two perspectives more accurately reflects the competing-risk phenomenon?

One possible strategy for addressing these questions is to see each $H_i(t)$, $i = 1, \ldots, k$, as the consequence of a covariate and to suppose that if the item were to experience covariate i alone, then its time to failure would coincide with the item at which $H_i(t)$ crossed its hazard potential X. With the item simultaneously experiencing k covariates, its survival function would be

$$\begin{aligned} P(T \geq t) &= P\big(H_1(t) \leq X, \ldots, H_k(t) \leq X\big) \\ &= P\big(X \geq \max\{H_1(t), \ldots, H_k(t)\}\big) \\ &= \exp\big(-\max\{H_1(t), \ldots, H_k(t)\}\big). \end{aligned} \tag{25}$$

Clearly, under the scenario of an item simultaneously experiencing k causes of failure (risks), the decomposition of $H(t)$ is not additive.

Whereas (25) could be new to the literature on competing risks, it is worth noting that the two scenarios discussed earlier—the traditional one involving a series system that leads to (24) and the one pertaining to the single item that leads to (25)—are related because considering a single HP X is tantamount to considering k HPs that are totally (and positively) dependent on one another. This leads to the following result.

Theorem 4. The survival function under any series system model for competing risks with positively dependent hazard potentials is bounded as

$$\exp\left(-\sum_{i=1}^{k} H_i(t)\right) \leq P(T \geq t) \leq \exp\big(-\max\{H_1(t), \ldots, H_k(t)\}\big).$$

This theorem shows that the two perspectives on competing-risk modeling can be reconciled through the notion of independent and dependent hazard potentials, with the left side of the inequality reflecting the former and the right side reflecting the latter.

7.1.1 Dependent Competing Risks and Competing Risk Processes. In our discussion thus far, the $H_i(t)$'s have been assumed known and specified. Consequently, the matter of independent or dependent competing risks was not germane; dependence and independence were embodied in the context of HPs. But the prevailing view of what constitutes dependent competing risks entails considering dependent lifetimes in the series system model mentioned earlier. We consider this approach circuitous. A proper framework for discussing dependent competing risks requires that the $H_i(t)$'s be random; a comprehensive way of doing this is to assume a stochastic process model $\{H_i(t); t \geq 0\}$, $i = 1, \ldots, k$, as was done in Section 6. We call such a model a competing-risk process, and call the k-variate process $\{H_1(t), \ldots, H_k(t); t \geq 0\}$ a dependent competing-risk process if the $H_i(t)$'s are interdependent. A unit fails when any one of the k marginal processes $\{H_i(t); t \geq 0\}$, $i = 1, \ldots, k$, hits the item's HP X. Interdependence of the $H_i(t)$'s will induce dependence between the corresponding lifetimes T_i, $i = 1, \ldots, k$. Thus the prevailing notion of what constitutes dependent competing risks will be sustained, albeit more as a consequence than as a fundamental construct. Viewing the competing-risk scenario from the standpoint of hitting the HP offers a convenient platform for appreciating the phenomenon of lifetimes under dependent competing risks.

Having stated the foregoing, the question still remains as to what would be suitable models for the k-variate process $\{H_1(t), \ldots, H_k(t); t \geq 0\}$, where the marginal processes $\{H_i(t); t \geq 0\}$, $i = 1, \ldots, k$, are such that each $H_i(t)$ is nondecreasing in t. One possibility would be to let each marginal process be a Brownian maximum process of Section 6.2.1 and deduce the interdependence between the marginal processes from the assumed dependence of the k-variate Brownian motion process that generate Brownian maxima processes. The specifics remain to be worked out. Another possibility, in the case where $k = 2$, is to assume that $\{H_1(t); t \geq 0\}$ is a nonnegative, nondecreasing, and right-continuous process of the type discussed in Section 6.2, but that the sample path of $\{H_2(t); t \geq 0\}$ is an impulse function of the form $H_2(t) = 0$ for all $t \neq t^*$, and $H_2(t^*) = \infty$, for some $t = t^* > 0$, where the rate of impulse occurrence depends on the state of the process $\{H_1(t); t \geq 0\}$. Such a model may be meaningful when the process $\{H_1(t); t \geq 0\}$ can be identified with, say, degradation and the process $\{H_2(t); t \geq 0\}$ can be identified with some form of trauma with a rate of occurrence depending on the state of the degradation process. Here degradation and trauma compete with each other for the lifetime of the system. Lemoine and Wenocur (1985) and Wenocur (1989) have proposed the foregoing as a framework for failure modeling, although not in the context of competing risks. With appropriate modifications, their results could be adapted for the competing-risk scenario.

7.2 Degradation and Aging Processes

Much has been written on what is known as "degradation modeling" and reliability assessment using degradation data. The thinking here has been that degradation is an observable phenomenon and that failure occurs when the level of degradation hits some threshold (see Doksum 1991). What the threshold should be and how it should be specified has not been made clear. Our review of the engineering and materials science literature on degradation suggests that this viewpoint is questionable. This is because degradation is viewed as the irreversible accumulation of damage throughout life that ultimately leads to failure (see Bogdanoff and Kozin 1985, p. 1). Whereas the term "damage" itself is not defined, it is claimed that damage manifests as cracks, corrosion, physical wear (depletion of material), and so on. Similarly, with regard to aging, a review of the literature on longevity and mortality indicates that aging pertains

to a unit's position in a state space in which the probabilities of failure are greater than in a former position and that the manifestations of aging are the biomedical and physical difficulties experienced by older individuals.

Thus it appears that both degradation and aging are abstract constructs that cannot be observed and thus cannot be measured. However, these constructs serve to describe a process that results in failure and can be viewed as the cause of observables such as crack growth and corrosion, which can be measured. Thus the question arises as to how one can mathematically model the degradation phenomenon and relate it to the observables mentioned earlier. Put another way, how can we mathematically describe the cause and effect phenomenon of degradation and the observables that it spawns? Our proposal is to treat the former as a cumulative hazard process and the latter as a covariate (or a marker) process that is influenced by the former (see, e.g., Whitmore, Crowder, and Lawless 1998). This viewpoint of view may fit well with Aalen's (1987) proposal that matters of causality be handled by stochastic process models. As before, the item fails when the cumulative hazard process hits the item's HP X. With the foregoing in mind, we define a degradation process as a bivariate stochastic process $\{H(t), Z(t); t \geq 0\}$, with $H(t)$ representing the unobserved cumulative hazard, and $Z(t)$ representing an observable marker that is a precursor to failure. In principle, $\{Z(t); t \geq 0\}$, the marker process, can also be a vector stochastic process. Whereas $H(t)$ is required to be nondecreasing, there is no such restriction on $Z(t)$; cracks can be repaired and sometimes do heal.

7.2.1 Specifying Degradation Processes. When the marker process can be meaningfully described by a Markov process, for which there is some precedence when the marker is crack growth (see Sobczyk 1987), the degradation process $\{H(t), Z(t); t \geq 0\}$ can be taken to be Cinlar's (1972) Markov additive process (MAP). When this is the case, $\{H(t); t \geq 0\}$ is a Lévy process with parameters depending on the state of the $\{Z(t); t \geq 0\}$ process. Another way to link the two processes in question is to use Cox's (1972) proportional hazards model or Aalen's (1989) additive hazards model, in which linkage is achieved through the processes $\{h(t); t \geq 0\}$ and $\{Z(t); t \geq 0\}$. The ramifications of the foregoing, as well as the MAP, remain to be explored. Our main purpose here is to propose a different approach for examining the degradation phenomenon and the role of the HP in analyzing it.

8. THE HAZARD GRADIENT AND CONDITIONAL HAZARD POTENTIALS

The purposes of this section are to obtain a generalization of Theorem 1 and to further explore the ramifications of dependent life-lengths and dependent HPs. We start with the notion of a "hazard gradient" and provide a strategy through which a collection of dependent lifetimes can be replaced by a collection of independent ones.

Let $T_1, \ldots, T_n$, be a collection of n lifetimes, and let $P(T_1 \geq t_1, \ldots, T_n \geq t_n) = R(t_1, \ldots, t_n)$ be its survival function. Let $\mathbf{t} = (t_1, \ldots, t_n)$ be such that $R(\mathbf{t}) > 0$. The quantity $H(\mathbf{t}) = \ln R(\mathbf{t})$ is the multivariate analog of $H(t)$. Suppose that $H(\mathbf{t})$ has a gradient $\mathbf{r}(\mathbf{t}) = (r_1(\mathbf{t}), \ldots, r_n(\mathbf{t}))$, where $r_i(\mathbf{t}) = \frac{\partial}{\partial t_i} H(\mathbf{t})$, $i = 1, \ldots, n$. The quantity $\mathbf{r}(\mathbf{t})$ is called the hazard gradient of $R(\mathbf{t})$ (see Marshall 1975a).

The relationship among $H(\mathbf{t})$, $R(\mathbf{t})$, and $\mathbf{r}(\mathbf{u})$ is expressed through

$$H(\mathbf{t}) = \int_{\mathbf{0}}^{\mathbf{t}} \mathbf{r}(\mathbf{u})\, d\mathbf{u} \tag{26}$$

and

$$P(T_1 \geq t_1, \ldots, T_n \geq t_n) = \exp\left(-\int_{\mathbf{0}}^{\mathbf{t}} \mathbf{r}(\mathbf{u})\, d\mathbf{u}\right). \tag{27}$$

Marshall (1975a) gave a decomposition of $H(\mathbf{t})$ that is noteworthy due to its role in allowing us to prove Theorem 5. Specifically,

$$H(\mathbf{t}) = \int_0^{t_1} r_1(u_1, 0, \ldots, 0)\, du_1 + \int_0^{t_2} r_2(t_1, u_2, 0, \ldots, 0)\, du_2 + \cdots + \int_0^{t_n} r_n(t_1, \ldots, t_{n-1}, u_n)\, du_n, \tag{28}$$

where $r_1(u_1, 0, \ldots, 0)$ is the failure rate of T_1 at u_1, and $r_i(t_1, \ldots, t_{i-1}, u_i, 0, \ldots, 0)$ is the (conditional) failure rate of T_i at u_i, were it so that $T_1 > t_1, \ldots, T_{i-1} > t_{i-1}$.

The first term on the right side of (28) is the cumulative hazard of T_1 at t_1 and is denoted by $H_1(t_1)$. The second term is the integral of the conditional hazard of T_2 at u_2 given that $T_1 \geq t_1$; it is denoted by $H_2(t_2|t_1)$. Similarly, the last term is denoted by $H_n(t_n|t_1, \ldots, t_{n-1})$. Thus

$$H(\mathbf{t}) = H_1(t_1) + H_2(t_2|t_1) + \cdots + H_n(t_n|t_1, \ldots, t_{n-1}),$$

and because $R(\mathbf{t}) = \exp(-H(\mathbf{t}))$,

$$P(T_1 \geq t_1, \ldots, T_n \geq t_n) = \exp[-H_1(t_1)] \exp[-H_2(t_2|t_1)] \cdots \times \exp[-H_n(t_n|t_1, \ldots, t_{n-1})]. \tag{29}$$

Clearly, $e^{-H_1(t_1)} = P(T_1 \geq t_1)$, and, using arguments that parallel those leading us to (1), we can see that for any $n \geq 2$,

$$\exp[-H_n(t_n|t_1, \ldots, t_{n-1})] = P(T_n \geq t_n | T_1 \geq t_1, \ldots, T_{n-1} \geq t_{n-1}). \tag{30}$$

Let $X_1, \ldots, X_n$, be the HPs corresponding to the lifetimes $T_1, \ldots, T_n$ and the cumulative hazards $H_1(t_1), \ldots, H_n(t_n)$. Then, a consequence of the relationship (29) is that

$$\begin{aligned} &P(T_n \geq t_n | T_1 \geq t_1, \ldots, T_{n-1} \geq t_{n-1}) \\ &= P\big(X_n \geq H_n(t_n) | X_1 \geq H_1(t_1), \ldots, X_{n-1} \geq H_{n-1}(t_{n-1})\big) \\ &= \exp[-H_n(t_n|t_1, \ldots, t_{n-1})]. \end{aligned} \tag{31}$$

Because $T_1, \ldots, T_n$ are not independent, the HPs $X_1, \ldots, X_n$ are, by virtue of Remark 1, also not independent. However, the hand side of (31) is the distribution function of an exponentially distributed random variable, say X_n^*, with a scale parameter of 1, evaluated at $H_n(t_n|t_1, \ldots, t_{n-1})$. Thus, from (30), we have the result that for all $n \geq 2$,

$$\begin{aligned} &P(T_n \geq t_n | T_1 \geq t_1, \ldots, T_{n-1} \geq t_{n-1}) \\ &= P\big(X_n \geq H_n(t_n) | X_1 \geq H_1(t_1), \ldots, X_{n-1} \geq H_{n-1}(t_{n-1})\big) \\ &= P(X_n^* \geq H_n(t_n|t_1, \ldots, t_{n-1})). \end{aligned}$$

The quantity X_n^* is called the conditional HP of the nth item; its unit exponential distribution is indexed by $H_n(t_n|t_1, \ldots, t_{n-1})$. In contrast, X_n, the HP of the nth item, has a unit exponential distribution indexed by $H_n(t_n)$.

Similarly, corresponding to each term on the right side of (28) except the first, there exist random variables $X_2^*, \ldots, X_{n-1}^*$, independent of one another, and also of X_n^*, such that

$$P(T_1 \geq t_1, \ldots, T_n \geq t_n) = P(X_1 \geq H_1(t_1))P(X_2^* \geq H_2(t_2|t_1)) \cdots \times P(X_n^* \geq H_n(t_n|t_1, \ldots, t_{n-1})).$$

We have now proved, as a multivariate analog to Theorem 1, the following results.

Theorem 5. Corresponding to every collection of nonnegative variables $T_1, \ldots, T_n$, having a survival function $R(t_1, \ldots, t_n)$, there exists a collection of n independent and exponentially distributed random variables $X_1, X_2^*, \ldots, X_n^*$, with scale parameter 1; X_1 is indexed on $H_1(t_1)$, and for $n \geq 2$, X_n^* is indexed on $H_n(t_n|t_1, \ldots, t_{n-1})$.

9. SUMMARY

In this article we have described a unifying perspective on the process leading to the failure of items that is context-independent. This perspective is made possible through the notion of an HP. Besides providing an alternative means of conceptualizing the failure process, the HP provides a means by which the nature of dependence between the lifetimes can be understood and exploited. With respect to the latter, we can generate (new) families of multivariate failure distributions using multivariate exponentials with unit exponential marginals as seeds. For items required to operate in dynamic environments, the HP provides a vehicle by which new families of univariate survival functions can be obtained. This is achieved by establishing a connection between the failure process and the killing times of continuous and increasing stochastic processes to a random barrier, which is the HP. The notion of a HP generalizes to a nonexponential distribution for the barrier and also to the multivariate case. To conclude, the importance of the notion of a HP stems from its ability to provide a different perspective on failure, a model for the cause of dependence of lifetimes, new multivariate models for failure, new univariate models for survival in dynamic environments, and a perspective on competing risks and degradation modeling.

This article is expository in the sense that it provides a feel for the foregoing possibilities. Clearly, more can be done. For one, stochastic processes other than those considered in Section 6.2 can be investigated. We may do more on considering covariates that drive the $\{H(t); t \geq 0\}$ process. Another possibility would be to consider bivariate processes and their killing times by interdependent barriers. In regard to the latter, one may also be able to leverage the idea for assessing competing risks by looking at the bivariate cumulative hazard process. Finally, there is a matter of statistical inference and model validation, topics that have not been touched on here. The possibilities of further capitalizing the notion of an HP are promising for reliability theorists, survival analysts, and actuarial scientists.

[Received November 2005. Revised June 2006.]

REFERENCES

Aalen, O. O. (1987), "Dynamic Modeling and Causality," *Scandinavian Actuarial Journal*, 12, 177–190.

——— (1989), "A Linear Regression Model for the Analysis of Life Times," *Statistics in Medicine*, 8, 907–925.

Aven, T., and Jensen, U. (1999), *Stochastic Models in Reliability*, New York: Springer-Verlag.

Bagdonavicius, V., and Nikulin, M. S. (1999), "Model Building in Accelerated Experiments," in *Statistical and Probabilistic Models in Reliability*, eds. D. C. Ionescu and N. Limnios, Boston: Birkhauser, pp. 51–73.

Barlow, R. E., and Proschan, F. (1975), *Statistical Theory of Reliability and Life Testing*, New York: Holt, Rinehart and Winston.

Blumenthal, R. M., Getoor, R. K., and McKean, H. P. (1962), "Markov Processes With Identical Hitting Distributions," *Illinois Journal of Mathematics*, 6, 402–420.

Bogdanoff, J. L., and Kozin, F. (1985), *Probabilistic Models of Cumulative Damage*, New York: Wiley.

Cinlar, E. (1972), "Markov Additive Processes II," *Zeitschrift für Wahrscheinlichkeitstheare und verwandte Gebiete*, 24, 94–121.

——— (1979), "On Increasing Continuous Processes," *Stochastic Processes and Their Applications*, 9, 147–154.

Cinlar, E., and Ozekici, S. (1987), "Reliability of Complex Devices in Random Environments," *Probability in the Engineering and Informatianal Sciences*, 1, 97–115.

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal af the Rayal Statistical Saciety*, Ser. B, 34, 187–220.

Doksum, K. A. (1991), "Degradation Models for Failure Time and Survival Data," *CWI Quarterly, Amsterdam*, 4, 195–203.

Duchesne, T., and Rosenthal, J. S. (2003), "On the Collapsibility of Lifetime Regression Models," *Advances in Applied Probability*, 35, 755–772.

Durham, S. D., and Padgett, W. J. (1997), "A Cumulative Damage Model for System Failure With Application to Carbon Fibers and Composites," *Technometrics*, 39, 34–44.

Esary, J. D., Marshall, A. W., and Proschan, F. (1973), "Shock Models and Wear Processes," *The Annals of Applied Probability*, 1, 627–649.

Gavrilov, L. A., and Gavrilova, N. S. (2001), "The Reliability Theory of Aging and Longrvity," *Jaurnal af Thearetical Bialogy*, 213, 527–545.

Gjessing, H. K., Aalen, O. O., and Hjort, N. (2003), "Frailty Models Based on Lévy Processes," *Advances in Applied Probability*, 35, 532–550.

Grandell, J. (1975), *Doubly Stochastic Poisson Processes*, New York: Springer-Verlag.

Gumbel, E. J. (1960), "Bivariate Exponential Distributions," *Jaurnal af the American Statistical Association*, 55, 698–707.

Kebir, Y. (1991), "On Hazard Rate Processes," *Naval Research Logistics*, 38, 865–876.

Kingman, J. F. C. (1964), "On Doubly Stochastic Poisson Processes," *Proceedings af the Cambridge Philosophical Society*, 60, 923–960.

Kotz, S., Balakrishman, N., and Johnson, N. L. (2000), *Continuous Multivariate Distributions*, New York: Wiley.

Lemoine, A. J., and Wenocur, M. L. (1985), "On Failure Modeling," *Naval Research Logistics Quarterly*, 32, 497–508.

Lindley, D. V., and Singpurwalla, N. D. (1986), "Multivariate Distributions for the Lifelengths of Environment," *Journal of Applied Prabability*, 23, 418–431.

Marshall, A. W. (1975a), "Some Comments on the Hazard Gradient," *Stachastic Processes and Their Applications*, 3, 293–300.

——— (1975b), "Multivariate Distributions With Monotone Hazard Rate," in *Reliability and Fault Tree Analysis*, eds. J. B. Fussell and N. D. Singpurwalla, Philadelphia: Society for Industrial and Applied Mathematics, pp. 259–284.

Marshall, A. W., and Olkin, I. (1967), "A Multivariate Exponential Distribution," *Jaurnal af the American Statistical Association*, 62, 30–44.

Nelson, R. B. (1995), "Copulas, Characterization, Correlation and Counterexamples," *Mathematics Magazine*, 68, 193–198.

Pettit, L. I., and Young, K. D. S. (1999), "Bayesian Analysis for Inverse Gaussian Lifetime Data With Measures of Degradation," *Jaurnal af Statistical Camputatian and Simulation*, 63, 217–234.

Pitman, J. W., and Speed, T. P. (1973), "A Note on Random Times," *Stachastic Processes and Their Applications*, 1, 369–374.

Protter, P. (1990), *Stochastic Integratian and Differential Equations: A New Appraach*, Berlin: Springer-Verlag.

Sedyakin, N. M. (1966), "On One Physical Principle in Reliability Theory," *Technical Cybernetics*, 3, 80–87.

Singpurwalla, N. D. (1995), "Survival in Dynamic Environments," *Statistical Science*, 10, 86–103.

Singpurwalla, N. D., and Kong, C. W. (2004), "Specifying Interdependence in Networked Systems," *IEEE Transactions in Reliability*, 53, 401–405.

Singpurwalla, N. D., and Wilson, S. P. (1995), "The Exponentiation Formula of Reliability and Survival: Does It Always Hold?" *Lifetime Data Analysis*, 1, 187–194.

Singpurwalla, N. D., and Youngren, M. A. (1993), "Multivariate Distributions Induced by Dynamic Environments," *Scandinavian Journal of Statistics*, 20, 251–261.

Sobczyk, K. (1987), "Stochastic Models for Fatigue Damage of Materials," *Advances in Applied Probability*, 19, 652–673.

Suppes, P. (1970), *A Probabilistic Theory of Causality*, Amsterdam: North-Holland.

Wenocur, M. L. (1989), "A Reliability Model Based on Gamma Process and Its Analytic Theory," *Advanced Applied Probability*, 21, 899–918.

Whitmore, G. A., Crowder, M. J., and Lawless, J. F. (1998), "Failure Inference From a Marker Process Based on a Bivariate Wiener Process," *Lifetime Data Analysis*, 4, 229–251.

Yang, Y., and Klutke, G. A. (2000), "Lifetime Characteristics and Inspection Schemes for Lévy Degradation Processes," *IEEE Transactions on Reliability*, 49, 377–382.

Yashin, A., and Arjas, E. (1988), "A Note on Random Intensities and Conditional Survival Functions," *Journal of Applied Probability*, 25, 630–635.

Yor, M. (1992), "On Some Exponential Functionals of Brownian Motion," *Advances in Applied Probability*, 24, 509–531.

(7)

Chapter 5

RELIABILITY AND SURVIVAL IN FINANCIAL RISK

Nozer D. Singpurwalla
Department of Statistics
The George Washington University, Washington, DC, U.S.A.
Email: nozer@gwu.edu

The aim of this paper is to create a platform for developing an interface between the mathematical theory of reliability and the mathematics of finance. This we are able to do because there exists an isomorphic relationship between the survival function of reliability, and the asset pricing formula of fixed income investments. This connection suggests that the exponentiation formula of reliability theory and survival analysis be reinterpreted from a more encompassing perspective, namely, as the *law of a diminishing resource*. The isomorphism also helps us to characterize the asset pricing formula in non-parametric classes of functions, and to obtain its crossing properties. The latter provides bounds and inequalities on investment horizons. More generally, the isomorphism enables us to expand the scope of mathematical finance and of mathematical reliability by importing ideas and techniques from one discipline to the other. As an example of this interchange we consider interest rate functions that are determined up to an unknown constant so that the set-up results in a Bayesian formulation. We may also model interest rates as "shot-noise processes", often used in reliability, and conversely, the failure rate function as a Lévy process, popular in mathematical finance. A consideration of the shot noise process for modelling interest rates appears to be new.

Key words: Asset pricing; Bayesian analysis; Failure rate; Interest rate; Non-parametric classes; Risk-free bond; Shot-noise process; Zero coupon bond.

1 Introduction

An area that has recently experienced an outburst of activity in the mathematical sciences is what is known as "financial risk analysis". However reliability theory, and survival analysis, which are some of the stalwart tools of risk analysis, have played little or no role in mathematical finance. There are many plausible causes behind the absence of a synergism between these two fields. One reason could be that the term "financial risk" needs to be better articulated and defined. Whereas mathematical finance has benefitted from topics in stochastic processes, statistical inference, and probabilistic modelling, the constructive role that reliability theory is able to play here remains to be exploited. The aim of this paper is to point out some avenues via which the above can be done. To do so, we use the well known asset pricing formula of a fixed income instrument (like a risk-free zero coupon bond) as a "hook". Underlying this formula is the use of an unknown (future) interest rate function. This can be a deterministic function or the realization of a stochastic process. We liken the interest rate function to a (deterministic or stochastic) failure rate function, and then using results from reliability theory explore its consequences on asset pricing. Among these consequences are bounds and inequalities on the investment horizon.

Associated with any failure rate function is a survival function. As currently interpreted, this function encapsulates the risk of failure of an item over time. By risk we mean here probability of failure. Since a risk-free zero coupon bond cannot (by definition) default, the survival function that results from looking at the interest rate as a failure rate cannot encapsulate the risk of a bond's default. This dilemma motivates us to seek alternate, more global, ways of interpreting the survival function. Our view is that the survival function be viewed as one that describes the phenomenon of a diminishing resource over time. In mathematical finance, this resource is a bond's present value; in reliability theory this resource is an item's "hazard potential" [cf. Singpurwalla (2004)] With the above perspective on a survival function the failure rate can be seen as the rate at which the item's hazard potential gets depleted, and the interest rate as the rate at which a bond's present value shrinks. This interpretation of the interest rate appears to be new and can be seen as one of the merits of the isomorphism between survival analysis and mathematical finance.

The remainder of this paper is organized as follows. In Section 2 we give an overview of the derivation of the asset pricing formula under both constant and varying interest rates and point out the relationship between this formula and the survival function of reliability theory. The material of this section is standard and can be found, for example, in Ross (1999). In Section 3 we present arguments that attempt to give a unifying perspective for the present value and the survival functions. Drawing from an analogy in physics about the decay of radioactive material, we claim that the two formulae in question describe the law of diminishing resource. In Section 4 we invoke several ideas and results from reliability theory to characterize present value functions into non-parametric classes and show how some of these results can be exploited for practical purposes. Section 5 pertains to a discussion of interest rate functions with unknown parameters, or as realizations of stochastic processes. In both cases we draw upon known results in reliability with a view of enhancing the state of the art in mathematical finance. Section 6 concludes the paper with some pointers for future research.

2 Asset Pricing of Risk Free Bonds: An Overview

The material of this section is for the benefit of those working in reliability and survival analysis whose familiarity with the various instruments of finance may be limited. The focus here is the derivation of the asset pricing formula for a risk free bond assuming a deterministic and known interest rate function. The section ends by pointing out the isomorphism between the asset pricing formula and the exponentiation formula of reliability and survival analysis.

A risk free *zero coupon bond* pays, with certainty, the buyer of the bond—the *bondholder*— \$1 at time T after the time of purchase; T is known as the *holding period* (of the buyer) or *maturity*. The bondholder purchases the bond at some calendar time t at a price $P(t,T)$, known as the *present value* at t. Clearly, $P(T,T) = 1$, and $P(t,T)$ decreases in T. Risk-free bonds are generally issued by governments and do not default because governments can always honor payments by "printing" their currency. The present value $P(t,T)$ depends on what the bond holder and the *bond issuer* think of the interest rate that will prevail during the period $(t, t+T]$.

2.1 *Interest Rates and Present Value Analysis*

To keep matters simple, suppose that an amount P is borrowed now, at time $t = 0$, for a period T with the understanding that at time T the amount returned is $P + rP = P(1+r)$. The amount P is known as the *principal* and r the *simple interest* rate per time T. When T is taken to be one year, r is the simple annual interest rate, and the compounding of interest is once per year. If the interest rate is compounded semi-annually, then the amount paid at the end of the year is $P(1+r/2)^2$. In this case r is called the

nominal interest rate. If the compounding is done n times per year then the amount paid at the end of the year is $P(1+r/n)^n$ and with *continuous compounding* the amount paid at year's end is $P \lim_{n\to\infty}(1+r/n)^n = Pe^r$.

With present value analysis we consider the reverse of the above process. Specifically, what should P be at time $t = 0$ so that at the end of the i-th period of compounding the amount paid (or *payoff*) is V, supposing that the nominal interest rate is r? it is easy to see that the principal $V(1+r)^{-i}$ would yield V at time i. The quantity $V(1+r)^{-i}$ is known as the *present value* at time $t = 0$ of the payoff V at time $t = i$.

2.1.1 *Present Value Under Varying Interest Rates*

Suppose that the nominal interest rate changes with time continuously, as $r(s)$, $s \geq 0$. The quantity $r(s)$ is called the *spot* (or *instantaneous*) interest rate at s. Consequently, an amount x invested at time s becomes $x(1+r(s)h)$ at time $s+h$—approximately—assuming that h is small. Let $D(T)$ denote the amount one has at time T if one invests one monetary unit at time 0. Then for h small and interest rate $r(s)$, $0 \leq s \leq T$

$$D(s+h) \approx D(s)(1+r(s)h),$$

or that the rate of change of the amount at time s is

$$\frac{D(s+h)-D(s)}{h} \approx D(s)r(s).$$

Taking the limit as $h \downarrow 0$, we have

$$\lim_{h\downarrow 0} \frac{D(s+h)-D(s)}{h} = D(s)r(s),$$

or that

$$r(s) = \frac{D'(s)}{D(s)},$$

where $D'(s)$ is the derivative of $D(s)$ at s, assuming it exists at any s.

Integrating over s from $[0,T]$, we have

$$\log(D(T)) - \log(D(0)) = \int_0^T r(s)ds.$$

Since $D(0) = 1$, the above can be written as

$$D((T))^{-1} = \exp\left[-\int_0^T r(s)ds\right].$$

But $(D(T))^{-1}$ is $P(0,T)$, the present value at time 0 of a bond that pays one monetary unit at time T. Thus in general we have the relationship

$$P(t,T) = \exp\left[-\int_t^{T+t} r(s)ds\right], \tag{1}$$

where$P(t,T)$ is the present value, at time t, of a risk free bond yielding one monetary unit at time $t+T$, under a continuously changing interest rate $r(s)$, s≥ 0. If we let $R(t,T)$ denote the exponent of the expression for $P(t,T)$, then

$$P(t,T) = \exp(-R(t,T)).$$

The average of the spot interest rate $r(s)$ is

$$\tilde{R}(t,T) = \frac{1}{T}\int_t^{T+t} r(s)ds; \tag{2}$$

it is called the *yield curve*.

2.2 *Isomorphism with the Survival Function*

Mathematically, Equation (1) is identical to the *exponentiation formula* of reliability theory and survival analysis with $r(s)$ as the failure rate function, and $P(t,T)$ as the survival function. Observe that $P(t,0) = 1$ and $P(t,T)$ is a decreasing function of T, which asymptotes to 0 as T increases to infinity. Similarly $R(t,T)$ can be identified with the *cumulative failure* (or hazard) rate at T, and $\tilde{R}(t,T)$—yield curve—with the *failure rate average*.

As two special cases, suppose that $r(s) = r$, a constant, for $s \geq t$, or that $r(s) = \alpha r(rs)^{\alpha-1}$, for $s \geq t$ and some constant $\alpha \geq 1$. Then $P(t,T) = \exp(-r(T-t))$ in the first case, and $P(t,T) = \exp(-r(T-t)^\alpha)$ in the second. These present value functions would correspond to the exponential and the Weibull survival functions, respectively.

3 Re-interpreting the Present Value and Survival Functions

In what follows, we set $t = 0$, so that $P(t,T)$ becomes $P(0,T) \stackrel{\text{def}}{=} P(T)$, $R(0,T) \stackrel{\text{def}}{=} R(T)$, and $\tilde{R}(0,T) \stackrel{\text{def}}{=} \tilde{R}(T)$. In the context of reliability and survival analysis the interpretation of $P(T)$ as a survival function, $r(s)$ as the failure rate function, and $\tilde{R}(T)$ as the failure rate average have an intuitive import that is embedded in the context of ageing and wear. How can one justify looking at $P(T)$, the present value function as a survival

function and the interest rate $r(s)$ as a hazard function? Alternatively put, how can one see the relationships

$$P(T) = \exp\left[-\int_0^T r(s)ds\right], \tag{3}$$

and

$$R(T) = \int_0^T r(s)ds, \tag{4}$$

from the perspective of hazard, risk and failure, especially since risk-free bonds do not default? More specifically, we may ask if there is a common theme—different from the ones in reliability and finance—that drives the likes of Equations (3) and (4)? The aim of this section is to show that there is indeed a common theme that is able to provide meaning to the above equations in a unified manner. This common theme causes us to look at the exponentiation formula of Equation (3) as encapsulating the phenomenon of a depleting resource. However, in order to do so, we need to first revisit the derivation of the exponentiation formula from first principles. The material that follows is standard and found in Barlow and Proschan (1975).

To keep our notation distinct, let X denote the time to failure of an item and let $F(x) = Pr(X \leq x)$. Suppose that $F(x)$ is absolutely continuous so that its derivative $\frac{dF(x)}{dx} \stackrel{\text{def}}{=} f(x)$ exists (almost everywhere). We now consider

$$Pr(x < X \leq x + dx | X > x) = \frac{F(x+dx) - F(x)}{\overline{F}(x)},$$

where $\overline{F}(x) = 1 - F(x)$. If we divide both sides of the above expression by dx, we get a rate in the sense that

$$\frac{1}{\overline{F}(x)} \frac{F(x+dx) - F(x)}{dx}$$

is the rate at which $F(x)$ increases at x, multiplied by $(\overline{F}(x))^{-1}$. Taking the limit as $dx \downarrow 0$, we have

$$\lim_{dx \downarrow 0} \frac{F(x+dx) - F(x)}{\overline{F}(x)dx} = \frac{f(x)}{\overline{F}(x)} \stackrel{\text{def}}{=} h(x). \tag{5}$$

The right hand side of Equation (5) is defined as the *failure* (or *hazard*) rate function, denoted here as $h(x)$. The qualifier "failure" is added because the function $F(x)$ whose rate of increase is being discussed represents the probability of failure by x. A motivation for referring to $h(x)$ as a rate has been given above. Namely, it is the rate at which the distribution function $F(x)$ increases in x. The exponentiation formula of Equation (3) is an immediate consequence of the relationship $h(x) = f(x)/\overline{F}(x)$.

It is important to note that the development above is not contingent on the fact that $F(x)$ necessarily be a probability distribution function. All that we require is for $F(x)$ to be absolutely continuous with respect to Lebesgue measure, and that for $h(x)$ to be non-negative $F(x)$ be non decreasing. To underscore this point, and also to pave the path for looking at the asset pricing formula from the point of risk and reliability, we turn to a scenario from physics, a scenario that does not involve failure nor does it involve the probability of failure. What we have in mind is the decay of radioactivity (as a function of time) of certain materials, say carbon 14. But before we do so, it is useful to note that Equation (5) may also be written as

$$\lim_{dx \downarrow 0} \frac{1}{\overline{F}(x)} \frac{\overline{F}(x+dx) - \overline{F}(x)}{dx} = -h(x),$$

so that $-h(x)$ encapsulates the rate at which $\overline{F}(x)$ decreases in x.

3.1 *The Exponentiation Formula as the Law of a Diminishing Resource*

Turning to the problem of radioactive decay, it has been claimed that for certain materials the amount of radioactivity decreases exponentially over time, so that if $H(t)$ denotes the level of radioactivity at time t, then $H(t) = \exp(-\lambda t)$, for some $\lambda > 0$. Note that $H(t)$ is absolutely continuous and behaves like a survival function. The rate at which this function decreases is $-\lambda \exp(-\lambda t)$, and so now our analogue of $h(t)$ is $\lambda \exp(-\lambda t)/H(t) = \lambda$, a constant. The exponentiation formula of Equation (3) holds here as well, though it does not have the interpretation used in reliability. Our position here is that the exponentiation formula is ubiqutious in any scenario involving an absolutely continuous monotonically decreasing function, the interpretation of the function being context dependent. In reliability, it is the item's survival function; in radioactivity it is the amount of radioactivity that is remaining, and in finance it is the present value at any time T.

3.1.1 *Interest Rate as a Proportion Loss in Present Value*

In the context of reliability, the quantity $h(x)dx$ is, approximately, the conditional probability of failure at x. In the context of radioactive decay, λdt is the proportion of radioactive loss in the time interval t, $t + dt$. This

interpretation will hold irrespective of the functional form of $H(t)$. The interpretation has a broader ramification in the sense that when $P(T)$ denotes the present value at time T and $r(s)$ is the interest rate, then $r(s)ds$ is the proportion loss of present value at time s in the interval s, $s + ds$. Thus one may liken the interest rate as a form of a hazard or risk posed to the present value of the function vis a vis its failure to maintain a particular value at any time. We now have at hand a point of view that unites the failure rate function and the interest rate function.

Our theme of interpreting interest rate as a proportion loss in present value has a synergetic effect in reliability. Specifically, since the survival function $\overline{F}(x)$ decreases in x from $\overline{F}(0) = 1$, the exponentiation formula of Equation (3) can be seen as a law which prescribes life-times as a consequence of some diminishing resource, with $\overline{F}(0) = 1$ interpreted as an item's initial resource. This resource gets depleted over time, with the proportion depleted at x being of the form $h(x)dx$. The amount of resource at x is given by the exponentiation formula of Equation (3).

Thus to recap, the well known exponentiation formula of reliability and survival can also be seen as a law governing a depletion of a resource, with the proportion loss at x governed by the failure[interest] rate $h(x)[r(x)]$. This interpretation is a consequence of the isomorphism between the survival and present value functions. We have now established a platform for discussing financial risk from the point of view of more traditional tools of risk analysis, namely, reliability theory and survival analysis. In what follows we show how this common platform enables us to import some ideas and notions from the latter to the former, and vice versa.

4 Characterizing Present Values Under Monotone Interest Rates

This section is mainly directed towards those working in mathematical finance. Its aim is to describe the qualitative behavior of the present value function $P(T)$ when the underlying interest function $r(s)$, $s \leq T$, or the yield curve $\widetilde{R}(T)$, is monotonic (increasing or decreasing) in T. By increasing (decreasing) we mean non-decreasing (non-increasing); thus a constant interest rate function is both increasing and decreasing. When a bond is issued, the precise nature of the interest rate that will prevail during the life of the bond will not be known. However, one can speculate its general nature as being edging upwards or downwards depending on ones view about the strength of the economy. Thus the objective here is to characterize the behavior of $P(T)$ when the interest rate function, or the yield curve is monotonic but not precisely known. The practical motivation for

characterizing present value functions will become clear in what follows. For now it suffices to say that such characterizations facilitate a comparison with present value functions under constant interest rate functions and enable one to obtain bounds and inequalities for investment horizons. The exercise here parallels that in reliability theory wherein comparison against the exponential survival function has proved to be valuable.

4.1 *Non-parametric Classes of Present Value Functions*

By a non parametric class of present value functions, we mean a class of functions whose precise form is unknown (i.e. they are not parametrically defined) but about which some general features can be specified.

Definition 1. The present value function $P(T)$ is defined to be IIR (DIR)—for increasing (decreasing) interest rate—if for each $\tau \geq 0$, $P(T + \tau)/P(T)$ is decreasing (increasing) in $T \geq 0$.

A consequence of Definition 1 is that when $P(T)$ is absolutely continuous the interest rate function $r(T)$ is increasing (decreasing) in T. Conversely, when $r(T)$ is increasing (decreasing) in T, $P(T)$ is IIR (DIR). When $r(t) = \lambda$, a constant greater than 0, $P(T) = \exp(-\lambda T)$, which is both IIR and DIR. All present value functions that display the IIR (DIR) property constitute a class that we label "IIR (DIR) class".

Interest rate functions are often not monotonic even though they may reflect a tendency to edge upwards. They may contain aberrations (or kinks) that are not too severe, in the sense that their average is monotone. In other words, whereas $r(T)$ is not monotone, the yield curve $\widetilde{R}(T)$ is monotone. To bring this feature into play we introduce

Definition 2. The present value function $P(T)$ is defined to be IAIR (DAIR)—for increasing (decreasing) average interest rate—if $-[\log P(T)]/T$ is increasing (decreasing) in $T \geq 0$.

A consequence of Definition 2 is that $P(T)$ IAIR (DAIR) is tantamount to $\widetilde{R}(T)$ increasing (decreasing) in $T \geq 0$. Analogous to IIR (DIR) class, we define the IAIR (DAIR) class as a collection of functions $P(T)$ that display the IAIR (DAIR) property. Verify that the IAIR class, denoted $\{IAIR\}$, encompass the IIR class—denoted $\{IIR\}$—so that $\{IIR\} \subseteq \{IAIR\}$. Similarly $\{DIR\} \subseteq \{DAIR\}$.

A further generalization of Definitions 1 and 2, a generalization whose merits will be pointed out later, is obtained via Definition 3 below.

Definition 3. The present value function $P(T)$ is said to display a NWO (NBO)—for new worse (better) than old—property if for each τ, $T \geq 0$,

$P(T+\tau) \leq (\geq)\, P(T)P(\tau)$.

It can be shown—details omitted [cf. Barlow and Proschan (1975)] – that

$$\{IIR\} \subseteq \{IAIR\} \subseteq \{NWO\},$$

and

$$\{DIR\} \subseteq \{DAIR\} \subseteq \{NBO\},$$

where the $\{NWO\}$ and the $\{NBO\}$ classes contain all present value functions that display the NWO and NBO property, respectively.

4.1.1 *Financial Interpretation of NBO (NWO) Feature*

Consider the case of equality in Definition 3. Now

$$P(T+\tau) = P(T)P(\tau), \tag{6}$$

and the above relationship holds if and only if $P(T) = \exp(-\lambda T)$, for some $\lambda \geq 0$ and $T \geq 0$. The interest rate function underlying this form of the present value function is $r(s) = \lambda$. Equation (6) also implies that

$$\frac{P(T) - P(T+\tau)}{P(T)} = 1 - P(\tau),$$

and since $P(0) = 1$, the above relationship can also be written as

$$\frac{P(T) - P(T+\tau)}{P(T)} = \frac{P(0) - P(\tau)}{P(0)}. \tag{7}$$

Because $P(T)$ is a decreasing function of T, the left hand side of Equation (7) describes the proportion loss in present value during a time interval $[0, \tau]$ at the time T, whereas the right hand side describes the proportion loss in the same time interval, but at time 0. This is an *analogue of the memoryless property* of the exponential distribution in the context of finance. Its practical consequence is that under a constant interest rate function, there is no reason to prefer one investment horizon over another, so long as the holding period is the same.

We now consider the case of strict inequality. Suppose that $P(T)$ is NWO, so that

$$P(T+\tau) < P(T)P(\tau),$$

and as a consequence

$$\frac{P(T) - P(T+\tau)}{P(T)} < \frac{P(0) - P(\tau)}{P(0)}. \tag{8}$$

This means that under Equation (8) the proportion loss in present value at some time $T > 0$ is always less than the proportion loss at time 0. Vice-versa when $P(T)$ is NBO and the inequality above is reversed. To a bondholder, the greater the drop in present value, the more attractive is the bond. Consequently, for $P(T)$'s that are NWO, an investment for any fixed holding period that is made early on in the life of the bond is more attractive than one (for the same holding period) that is made later on. In the IIR or the IAIR case, the above claim makes intuitive sense because the aforementioned properties are a manifestation of increasing interest rates and increasing yield curves, and $\{IIR\} \subseteq \{IAIR\} \subseteq \{NWO\}$. A similar claim can be made in the case of $P(T)$ that is NBO.

It is of interest to note that our definition of NWO and NBO is a reverse of that used in reliability theory, namely, the NBU and NWU classes. This makes sense, because a decrease of the present value function is a consequence of an earned resource (namely interest) whereas the decrease of the survival function is a consequence of a depleted resource.

4.2 *Present Value Functions that are Log Concave and PF_2*

Suppose that the present value function $P(T)$ belong to one of the several non-parametric classes introduced in Section 4.1, and suppose that the spot interest rate at time of issue of bond is $\lambda > 0$. Were the interest rate over the investment horizon T to remain a constant at λ, then the present value function should be of the form $\exp(-\lambda T)$, $T \geq 0$. The purpose of this section is to compare $P(T)$ and $\exp(-\lambda T)$. Such a comparison could provide new insights about desirable asset pricing investment horizons. To do so, we need to introduce the notions of log concavity and Polya Frequency Functions of Order 2 – abbreviated PF_2. These notions have turned out to be useful in reliability theory.

Definition 4. A function $h(x)$, $-\infty < x < \infty$ is said to be PF_2 if: $h(x) \geq 0$ for $-\infty < x < \infty$, and

$$\begin{vmatrix} h(x_1 - y_1) & h(x_1 - y_2) \\ h(x_2 - y_1) & h(x_2 - y_2) \end{vmatrix} \geq 0$$

for all $-\infty < x_1 < x_2 < \infty$ and $-\infty < y_1 < y_2 < \infty$, or equivalently $\log h(x)$ is concave on $(-\infty, +\infty)$, or equivalently for fixed $\Delta > 0$, $h(x + \Delta)/h(x)$ is decreasing in x for $a \leq x \leq b$, where

$$a = \inf_{h(y)>0} y \quad \text{and} \quad b = \sup_{h(y)>0} y.$$

The above equivalencies are given in Barlow and Proschan (1975, p.76). Log concavity and PF_2 enable us to establish crossing properties of $P(\bullet)$.

To start with, suppose that $P(\bullet)$ is IIR (DIR). Then, from Definition 1 we have that for each $\tau \geq 0$, $P(T+\tau)/P(T)$ is decreasing (increasing) in $T \geq 0$. As a consequence we have:

Claim 1: $P(\bullet)$ IIR is equivalent to $P(\bullet)$ being both log-concave and PF_2.

Since $P(\bullet)$ IIR is equivalent to an increasing interest rate function $r(\bullet)$, and vice-versa, the essence of Claim 4.1 is that increasing interest rate functions lead to log-concave present value functions. What is the behavior of $P(\bullet)$ if instead of the interest rate function being increasing it is the yield curve that is increasing? More generally, suppose that $P(\bullet)$ is IAIR (DAIR). Then, $P^{1/T} \downarrow (\uparrow) T$, for $T \geq 0$; see Definition 4.2. Consequently we have

Claim 2: *$P(\bullet)$ IAIR (DAIR) implies that for all $T \geq 0$ and any α, $0 < \alpha < 1$,*

$$P(\alpha T) \geq (\leq) P^{\alpha}(T). \tag{9}$$

To interpret Equation (9), let $Q(T) = 1/P(T)$. Then $Q(T)$ is the amount received at time T for every unit of money invested at time $T = 0$. Consequently taking reciprocals in Equation (4.4), we have

$$Q(T/2) \leq (\geq)(Q(T))^{1/2}.$$

Thus, here again, long investment horizons yield more bang for a buck than short horizons when the yield curve is monotonic increasing, and vice-versa when the yield curve is monotone decreasing. Claim 2 prescribes how the investment horizon scales.

To explore the crossing properties of present value functions that are IAIR (DAIR), we introduce

Definition 5. A function $h(x)$, $0 \leq x \leq \infty$ is said to be star-shaped if $h(x)/x$ is increasing in x. Otherwise, it is said to be anti star-shaped. Equivalently, $h(x)$ is star-shaped (anti star-shaped), if for all α, $0 \leq \alpha \leq 1$,

$$h(\alpha x) \leq (\geq) \alpha h(x).$$

It is easy to verify that any convex function passing through the origin is star-shaped. [cf. Barlow and Proschan (1975, p.90)]

Since $P(\bullet)$ IAIR (DAIR) implies — see Definition 2 — that $-[\log P(T)]/T$ is increasing (decreasing) in $T \geq 0$, it now follows that

Claim 3: *$P(\bullet)$ IAIR (DAIR) implies that $T(\tilde{R}(T))$ is star-shaped (anti star-shaped).*

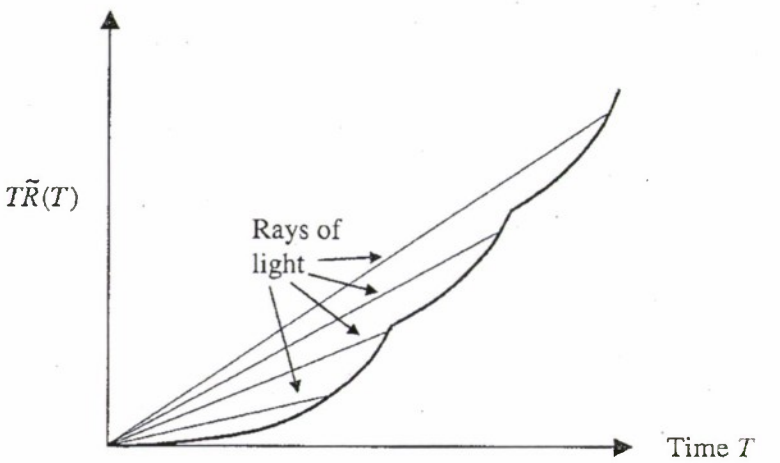

Figure 1 Star-Shapedness of $T(\tilde{R}(T))$ when $P(\cdot)$ is IAIR

Recall that $\tilde{R}(T)$ is the yield curve. The star-shapedness property, illustrated above, is useful for establishing Theorem 4.1 which gives bounds on $P(\bullet)$. The essence of the star-shapedness property is that there exists a point from which a ray of light can be drawn to all points of the star-shaped function $T(\tilde{R}(T)) = \int_0^T r(u)du$, with the origin as the point from which the rays of light can be drawn.

It is clear from an examination of Figure 1, that a star-shaped function can cross a straight line from the origin at most once, and that if it does so, it will do it from below. Thus we have

Theorem 1. *The present value function $P(\bullet)$ is IAIR (DAIR) iff for $T \geq 0$ and each $\lambda > 0$, $(P(T)-\exp(-\lambda T))$, has at most one change of sign, and if a change of sign actually occurs, it occurs from $+$ to $-$ (from $-$ to $+$).*

A formal proof of this theorem is in Barlow and Proschan (1975, p. 90). Its import is that the present value function under a monotonically increasing yield curve will cross the present value function under a constant interest rate λ—namely $\exp(-\lambda T)$—at most once, and that if it does cross it will do so from above. The reverse is true when the yield curve decreases monotonically.

Figure 2 illustrates the aforementioned crossing feature for the case of $P(\bullet)$ IAIR, showing a crossing at some time T^*. In general, T^* is unknown;

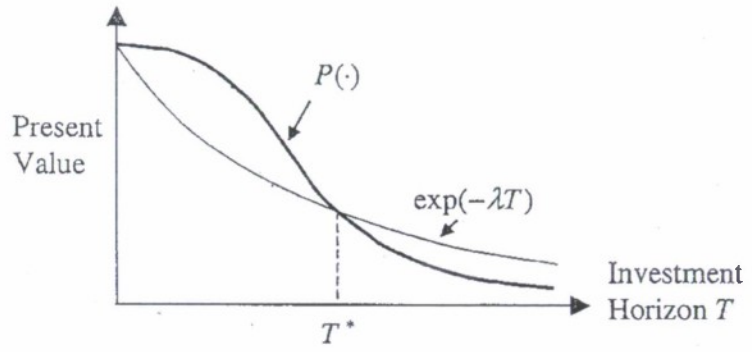

Figure 2 Crossing Properties of an IAIR Present Value Function

it will be known only when a specific functional form is assumed for $P(\bullet)$.

The essence of Figure 2 is that when the yield curve is predicted to be monotone increasing, and having a spot interest rate $\lambda > 0$ at $T = 0$, then the investment horizon should be at least T^*. Investment horizons smaller than T^* will result in smaller total yields than those greater than T^*. The investment horizon of T^* is an equilibrium point.

The illustration of Figure 2 assumes that $P(T)$ and $\exp(-\lambda T)$ cross, whereas Theorem 1 asserts that there is at most one crossing. Thus we need to explore the conditions under which a crossing necessarily occurs and the point at which the crossing occurs. That is, we need to find T^*, assuming that $T^* < \infty$. For this, we need to introduce the notion of "star-ordering".

Definition 6. Let $F(T) = 1 - P(T)$ and $G(T) = (1 - e^{-\lambda T})$, for $\lambda > 0$ and $T \geq 0$. Clearly, $F(0) \equiv G(0) = 0$.Then $F(T)$ is said to be star-ordered with respect to $G(T)$, written $F \underset{*}{<} G$, if $G^{-1}[F(T)]$is star-shaped; i.e. $G^{-1}[F(T)]/T$ is increasing in $T \geq 0$.

With the above definition in place, we have the following as a theorem. It is compiled from a collection of results on pages 107-110 in Barlow and Proschan (1975).

Theorem 2. *Let* $F \underset{*}{<} G$. *Then*

i) $P(\bullet)$ *is IAIR, and*

ii) $P(T)$ *crosses* $\exp(-\lambda T)$ *at most once, and from above, as* $T \uparrow \infty$, *for each* $\lambda > 0$. *Furthermore if* $\int_0^\infty P(u)du = 1/\lambda$, *then*

iii) *A single crossing must occur, and* T^*, *the point at which the crossing occurs is greater than* $1/\lambda$. *Finally a crossing will necessarily occur at* $T^* = 1/\lambda$, *if*

iv) $P(u)$ *is DIR and*

$$\int_0^\infty P(u)du = 1/\lambda.$$

Under *iv)* above, the interest rate is monotonically decreasing; in this case the investment horizon should be no more than T^*.

In parts *ii)* and *iv)* of Theorem 2, we have imposed the requirement that

$$\int_0^\infty P(u)du = 1/\lambda. \tag{10}$$

How must we interpret the condition of Equation (10)? To do so, we appeal to the isomorphism of Section 2. Since $P(u)$ behaves like a survival function, with $P(0) = 1$ and $P(T)$ decreasing in T, we may regard T as a random variable with distribution function $(1 - P(\bullet))$. Consequently, the left hand side of Equation (10) is the expected value of T. With this as an interpretation, we may regard the investment horizon as an unknown quantity whose distribution is prescribed by the present value function, and whose mean is $1/\lambda$.

5 Present Value Functions Under Stochastic Interest Rates

The material of Section 4 was based on the premise that whereas the spot interest rate over the holding period of a bond is unknown, its general nature—a monotonic increase or decrease—can be speculated. Such speculations may be meaningful for small investment horizons; over the long run interest rates cannot be assumed to be monotonic. In any case, the scenario of Section 4 pertains to the case of deterministic but partially specified interest rates. In this section we consider the scenario of interest rate functions that are specified up to some unknown constants, or are the realization of a stochastic process. An analogue of the above two scenarios in reliability theory is a consideration of hazard functions that are stochastic about which much has been written. A recent overview is given by Yashin and Manton (1997).

5.1 *Interest Rate Functions with Random Coefficients*

Recall [see Equation (3)] the exponentiation formula for the present value function under a specified interest rate function $r(s)$, $s \geq 0$, as

$$P(T) = \exp(-R(T)), \tag{11}$$

where $R(T)$ is the cumulative interest rate function. Suppose now that $r(s)$, $s \geq 0$ cannot be precisely specified. Then the $R(T)$ of Equation (11) becomes a random quantity. Let $\pi[R(T)]$ describe our uncertainty about $R(T)$ for any fixed $T \geq 0$. We require that $\pi(\bullet)$ be assessed and specified. Thus our attention now centers around assessing $P(T;\pi)$, the present value function when $\pi[R(T)]$ can be specified for any desired value of T. In other words, $P(T;\pi)$ refers to the fact that the present value function depends on π. In what follows we shall show that

$$P(T;\pi) = E_\pi[\exp(-R(T))], \tag{12}$$

where E_π denotes the expectation with respect to $\pi(\bullet)$. To see why, we may use a strategy used in reliability theory which which begins by noting that the right-hand side of Equation (11) can also be written as

$$\exp(-R(T)) = Pr(X \geq R(T)),$$

where X is a random variable whose distribution function is a unit exponential. Consequently when $R(T)$ is random

$$\begin{aligned} P(T;\pi) &= \int_0^\infty Pr(X \geq R(T)|R(T))\pi[R(T)]dR(T) \\ &= \int_0^\infty \exp(-R(T))\pi[R(T)]dR(T) \\ &= E_\pi[\exp(-R(T))]. \end{aligned}$$

Thus in order to obtain the present value function for any investment horizon T, when we are uncertain about interest rate function over the horizon $[0,T]$, all we need do is specify our uncertainty about the cumulative interest rate at T, via $\pi[R(T)]$. What is noteworthy here is that the functional form of $R(T)$, $T \geq 0$ does not matter. All that matters is the value of $R(T)$.

5.1.1 *Consideration of Special Cases*

As an illustration of how we may put Equation (12) to work, suppose that $r(s) = \lambda$, $s \geq 0$, but that λ is unknown. This means that at time 0^+, the spot interest rate is to take some value λ, $\lambda \geq 0$ that is unknown at time 0

when the bond is purchased and that the interest rate is to remain constant over the life of the bond.

Suppose further that our uncertainty about λ is described by a gamma distribution with scale parameter α and a shape parameter β. Then $U \stackrel{\text{def}}{=} \lambda T$ has a density at u of the form

$$\pi(u;\alpha,\beta) = \frac{\exp(-\alpha u)\alpha^\beta u^{\beta-1}}{T^\beta \Gamma(\beta)},$$

from which it follows that the present value function is

$$P(T;\alpha,\beta) = \left(\frac{\alpha}{T+\alpha}\right)^\beta, \tag{13}$$

which is of the same form as the survival function of a Pareto distribution. In reliability, such functions are a consequence of doing a Bayesian analysis of lifetimes.

The argument carries forward to a higher level of sophistication wherein one assigns a prior to the survival function itself, the classic examples being the *Dirichlet process prior* of Ferguson [cf. Ferguson, Phadia and Tiwari (1992)], the *Tailfree and Neutral to the Right Process priors* of Doksum (1974), and the *Beta process priors* of Hjort (1990). Invoking the above ideas in the context of financial risk analysis could lead to interesting possibilities.

It can be verified that the present value function of Equation (13) belongs to the DIR class of functions of Definition 11. For this class we are able to provide an upper bound on $P(T)$; see Theorem 3 below. The implication for this theorem is that for scenarios of the type considered here, short investment horizons are to be preferred over long ones.

Theorem 3. [**Barlow and Proschan [1], p.116**] *If $P(T)$ is DIR with mean μ, then*

$$P(T;\mu) \leq \begin{cases} \exp(-T/\mu), & \text{for } T \leq \mu, \\ \frac{\mu}{T}e^{-1}, & \text{for } T \geq \mu; \end{cases} \tag{14}$$

this bound is sharp.

The dark line of Figure 3 illustrates the behavior of this bound. It shows that the decay in present value for time horizons smaller than μ is greater than the decay in present value for time horizons greater than μ.

The dotted line of Figure 3 shows the behavior of the upper bound had its decay been of the form $\exp(-T/\mu)$ for all values of T. Clearly investment horizons greater than μ would not be of advantage to a holder of the bond.

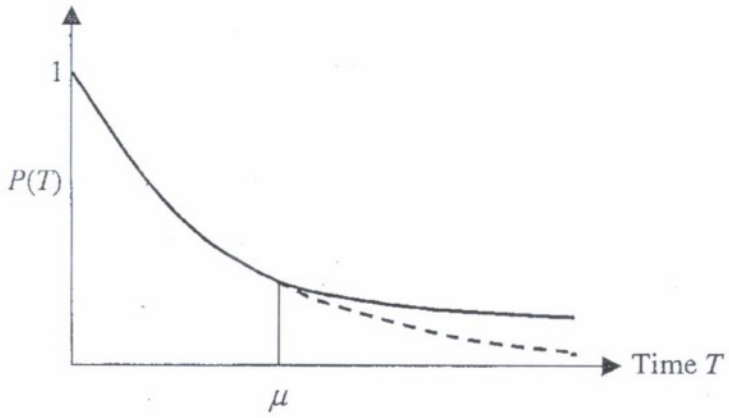

Figure 3 Upper Bound on $P(T)$ when $P(T)$ is DIR

For the special case considered here, namely λ unknown with its uncertainty described by $\pi(\lambda;\alpha,\beta)$, $P(T;\alpha,\beta) = (\alpha/(T+\alpha))^{\beta}$. Were $P(T;\alpha,\beta)$ be interpreted as a survival function, then the μ of Theorem 3 would be of the form

$$\mu = \int_0^{\infty} \left(\frac{\alpha}{T+\alpha}\right)^{\beta} dT = \frac{\alpha}{\beta-1};$$

it exists if $\beta > 1$. Consequently, under this $P(T;\alpha,\beta)$ the investment horizon should not exceed $\alpha/(\beta-1)$.

Recall that were λ to be known with certainty, $P(T)$ would be $\exp(-\lambda T)$, $\lambda > 0$, $T \geq 0$, and that there would be no restrictions on the investment horizon so that a bond holder could choose any value of T as an investment horizon. With λ unknown, the net effect is to choose shorter investment horizons, namely those that are at most $\alpha/(\beta-1)$. A similar conclusion can also be drawn in the case wherein $\pi(\lambda;\alpha,\beta)$ be a uniform over $[\alpha,\beta]$. It can be verified that in the uniform case

$$P(T;\alpha,\beta) = \frac{e^{-T\alpha} - e^{-T\beta}}{T(\beta-\alpha)},$$

and that $P(T;\alpha,\beta)$ is again DIR.

Whereas the above conclusions regarding uncertainty about $r(s)$, $s \geq 0$ causing a lowering of the investment horizon have been made based on a consideration of a special case, namely $r(s) = \lambda$, $\lambda > 0$, $s \geq 0$, the question arises about the validity of this claim, were $r(s)$ to be any other

function of s, say $r(s) = \alpha\lambda(\lambda s)^{\alpha-1}$, for some $\lambda > 0$, and $\alpha > 0$. When α is assumed known, and uncertainty about λ is described by $\pi(\lambda;\bullet)$, then Equation (12) would be a scale mixture of exponentials and by Theorem 4.7 of Barlow and Proschan (1975, p.103), it can be seen that $P(T;\bullet)$ is DIR, so that Theorem 5.1 comes into play and the inequalities of Equation (14) hold. Thus once again, uncertainty about λ causes a lowering of the investment horizon. Indeed, the essence of Theorem 3 will always hold if the cumulative interest rate $R(T)$ is such that any function of T does not entail unknown parameters.

5.2 *Interest Rates as the Realization of a Stochastic Process*

In this section we consider the case of interest rates that are the realization of a stochastic process. A consideration of stochastic processes for describing interest rate function is not new to the literature in mathematical finance. Indeed much has been written and developed therein; so much so, that some of the results can be profitably imported for use in reliability theory, wherein a consideration of stochastic failure rate functions has proven to be of value [cf. Singpurwalla (1995)]. One such example, is to describe the failure rate function by a Lévy process and to explore the hitting time of this process to a random threshold so that survival function can be introduced; the details are in Singpurwalla (2004).

The focus of this section, however, is to describe the use of a shot-noise process for modelling interest rates and to explore its consequences on the present value function. A use of the shot-noise process for describing the failure rate function has been considered by Singpurwalla and Youngreen (1993). Given below is the adaptation of this process for describing the interest rate function and some justification as to why this could be a meaningful thing to do.

We start by first noting that when the interest rate function is the realization of a stochastic process, say $\{r(s); s \geq 0\}$, then as a consequence of an argument on "randomized stopping times" by Pitman and Speed (1973), the present value function $P(T)$ is of the form $E[\exp(-R(T)]$. Here $\{R(T); T \geq 0\}$ is the *cumulative interest rate process* with $R(T) = \int_0^T r(u)du$, and as in Equation (5.2) the expectation is with respect to the distribution of $R(T)$. Clearly, an evaluation of $P(T)$ would be dependent on the ease with which $E[\exp(-R(T))]$ can be computed. With that in mind, we consider below as a special case a shot-noise process for $\{r(s); s \geq 0\}$.

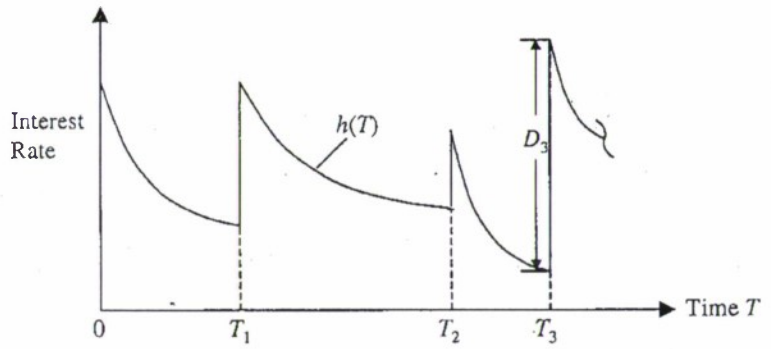

Figure 4 Sample Path of a Shot-Noise Process

5.2.1 *The Shot-Noise Process for Interest Rates*

The shot-noise process of physics is an attractive model for describing the fluctuations of the interest rate function. Our rationale for doing so is that interest rates take an upward jump when certain deleterious economic events occur. Subsequent to their upward jump, the interest rates tend to come down—or even remain constant—until the next deleterious event occurs. In Figure 4 the deleterious events are shown to occur at times $T_1, T_2, T_3, \ldots$. Such events are assumed to occur at random and are governed by say a Poisson process with rate m, $m > 0$. The amount by which the interest rate jumps upward at time T_i is supposed to be random; let this be denoted by a random variable D_i. Finally, suppose that the rate at which the interest rate decays is governed by a function, $h(s)$, $s \geq 0$; this function is called the *attenuation function*. Then, it is easy to see that for any time $T \geq 0$,

$$r(T) = \sum_{i=1}^{\infty} D_i h(T - T_i),$$

with $h(u) = 0$ whenever $u < 0$.

In what follows, we suppose that the T_i's and the D_i's are serially and contemporaneously independent. We also suppose that the D_i's are identically distributed as a random variable D.

If $D = d$, a constant, and if the attenuation function is of the form $h(u) = (1+u)^{-1}$—i.e. the interest rate decays slowly, then it can be shown that the present value function takes the form

$$P(T; m) = \exp(-mT)(1+T)^m. \qquad (15)$$

If, on the contrary, D has an exponential distribution with scale parameter b, and $h(u) = \exp(-au)$—that is, the interest rate decays exponentially, then

$$P(T; m, a, b) = \exp\left(-\frac{mbT}{1+ab}\right)\left(\frac{1+ab-\exp(-aT)}{ab}\right)^{mb/(1+ab)}. \qquad (16)$$

The $P(T)$ of Equation (15) is the survival function of a Pareto distribution. If in Equation (16) we set $a = b = 1$, and $m = 2$, then a change of time scale from T to $\exp(T)$ would result in the present value function having the form of the survival function of a beta distribution on $(0, 1)$ with parameters 1 and 2.

Thus to summarize, the consideration of a shot-noise process for the interest rate function results in some interesting forms of the present value function. A possible drawback of describing the interest rate by a shot-noise process is that except for the random times at which the interest rate shoots up by a random amount, the process is essentially deterministic.

6 Summary, Conclusions, and Future Work

Equations (13) through (16) were originally obtained in the context of reliability under dynamic environments. The isomorphism of Section 2 has enabled us to invoke them in the context of finance, and what is given in Section 5 barely scratches the surface. Much more can be done along these lines. For example, a hierarchical modelling of interest rate is one possibility. Another possibility, and one that is motivated by work of Dykstra and Laud (1981) is to describe the cumulative interest rate by a gamma process or to look at the present value functions as Dirichlet or neutral to the right processes. Another possibility, and one that is motivated by the enormous literature in survival analysis is to model interest rates as a function of covariates and markers. The Markov Additive Process of Cinlar (1972) presents an opportunity for doing the above. The purpose of this paper is mainly to open the door to other possibilities by creating a suitable platform, which we feel has been done.

But, as correctly pointed out by a referee, our discussion here has been one-sided. We have pointed out how results in reliability and survival analysis can be brought to bear on mathematical finance. It would be a folly not to acknowledge that the reverse can also be true. Indeed, this is something that has already been done by us [see Singpurwalla (2004)], where we capitalize on the several results on hitting times of stochastic processes – such as the Lévy – that can be used to generate new families of survival functions for items experiencing dynamic environments.

Acknowledgements

Research supported by Grants DAAD 19-01-1-0502 under a MURI and DAAD 19-02-01-0195, The U. S. Research Office. Some detailed comments by a referee and several conceptual issues raised by Professor Dabrowska have helped sharpen this paper. My thanks to both.

References

1. Barlow, R. E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing.* Holt, Rinehart and Winston, Inc., New York.
2. Cinlar, E. (1972). Markov additive processes. *II. Z. Wahrsch. Verw. Gebiete* **24** 94–121.
3. Doksum, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions". *Ann. Prob.* **2** 183–201.
4. Dykstra, R. L. and Laud, P. W. (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.* **9** 356–367.
5. Ferguson, T. S., Phadia, E. G. and Tiwari, R. C. (1992). Bayesian nonparametric inference. *Current Issues in Statistical Inference: Essays in Honor of D. Basu.* **17** 127–150.
6. Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18** 1259–1294.
7. Pitman, J. W. and Speed, T. P. (1973). A note on random times. *Stochastic Process. Appl.* **1** 369–374.
8. Ross, S. M. (1999). *An Introduction to Mathematical Finance.* Cambridge University Press, U. K.
9. Singpurwalla, N. D. (1995). Survival in dynamic environments. *Statist. Sci.* **10** 86–103.
10. Singpurwalla, N. D. (2004). The hazard potential of items and individuals. *Technical Report GWU/IRRA/TR-00/2.* The George Washington University.
11. Singpurwalla, N. D. and Youngreen M.A. (1993). Multivariate distributions induced by dynamic environments. *Scand. J. Statist.* **20** 251–261.
12. Yashin, A. I. and Manton, K. G. (1997). Effects of unobserved and partially observed covariate processes on system failure: a review of models and estimation strategies. *Statist. Sci.* **12** 20–34.

Chapter 6

SIGNATURE-RELATED RESULTS ON SYSTEM LIFETIMES

Henry W. Block, Michael R. Dugas, and Francisco J. Samaniego

Department of Statistics
University of Pittsburgh, Pittsburgh, PA, U.S.A.

Department of Statistics
University of California, Davis, CA, U.S.A

E-mails: hwb@stat.pitt.edu, mike@dugas.com & fjsamaniego@ucdavis.edu

The performance (lifetime, failure rate, etc.) of a coherent system in iid components is completely determined by its "signature" and the common distribution of its components. A system's signature, defined as a vector whose i^{th} element is the probability that the system fails upon the i^{th} component failure, was introduced by Samaniego (1985) as a tool for indexing systems in iid components and studying properties of their lifetimes. In this paper, several new applications of the signature concept are developed for the broad class of mixed systems, that is, for stochastic mixtures of coherent systems in iid components. Kochar, Mukerjee and Samaniego (1999) established sufficient conditions on the signatures of two competing systems for the corresponding system lifetimes to be stochastically ordered, hazard-rate ordered or likelihood-ratio ordered, respectively. Partial results are obtained on the necessity of these conditions, but all are shown not to be necessary in general. Necessary and sufficient conditions (NASCs) on signature vectors for each of the three order relations above to hold are then discussed. Examples are given showing that the NASCs can also lead to information about the precise number and locations of crossings of the systems' survival functions or failure rates in $(0, \infty)$ and about intervals over which the likelihood ratio is monotone. New results are established relating the asymptotic behavior of a system's failure rate, and the rate of convergence to zero of a system's survival function, to the signature of the system.

Key words: Survival function; Failure rate; Coherent system; Mixed system; Stochastic ordering; Hazard rate ordering; Likelihood ratio ordering; Temporal asymptotics.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 229–240

DOI: 10.1214/074921706000000473

On competing risk and degradation processes

Nozer D. Singpurwalla[1,*]

The George Washington University

Abstract: Lehmann's ideas on concepts of dependence have had a profound effect on mathematical theory of reliability. The aim of this paper is two-fold. The first is to show how the notion of a "hazard potential" can provide an explanation for the cause of dependence between life-times. The second is to propose a general framework under which two currently discussed issues in reliability and in survival analysis involving interdependent stochastic processes, can be meaningfully addressed via the notion of a *hazard potential*. The first issue pertains to the failure of an item in a dynamic setting under multiple interdependent risks. The second pertains to assessing an item's life length in the presence of observable surrogates or markers. Here again the setting is dynamic and the role of the marker is akin to that of a leading indicator in multiple time series.

1. Preamble: Impact of Lehmann's work on reliability

Erich Lehmann's work on non-parametrics has had a conceptual impact on reliability and life-testing. Here two commonly encountered themes, one of which bears his name, encapsulate the essence of the impact. These are: the notion of a ***Lehmann Alternative***, and his exposition on ***Concepts of Dependence***. The former (see Lehmann [4]) comes into play in the context of accelerated life testing, wherein a Lehmann alternative is essentially a model for accelerating failure. The latter (see Lehmann [5]) has spawned a large body of literature pertaining to the reliability of complex systems with interdependent component lifetimes. Lehmann's original ideas on characterizing the nature of dependence has helped us better articulate the effect of failures that are *causal* or *cascading*, and the consequences of lifetimes that exhibit a negative correlation. The aim of this paper is to propose a framework that has been inspired by (though not directly related to) Lehmann's work on dependence. The point of view that we adopt here is "dynamic", in the sense that what is of relevance are dependent stochastic processes. We focus on two scenarios, one pertaining to competing risks, a topic of interest in survival analysis, and the other pertaining to degradation and its markers, a topic of interest to those working in reliability. To set the stage for our development we start with an overview of the notion of a *hazard potential*, an entity which helps us better conceptualize the process of failure and the cause of interdependent lifetimes.

*Research supported by Grant DAAD 19-02-01-0195, The U. S. Army Research Office.
[1]Department of Statistics, The George Washington University, Washington, DC 20052, USA, e-mail: nozer@gwu.edu
AMS 2000 subject classifications: primary 62N05, 62M05; secondary 60J65.
Keywords and phrases: biomarkers, dynamic reliability, hazard potential, interdependence, survival analysis, inference for stochastic processes, Wiener maximum processes.

2. Introduction: The hazard potential

Let T denote the time to failure of a unit that is scheduled to operate in some specified static environment. Let $h(t)$ be the hazard rate function of the survival function of T, namely, $P(T \geq t), t \geq 0$. Let $H(t) = \int_0^t h(u)du$, be the cumulative hazard function at t; $H(t)$ is increasing in t. With $h(t), t \geq 0$ specified, it is well known that

$$\Pr(T \geq t; h(t), t \geq 0) = \exp(-H(t)).$$

Consider now an exponentially distributed random variable X, with scale parameter λ, $\lambda \geq 0$. Then for some $H(t) \geq 0$,

$$\Pr(X \geq H(t)|\lambda = 1) = \exp(-H(t));$$

thus

$$\Pr(T \geq t; h(t), t \geq 0) = \exp(-H(t)) = \Pr(X \geq H(t)|\lambda = 1). \tag{2.1}$$

The right hand side of the above equation says that the item in question will fail when its cumulative hazard $H(t)$ crosses a threshold X, where X has a unit exponential distribution. Singpurwalla [11] calls X the ***Hazard Potential*** of the item, and interprets it as an unknown resource that the item is endowed with at inception. Furthermore, $H(t)$ is interpreted as the amount of resource consumed at time t, and $h(t)$ is the rate at which that resource gets consumed. Looking at the failure process in terms of an endowed and a consumed resource enables us to characterize an environment as being *normal* when $H(t) = t$, and as being *accelerated* (*decelerated*) when $H(t) \geq (<)\ t$. More importantly, with X interpreted as an unknown resource, we are able to interpret dependent lifetimes as the consequence of dependent hazard potentials, the later being a manifestation of commonalities of design, manufacture, or genetic make-up. Thus one way to generate dependent lifetimes, say T_1 and T_2 is to start with a bivariate distribution (X_1, X_2) whose marginal distributions are exponential with scale parameter one, and which is not the product of exponential marginals. The details are in Singpurwalla [11].

When the environment is dynamic, the rate at which an item's resource gets consumed is random. Thus $h(t); t \geq 0$ is better described as a stochastic process, and consequently, so is $H(t), t \geq 0$. Since $H(t)$ is increasing in t, the ***cumulative hazard process*** $\{H(t); t \geq 0\}$ is a continuous increasing process, and the item fails when this process hits a random threshold X, the item's hazard potential. Candidate stochastic processes for $\{H(t); t \geq 0\}$ are proposed in the reference given above, and the nature of the resulting lifetimes described therein. Noteworthy are an increasing Lévy process, and the maxima of a Wiener process.

In what follows we show how the notion of a hazard potential serves as a unifying platform for describing the competing risk phenomenon and the phenomenon of failure due to ageing or degradation in the presence of a marker (or a bio marker) such as crack size (or a CD4 cell count).

3. Dependent competing risks and competing risk processes

By "competing risks" one generally means failure due to agents that presumably compete with each other for an item's lifetime. The traditional model that has been used for describing the competing risk phenomenon has been the reliability of a series system whose component lifetimes are independent or dependent. The idea

here is that since the failure of any component of the system leads to the failure of the system, the system experiences multiple risks, each risk leading to failure. Thus if T_i denotes the lifetime of component i, $i = 1, \ldots, k$, say, then the cause of system failure is that component whose lifetime is smallest of the k lifetimes. Consequently, if T denotes a system's lifetime, then

$$\Pr(T \geq t) = P(H_1(t) \leq X_1, \ldots, H_k(t) \leq X_k), \tag{3.1}$$

where X_i is the hazard potential of the i-th component, and $H_i(t)$ its cumulative hazard (or the risk to component i) at time t. If the X_i's are assumed to be independent (a simplifying assumption), then (3.1) leads to the result that

$$\Pr(T \geq t) = \exp[-(H_1(t) + \cdots + H_k(t))], \tag{3.2}$$

suggesting an additivity of cumulative hazard functions, or equivalently, an additivity of the risks. Were the X_i's assumed dependent, then the nature of their dependence will dictate the manner in which the risks combine. Thus for example if for some θ, $0 \leq \theta \leq 1$, we suppose that

$$\Pr(X_1 \geq x_1, X_2 \geq x_2|\theta) = \exp(-x_1 - x_2 - \theta x_1 x_2),$$

namely one of Gumbel's bivariate exponential distributions, then

$$\Pr(T \geq t|\theta) = \exp[-(H_1(t) + H_2(t) + \theta H_1(t) H_2(t))].$$

The cumulative hazards (or equivalently, the risks) are no longer additive.

The series system model discussed above has also been used to describe the failure of a single item that experiences several failure causing agents that compete with each other. However, we question this line of reasoning because a single item posseses only one unknown resource. Thus the $X_1, \ldots, X_k$ of the series system model should be replaced by a single X, where $X_1 = X_2 = \cdots = X_k = X$ (in probability). To set the stage for the single item case, suppose that the item experiences k agents, say $C_1, \ldots, C_k$, where an agent is seen as a cause of failure; for example, the consumption of fatty foods. Let $H_i(t)$ be the consequence of agent C_i, were C_i be the only agent acting on the item. Then under the simultaneous action by all of the k agents the item's survival function

$$\begin{aligned} &\Pr(T \geq t; h_1(t), \ldots, h_k(t)) \\ &\quad = P(H_1(t) \leq X, \ldots, H_k(t) \leq X) \\ &\quad = \exp(-\max(H_1(t), \ldots, H_k(t))). \end{aligned} \tag{3.3}$$

Here again, the cumulative hazards are not additive.

Taking a clue from the fact that dependent hazard potentials lead us to a non-additivity of the cumulative hazard functions, we observe that the condition $X_1 \overset{P}{=} X_2 \overset{P}{=} \cdots \overset{P}{=} X_k \overset{P}{=} X$ (where $X_1 \overset{P}{=} X_2$ denotes that $X_1$and X_2 are equal in probability) implies that $X_1, \ldots, X_k$ are *totally positively dependent*, in the sense of Lehmann (1966). Thus (3.2) and (3.3) can be combined to claim that in general, under the series system model for competing risks, $P(T \geq t)$ can be bounded as

$$\exp(-\sum_{1}^{k} H_i(t)) \leq P(T \geq t) \leq \exp(-\max(H_1(t), \ldots, H_k(t))). \tag{3.4}$$

Whereas (3.4) above may be known, our argument leading up to it could be new.

3.1. Competing risk processes

The prevailing view of what constitutes dependent competing risks entails a consideration of dependent component lifetimes in the series system model mentioned above. By contrast, our position on a proper framework for describing dependent competing risks is different. Since it is the $H_i(t)$'s that encapsulate the notion of risk, dependent competing risks should entail interdependence between $H_i(t)$'s, $i = 1, \ldots, k$. This would require that the $H_i(t)$'s be random, and a way to do so is to assume that each $\{H_i(t); t \geq 0\}$ is a stochastic process; we call this a ***competing risk process***. The item fails when any one of the $\{H_i(t); t \geq 0\}$ processes first hits the items hazard potential X. To incorporate interdependence between the $H_i(t)$'s, we conceptualize a k-variate process $\{H_1(t), \ldots, H_k(t); t \geq 0\}$, that we call a ***dependent competing risk process***. Since $H_i(t)$'s are increasing in t, one possible choice for each $\{H_i(t); t \geq 0\}$ could be a *Brownian Maximum Process*. That is $H_i(t) = \sup_{0<s\leq t}\{W_i(s); s \geq 0\}$, where $\{W_i(s); s \geq 0\}$ is a standard Brownian motion process. Dependence between the $H_i(t)$'s can be induced via a dependence between the $\{W_i(s); s \geq 0\}$ processes. Thus for example, in the bivariate case, if ρ denotes the correlation between two standard Brownian motion processes, then

$$\Pr(T \geq t) = \int_0^\infty P\left(H_1(t) \leq x, H_2(t) \leq x\right) e^{-x} dx$$

and it can be shown (details omitted) that,

$$\Pr(T \geq t) = \frac{\int_0^t \int_0^t \exp\left[-\frac{(a^2+b^2-2\rho ab)}{2t(1-\rho^2)}\right] dadb}{\int_0^\infty \int_0^\infty \exp\left[-\frac{(u^2+v^2-2\rho uv)}{2t(1-\rho^2)}\right] dudv}. \tag{3.5}$$

Another possibility, again for the case of $k = 2$, is to assume that $\{H_1(t); t \geq 0\}$ is some non-negative, non-decreasing, right-continuous process, but that $\{H_2(t); t \geq 0\}$ has a sample path which is an impulse function of the form $H_2(t) = 0$ for all $t < t^*$, and that $H_2(t^*) = \infty$ for some $t^* > 0$, where the rate of occurrence of the impulse at time t depends on $H_1(t)$. The process $\{H_2(t); t \geq 0\}$ can be identified with some sort of a traumatic event that competes with the process $\{H_1(t); t \geq 0\}$ for the lifetime of the item. In the absence of trauma the item fails when the process $\{H_1(t); t \geq 0\}$ hits the item's hazard potential. This scenario parallels the one considered by Lemoine and Wenocur [6], albeit in a context that is different from ours. By assuming that the probability of occurrence of an impulse in the time interval $[t, t + h)$, given that $H_1(t) = \omega$, is $1 - \exp(-\omega h)$, Lemoine and Wenocur [6] have shown that for $X = x$, the probability of survival of an item to time t is of the form:

$$\Pr(T \geq t) = E\left[\exp\left(\int_0^t H_1(s)ds\right) I_{[0,x)}\left(H_1(t)\right)\right], \tag{3.6}$$

where $I_A(\bullet)$ is the indicator of a set A, and the expectation is with respect to the distribution of the process $\{H_1(t); t \geq 0\}$. As a special case, when $\{H_1(t); t \geq 0\}$ is a gamma process (see Singpurwalla [10]), and x is infinite, so that $I_{[0,\infty)}\left(H_1(t)\right) = 1$ for $H_1(t) \geq 0$, the above equation takes the form

$$\Pr(T \geq t) = \exp(-(1+t)\log(1+t) + t). \tag{3.7}$$

The closed form result of (3.7) suffers from the disadvantage of having the effect of the hazard potential de facto nullified. The more realistic case of (3.6) will call for numerical or simulation based approaches. These remain to be done; our aim here has been to give some flavor of the possibilities.

4. Biomarkers and degradation processes

A topic of current interest in both reliability and survival analysis pertains to assessing lifetimes based on observable surrogates, such as crack length, and biomarkers like CD4 cell counts. Here again the hazard potential provides a unified perspective for looking at the interplay between the unobservable failure causing phenomenon, and an observable surrogate. It is an assumed dependence between the above two processes that makes this interplay possible.

To engineers (cf. Bogdanoff and Kozin [1]) degradation is the irreversible accumulation of damage throughout life that leads to failure. The term "damage" is not defined; however it is claimed that damage manifests itself via surrogates such as cracks, corrosion, measured wear, etc. Similarly, in the biosciences, the notion of "ageing" pertains to a unit's position in a state space wherein the probabilities of failure are greater than in a former position. Ageing manifests itself in terms of biomedical and physical difficulties experienced by individuals and other such biomarkers.

With the above as background, our proposal here is to conceptualize ageing and degradation as unobservable constructs (or latent variables) that serve to describe a process that results in failure. These constructs can be seen as the cause of observable surrogates like cracks, corrosion, and biomarkers such as CD4 cell counts. This modelling viewpoint is not in keeping with the work on degradation modelling by Doksum [3] and the several references therein. The prevailing view is that degradation is an observable phenomenon that reveals itself in the guise of crack length and CD4 cell counts. The item fails when the observable phenomenon hits some threshold whose nature is not specified. Whereas this may be meaningful in some cases, a more general view is to separate the observable and the unobservable and to attribute failure as a consequence of the behavior of the unobservable.

To mathematically describe the cause and effect phenomenon of degradation (or ageing) and the observables that it spawns, we view the (unobservable) cumulative hazard function as degradation, or ageing, and the biomarker as an observable process that is influenced by the former. The item fails when the cumulative hazard function hits the item's hazard potential X, where X has exponential (1) distribution. With the above in mind we introduce the ***degradation process*** as a bivariate stochastic process $\{H(t), Z(t), t \geq 0\}$, with $H(t)$ representing the unobservable degradation, and $Z(t)$ an observable marker. Whereas $H(t)$ is required to be non-decreasing, there is no such requirement on $Z(t)$. For the marker to be useful as a predictor of failure, it is necessary that $H(t)$ and $Z(t)$ be related to each other. One way to achieve this linkage is via a *Markov Additive Process* (cf. Cinlar [2]) wherein $\{Z(t); t \geq 0\}$ is a Markov process and $\{H(t); t \geq 0\}$ is an increasing Lévy process whose parameters depend on the state of the $\{Z(t); t \geq 0\}$ process. The ramifications of this set-up need to be explored.

Another possibility, and one that we are able to develop here in some detail (see Section 5), is to describe $\{Z(t); t \geq 0\}$ by a Wiener process (cracks do heal and CD4 cell counts do fluctuate), and the unobservable degradation process $\{H(t); t \geq 0\}$

by a ***Wiener Maximum Process***, namely,

$$H(t) = \sup_{0<s\leq t} \{Z(s); s \geq 0\}. \tag{4.1}$$

What makes the topic of analyzing degradation processes attractive is not just the modeling part; the statistical and computational issues that the set-up creates are quite challenging. Since $\{Z(t); t \geq 0\}$ is an observable process, how may one use observations on this process until some time, say t^*, to make inferences about the process of interest $H(t)$, for $t > t^*$? In other words, how does one assess $\Pr(T > t|\{Z(s); 0 < s \leq t^* < t\})$, where T is an item's time to failure? Furthermore, as is often the case, the process $\{Z(s); s \geq 0\}$ cannot be monitored continuously. Rather, what one is able to do is observe $\{Z(s); s \geq 0\}$ at k discrete time points and use these as a basis for inference about $\Pr(T > t|\{Z(s); 0 < s \leq t^* < t\})$. These and other matters are discussed next in Section 5, which could be viewed as a prototype of what else is possible using other models for degradation.

5. Inference under a Wiener maximum process for degradation

We start with some preliminaries about a Wiener process and its hitting time to a threshold. The notation used here is adopted from Doksum [3].

5.1. Hitting time of a Wiener maximum process to a random threshold

Let Z_t denote an observable marker process $\{Z(t); t \geq 0\}$, and H_t an unobservable degradation process $\{H(t); t \geq 0\}$. The relationship between these two processes is prescribed by (4.1). Suppose that Z_t is described by a Wiener process with a drift parameter η and a diffusion parameter $\sigma^2 > 0$. That is, $Z(0) = 0$ and Z_t has independent increments. Also, for any $t > 0$, $Z(t)$ has a Gaussian distribution with $E(Z(t)) = \eta t$, and for any $0 \leq t_1 < t_2$, $\mathrm{Var}[Z(t_2) - Z(t_1)] = (t_2 - t_1)\sigma^2$. Let T_x denote the first time at which Z_t crosses a threshold $x > 0$; that is, T_x is the hitting time of Z_t to x. Then, when $\eta = 0$,

$$\begin{aligned} \Pr(Z(t) \geq x) &= \Pr(Z(t) \geq x|T_x \leq t)\Pr(T_x \leq t) \\ &\quad + \Pr(Z(t) \geq x|T_x > t)\Pr(T_x > t), \end{aligned} \tag{5.1}$$

$$\Pr(T_x \leq t) = 2\Pr(Z(t) \geq x). \tag{5.2}$$

This is because $\Pr(Z(t) \geq x|T_x \leq t)$ can be set to $1/2$, and the second term on the right hand side of (5.1) is zero. When $Z(t)$ has a Gaussian distribution with mean ηt and variance $\sigma^2 t$, $\Pr(Z(t) \geq x)$ can be similarly obtained, and thence $\Pr(T_x \leq t) \stackrel{def}{=} F_x(t|\eta, \sigma)$. Specifically it can be seen that

$$F_x(t|\eta, \sigma) = \Phi\left(\frac{\sqrt{\lambda}}{\mu}\sqrt{t} - \frac{\sqrt{\lambda}}{\sqrt{t}}\right) + \Phi\left(-\frac{\sqrt{\lambda}}{\mu}\sqrt{t} - \frac{\sqrt{\lambda}}{\sqrt{t}}\right)\exp\left(\frac{2\lambda}{\mu}\right), \tag{5.3}$$

where $\mu = x/\eta$ and $\lambda = x^2/\sigma^2$. The distribution $F_x$is the ***Inverse Gaussian Distribution*** (IG-Distribution) with parameters μ and λ, where $\mu = E(T_x)$ and $\lambda\mu^2 = \mathrm{Var}(T_x)$. Observe that when $\eta = 0$, both $E(T_x)$ and $\mathrm{Var}(T_x)$ are infinite, and thus for any meaningful description of a marker process via a Wiener process, the drift parameter η needs to be greater than zero.

The probability density of F_x at t takes the form:

$$f_x(t|\eta,\sigma) = \sqrt{\frac{\lambda}{2\pi t^3}} \exp\left[-\frac{\lambda}{2\mu^2}\frac{(t-\mu)^2}{t}\right], \tag{5.4}$$

for $t, \mu, \lambda > 0$.

We now turn attention to H_t, the process of interest. We first note that because of (4.1), $H(0) = 0$, and $H(t)$ is non-decreasing in t; this is what was required of H_t. An item experiencing the process H_t fails when H_t first crosses a threshold X, where X is unknown. However, our uncertainty about X is described by an exponential distribution with probability density $f(x) = e^{-x}$. Let T denote the time to failure of the item in question. Then, following the line of reasoning leading to (5.1), we would have, in the case of $\eta = 0$,

$$\Pr(T \le t) = 2\Pr(H(t) \ge x).$$

Furthermore, because of (4.1), the hitting time of H_t to a random threshold X will coincide with T_x, the hitting time of Z_t (with $\eta > 0$) to X. Consequently,

$$\begin{aligned}\Pr(T \le t) = \Pr(T_x \le t) &= \int_0^\infty \Pr(T_x \le t | X = x) f(x) dx \\ &= \int_0^\infty \Pr(T_x \le t) e^{-x} dx = \int_0^\infty F_x(t|\eta,\sigma) e^{-x} dx.\end{aligned}$$

Rewriting $F_x(t|\eta,\sigma)$ in terms of the marker process parameters η and σ, and treating these parameters as known, we have

$$\begin{aligned}\Pr(T \le t|\eta,\sigma) &\stackrel{def}{=} F(t|\eta,\sigma) \\ &= \int_0^\infty \left[\Phi\left(\frac{\eta}{\sigma}\sqrt{t} - \frac{x}{\sigma\sqrt{t}}\right) + \Phi\left(-\frac{\eta}{\sigma}\sqrt{t} - \frac{x}{\sigma\sqrt{t}}\right)\right] \\ &\quad \times \exp\left(x\left(\frac{2\eta}{\sigma^2} - 1\right)\right) dx,\end{aligned} \tag{5.5}$$

as our assessment of an item's time to failure with η and σ assumed known. It is convenient to summarize the above development as follows

Theorem 5.1. *The time to failure T of an item experiencing failure due to ageing or degradation described by a* ***Wiener Maximum Process*** *with a drift parameter $\eta > 0$, and a diffusion parameter $\sigma^2 > 0$, has the distribution function $F(t|\eta,\sigma)$ which is a location mixture of Inverse Gaussian Distributions. This distribution function, which is also the hitting time of the process to an exponential* (1) *random threshold, is given by* (5.5).

In Figure 1 we illustrate the behavior of the IG-Distribution function $F_x(t)$, for $x = 1, 2, 3, 4$, and 5, when $\eta = \sigma = 1$, and superimpose on these a plot of $F(t|\eta = \sigma = 1)$ to show the effect of averaging the threshold x. As can be expected, averaging makes the S-shapedness of the distribution functions less pronounced.

5.2. Assessing lifetimes using surrogate (biomarker) data

The material leading up to Theorem 5.1 is based on the thesis that η and σ^2 are known. In actuality, they are of course unknown. Thus, besides the hazard potential

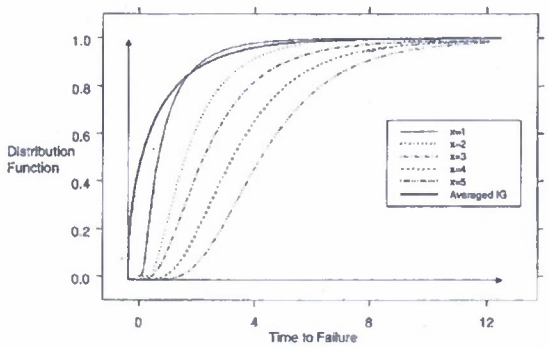

FIG 1. *The IG-Distribution with thresholds* $x = 1, \ldots, 5$ *and the averaged IG-Distribution.*

X, the η and σ^2 constitute the unknowns in our set-up. To assess η and σ^2 we may use prior information, and when available, data on the underlying processes Z_t and H_t. The prior on X is an exponential distribution with scale one, and this prior can also be updated using process data. In the remainder of this section, we focus attention on the case of a single item and describe the nature of the data that can be collected on it. We then outline an overall plan for incorporating these data into our analyses.

In Section 5.3 we give details about the inferential steps. The scenario of observing several items to failure in order to predict the lifetime of a future item will not be discussed.

In principle, we have assumed that H_t is an unobservable process. This is certainly true in our particular case when the observable marker process Z_t cannot be continuously monitored. Thus it is not possible to collect data on H_t. Contrast our scenario to that of Doksum [3], Lu and Meeker [7], and Lu, Meeker and Escobar [8], who assume that degradation is an observable process and who use data on degradation to predict an item's lifetime. We assume that it is the surrogate (or the biomarker) process Z_t that is observable, but only prior to T, the item's failure time. In some cases we may be able to observe Z_t at $t=T$, but doing so in the case of a single item would be futile, since our aim is to assess an unobserved T. Data on Z_t will certainly provide information about η and σ^2, but also about X; this is because for any $t < T$, we know that $X > Z(t)$. Thus, as claimed by Nair [9], data on (the observable surrogates of) degradation helps sharpen lifetime assessments, because a knowledge of η, σ^2 and X translates to a knowledge of T.

It is often the case – at least we assume so – that Z_t cannot be continuously monitored, so that observations on Z_t could be had only at times $0 < t_1 < t_2 < \cdots < t_k < T$, yielding $\mathbf{Z} = (Z(t_1), \ldots, Z(t_k))$ as data. Furthermore, based on $Z(t_k)$, we are able to assert that $X > Z(t_k)$. This means that our updated uncertainty about X will be encapsulated by a shifted exponential distribution with scale parameter one, and a location (or shift) parameter $Z(t_k)$.

Thus for an item experiencing failure due to degradation, whose marker process yields $\mathbf{Z}$ as data, our aim will be to assess the item's *residual life* $(T - t_k)$. That is, for any $u > 0$, we need to know $\Pr(T > t_k + u; \mathbf{Z}) = \Pr(T > t_k + u; T > t_k)$, and this under a certain assumption (cf. Singpurwalla [12]) is tantamount to knowing

$$\frac{\Pr(T > t_k + u)}{\Pr(T > t_k)}, \tag{5.6}$$

for $0 < u < \infty$. To assess the two quantities in the above ratio, we need to consider the quantity $\Pr(T > t; \mathbf{Z})$, for some $t > 0$. Let $\pi(\eta, \sigma^2, x; \mathbf{Z})$ encapsulate our uncertainty about η, σ^2 and X in the light of the data $\mathbf{Z}$. In Section 5.3 we describe our approach for assessing $\pi(\eta, \sigma^2, x; \mathbf{Z})$. Now

$$\begin{aligned}(5.7)\quad \Pr(T > t; \mathbf{Z}) &= \int_{\eta,\sigma^2,x} \Pr(T > t|\eta, \sigma^2, x; \mathbf{Z})\pi(\eta, \sigma^2, x; \mathbf{Z})(d\eta)(d\sigma^2)(dx) \\ &= \int_{\eta,\sigma^2,x} \Pr(T_x > t|\eta, \sigma^2)\pi(\eta, \sigma^2, x; \mathbf{Z})(d\eta)(d\sigma^2)(dx) \\ (5.8)\qquad &= \int_{\eta,\sigma^2,x} F_x(t|\eta, \sigma)\pi(\eta, \sigma^2, x; \mathbf{Z})(d\eta)(d\sigma^2)(dx),\end{aligned}$$

where $F_x(t|\eta, \sigma)$ is the IG-Distribution of (5.3).

Implicit to going from (5.7) to (5.8) is the assumption that the event $(T > t)$ is independent of $\mathbf{Z}$ given η, σ^2 and X. In Section 5.3 we will propose that η be allowed to vary between a and b; also, $\sigma^2 > 0$, and having observed $Z(t_k)$, it is clear that x must be greater than $Z(t_k)$. Consequently, (5.8) gets written as

$$(5.9)\qquad \Pr(T > t; \mathbf{Z}) = \int_a^b \int_0^\infty \int_{Z(t_k)}^\infty F_x(t|\eta, \sigma)\pi(\eta, \sigma^2, x; \mathbf{Z})(d\eta)(d\sigma^2)(dx),$$

and the above can be used to obtain $\Pr(T > t_k + u; \mathbf{Z})$ and $\Pr(T > t_k; \mathbf{Z})$. Once these are obtained, we are able to assess the residual life $\Pr(T > t_k + u|T > t_k)$, for $u > 0$.

We now turn our attention to describing a Bayesian approach specifying $\pi(\eta, \sigma^2, x; \mathbf{Z})$.

5.3. Assessing the posterior distribution of η, σ^2 and X

The purpose of this section is to describe an approach for assessing $\pi(\eta, \sigma^2, x; \mathbf{Z})$, the posterior distribution of the unknowns in our set-up. For this, we start by supposing that $\mathbf{Z}$ is an unknown and consider the quantity $\pi(\eta, \sigma^2, x|\mathbf{Z})$. This is done to legitimize the ensuing simplifications. By the multiplication rule, and using obvious notation

$$\pi(\eta, \sigma^2, x|\mathbf{Z}) = \pi_1(\eta, \sigma^2|X, \mathbf{Z})\pi_2(X|\mathbf{Z}).$$

It makes sense to suppose that η and σ^2 do not depend on X; thus

$$(5.10)\qquad \pi(\eta, \sigma^2, x|\mathbf{Z}) = \pi_1(\eta, \sigma^2|\mathbf{Z})\pi_2(X|\mathbf{Z}).$$

However, $\mathbf{Z}$ is an observed quantity. Thus (5.10) needs to be recast as:

$$(5.11)\qquad \pi(\eta, \sigma^2, x; \mathbf{Z}) = \pi_1(\eta, \sigma^2; \mathbf{Z})\pi_2(X; \mathbf{Z}).$$

Regarding the quantity $\pi_2(X; \mathbf{Z})$, the only information that $\mathbf{Z}$ provides about X is that $X > Z(t_k)$. Thus $\pi_2(X; \mathbf{Z})$ becomes $\pi_2(X; Z(t_k))$. We may now invoke Bayes' law on $\pi_2(X; Z(t_k))$ and using the facts that the prior on X is an exponential (1) distribution on $(0, \infty)$, obtain the result that the posterior of X is also an exponential (1) distribution, but on $(Z(t_k), \infty)$. That is, $\pi_2(X; Z(t_k))$ is a shifted exponential distribution of the form $\exp(-(x - Z(t_k)))$, for $x > Z(t_k)$.

Turning attention to the quantity $\pi_1(\eta, \sigma^2; \mathbf{Z})$ we note, invoking Bayes' law, that

$$(5.12)\qquad \pi_1(\eta, \sigma^2; \mathbf{Z}) \propto \mathcal{L}(\eta, \sigma^2; \mathbf{Z})\pi^*(\eta, \sigma^2),$$

where $\mathcal{L}(\eta, \sigma^2; \mathbf{Z})$ is the likelihood of η and σ^2 with $\mathbf{Z}$ fixed, and $\pi^*(\eta, \sigma^2)$ our prior on η and σ^2. In what follows we discuss the nature of the likelihood and the prior.

The Likelihood of η and σ^2

Let $Y_1 = Z(t_1)$, $Y_2 = (Z(t_2) - Z(t_1)), \ldots, Y_k = (Z(t_k) - Z(t_{k-1}))$, and $s_1 = t_1$, $s_2 = t_2 - t_1, \ldots, s_k = t_k - t_{k-1}$. Because the Wiener process has independent increments, the y_i's are independent. Also, $y_i \sim N(\eta s_i, \sigma^2 s_i)$, $i = 1, \ldots, k$, where $N(\mu, \xi^2)$ denotes a Gaussian distribution with mean μ and variance ξ^2. Thus, the joint density of the y_i's, $i = 1, \ldots, k$, which is useful for writing out a likelihood of η and σ^2, will be of the form

$$\prod_{i=1}^{k} \phi\left(\frac{y_i - \eta s_i}{\sigma^2 s_i}\right);$$

where ϕ denotes a standard Gaussian probability density function. As a consequence of the above, the likelihood of η and σ^2 with $\mathbf{y} = (y_1, \ldots, y_k)$ fixed, can be written as:

$$\mathcal{L}(\eta, \sigma^2; \mathbf{y}) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi s_i}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_i - \eta s_i}{\sigma^2 s_i}\right)^2\right]. \tag{5.13}$$

The Prior on η and σ^2

Turning attention to $\pi^*(\eta, \sigma^2)$, the prior on η and σ^2, it seems reasonable to suppose that η and σ^2 are not independent. It makes sense to suppose that the fluctuations of Z_t depend on the trend η. The larger the η, the bigger the σ^2, so long as there is a constraint on the value of η. If η is not constrained the marker will take negative values. Thus, we need to consider, in obvious notation

$$\pi^*(\eta, \sigma^2) = \pi^*(\sigma^2|\eta)\pi^*(\eta). \tag{5.14}$$

Since η can take values in $(0, \infty)$, and since $\eta = \tan\theta$ – see Figure 2 – θ must take values in $(0, \pi/2)$.

To impose a constraint on η, we may suppose that θ has a *translated beta* density on (a, b), where $0 < a < b < \pi/2$. That is, $\theta = a + (b - a)W$, where W has a beta distribution on $(0, 1)$. For example, a could be $\pi/8$ and b could be $3\pi/8$. Note that were θ assumed to be uniform over $(0, \pi/2)$, then η will have a density of the form $2/[\pi(1 + \eta^2)]$ – which is a *folded Cauchy*.

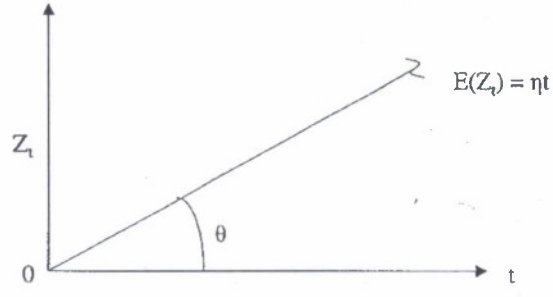

FIG 2. *Relationship between Z_t and η.*

The choice of $\pi^*(\sigma^2|\eta)$ is trickier. The usual approach in such situations is to opt for natural conjugacy. Accordingly, we suppose that $\psi \stackrel{def}{=} \sigma^2$ has the prior

$$\pi^*(\psi|\eta) \propto \psi^{-\left(\frac{\nu}{2}+1\right)} \exp\left(-\frac{\eta}{2\psi}\right), \tag{5.15}$$

where ν is a parameter of the prior.

Note that $E(\psi|\eta,\nu) = \eta/(\nu-2)$, and so $\psi = \sigma^2$ increases with η, and η is constrained over a and b. Thus a constraint on σ^2 as well.

To pin down the parameter ν, we anchor on time $t = 1$, and note that since $E(Z_1) = \eta$ and $\mathrm{Var}(Z_1) = \sigma^2 = \psi$, σ should be such that $\Delta\sigma$ should not exceed η for some $\Delta = 1, 2, 3, \ldots$; otherwise Z_1 will become negative. With $\Delta = 3$, $\eta = 3\sigma$ and so $\psi = \sigma^2 = \eta^2/9$. Thus ν should be such that $E(\sigma^2|\eta,\nu) \approx \eta^2/9$. But $E(\sigma^2|\eta,\nu) = \eta/(\nu-2)$, and therefore by setting $\eta/(\nu-2) = \eta^2/9$, we would have $\nu = 9/\eta + 2$. In general, were we to set $\eta = \Delta\sigma$, $\nu = \Delta^2/\eta + 2$, for $\Delta = 1, 2, \ldots$. Consequently, $\nu/2 + 1 = (\Delta^2/\eta + 2)/2 + 1 = \Delta^2/2\eta + 2$, and thus

$$\pi^*(\psi|\eta;\Delta) = \psi^{-\left(\frac{\Delta^2}{2\eta}+2\right)} \exp\left(-\frac{\eta}{2\psi}\right), \tag{5.16}$$

would be our prior of σ^2, conditioned on ψ, and $\Delta = 1, 2, \ldots$, serving as a prior parameter. Values of Δ can be used to explore sensitivity to the prior.

This completes our discussion on choosing priors for the parameters of a Wiener process model for Z_t. All the necessary ingredients for implementing (5.9) are now at hand. This will have to be done numerically; it does not appear to pose major obstacles. We are currently working on this matter using both simulated and real data.

6. Conclusion

Our aim here was to describe how Lehmann's original ideas on (positive) dependence framed in the context of non-parametrics have been germane to reliability and survival analysis, and even so in the context of survival dynamics. The notion of a hazard potential has been the "hook" via which we can attribute the cause of dependence, and also to develop a framework for an appreciation of competing risks and degradation. The hazard potential provides a platform through which the above can be discussed in a unified manner. Our platform pertains to the hitting times of stochastic processes to a random threshold. With degradation modeling, the unobservable cumulative hazard function is seen as the metric of degradation (as opposed to an observable, like crack growth) and when modeling competing risks, the cumulative hazard is interpreted as a risk. Our goal here was not to solve any definitive problem with real data; rather, it was to propose a way of looking at two commonly encountered problems in reliability and survival analysis, problems that have been well discussed, but which have not as yet been recognized as having a common framework. The material of Section 5 is purely illustrative; it shows what is possible when one has access to real data. We are currently persuing the details underlying the several avenues and possibilities that have been outlined here.

Acknowledgements

The author acknowledges the input of Josh Landon regarding the hitting time of a Brownian maximum process, and Bijit Roy in connection with the material of

Section 5. The idea of using Wiener Maximum Processes for the cumulative hazard was the result of a conversation with Tom Kurtz.

References

[1] Bogdanoff, J. L. and Kozin, F. (1985). *Probabilistic Models of Cumulative Damage.* John Wiley and Sons, New York.

[2] Cinlar, E. (1972). Markov additive processes. II. *Z. Wahrsch. Verw. Gebiete* **24**, 94–121. MR0329047

[3] Doksum, K. A. (1991). Degradation models for failure time and survival data. *CWI Quarterly, Amsterdam* **4**, 195–203.

[4] Lehmann, E. L. (1953). The power of rank tests. *Ann. Math. Stat.* **24**, 23–43. MR0054208

[5] Lehmann, E. L. (1966). Some concepts of dependence. *Ann. Math. Stat.* **37**, 1137–1135. MR0202228

[6] Lemoine, A. J. and Wenocur, M. L. (1989). On failure modeling. *Naval Research Logistics Quarterly* **32**, 497–508. MR0802029

[7] Lu, C. J. and Meeker, W. Q. (1993). Using degradation measures to estimate a time-to-failure distribution. *Technometrics* **35**, 161–174. MR1225093

[8] Lu, C. J., Meeker, W. Q. and Escobar, L. A. (1996). A comparison of degradation and failure-time analysis methods for estimating a time-to-failure distribution. *Statist. Sinica* **6**, 531–546. MR1410730

[9] Nair, V. N. (1988). Discussion of "Estimation of reliability in field-performance studies" by J. D. Kalbfleisch and J. F. Lawless. *Technometrics* **30**, 379–383. MR0970998

[10] Singpurwalla, N. D. (1997). Gamma processes and their generalizations: An overview. In *Engineering Probabilistic Design and Maintenance for Flood Protection* (R. Cook, M. Mendel and H. Vrijling, eds.). Kluwer Acad. Publishers, 67–73.

[11] Singpurwalla, N. D. (2005). Betting on residual life. Technical report. The George Washington University.

[12] Singpurwalla, N. D. (2006). The hazard potential: Introduction and overview. *J. Amer. Statist. Assoc.*, to appear.

(1)

Reliability and Risk

A Bayesian Perspective

Nozer D. Singpurwalla
The George Washington University, Washington DC, USA

John Wiley & Sons, Ltd

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13 978-0-470-85502-7 (HB)
ISBN-10 0-470-85502-9 (HB)

Typeset in 9.5/11.5pt Times by Integra Software Services Pvt. Ltd, Pondicherry, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Acknowledgements

Work on this book began in February 1994 at the Belagio Study and Conference Center of the Rockefeller Foundation in Belagio, Italy. The author thanks the Foundation for providing an environment conducive to jump-starting this project. Subsequent places wherein the author found habitats to work on this book have been The Santa Fe Institute, The Los Alamos National Laboratory, and the Departments of Statistics, University of Oxford, England, and the Université de Brentagc-Sud, France. The author acknowledges with thanks the hospitality provided by these institutions, orchestrated by Sallie Keller-McNulty, Mary and Dan Lunn, and Mounir Mesbah. Of course, the author's home institution, The George Washington University, warrants a special acknowledgement for nurturing his interests, and for providing an atmosphere conducive to their development. The author's deep gratitude also goes to the sponsors of his research, The Office of Naval Research and the Army Research Office, particularly the latter for its continuous sponsorship over the past several years. Much of this work is embedded in the material presented here.

Since its beginnings in 1994, this book project has gone through several changes in title and publishing house. The project was initiated by John Kimmel, and the initial encouragement and guidance came from Sir David Cox of Oxford; for this I thank him. The book project underwent several publishing house changes until Adrian Smith navigated its safe landing in the hands of Sian Jones of John Wiley, UK; thanks to both. The persistent nags of Wiley's Kathryn Sharples forced the author to accelerate the writing in earnest and bring about the book's closure. Kathryn deserves an applause.

There are others who have directly or indirectly contributed to the completion of this project that the author acknowledges. The late Professor Louis Nanni got him interested in statistics, and Professors John Kao, Richard Barlow and (the late) Frank Proschan got him interested in reliability. Professors Denis Lindley and Jay Sethuraman contributed much to the author's appreciation of probability, which is really what the book is all about. Both occupy a special place in the author's heart and mind.

The nitty-gritty aspects of this book would not have been taken care of without the tremendous and dedicated help of Josh Landon. Later on, Josh was joined by Bijit Roy, and the two being masters in the art of manipulating equations and harnessing computers, provided invaluable support toward the book's completion; thanks to both.

Last but not the least, the author singles out two members of his family for their understanding and unconditional support. His sister Khorshed bore the brunt of his domestic responsibilities in India and freed him to pursue his professional career. His wife Norah did the same here in the US and spent, without anger or resentment, many an evening and weekend in isolation when the author sequestered himself in his basement cocoon. Thank you Norah; you can freely spend the royalties – if any!

(2)

MATHEMATICAL RELIABILITY: AN EXPOSITORY PERSPECTIVE

Edited by
REFIK SOYER
Department of Management Science
The George Washington University
Washington, DC 20052

THOMAS A. MAZZUCHI
Department of Engineering Management and Systems Engineering
The George Washington University
Washington, DC 20052

NOZER D. SINGPURWALLA
Department of Statistics
The George Washington University
Washington, DC 20052

Kluwer Academic Pubtishers
Boston/Dordrecht/London

Distributors for North, Central and South America:
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA
Telephone (781) 871-6600
Fax (781) 871-6528
E-Mail <kluwer@wkap.com>

Distributors for all other countries:
Kluwer Academic Publishers Group
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS
Telephone 31 78 6576 000
Fax 31 78 6576 474
E-Mail <orderdept@wkap.nl>

 Electronic Services <http://www.wkap.nl>

Library of Congress Cataloging-in-Publication

Soyer, Refik/ Mazzuchi, Thomas A./ Singpurwalla, Nozer D.
Mathematical Reliability: An Expository Perspective
ISBN 1-4020-7697-5

Printed on acid-free paper.

Printed in the United States of America

Foreword

The entries in this volume have been categorized into seven parts, each part emphasizing a theme that in our judgment seems poised for the future development of reliability as an academic discipline with relevance. The seven parts: are *Networks and Systems*; *Recurrent Events*; *Information and Design*; *The Failure Rate Function and Burn-in*; *Software Reliability and Random Environments*; *Reliability in Composites and Orthopedics*, and *Reliability in Finance and Forensics*. Embedded within the above are some of the other currently active topics such as causality, cascading, exchangeability, expert testimony, hierarchical modeling, optimization and survival analysis. Collectively, these when linked with utility theory constitute the science base of risk analysis.

Part I on *Networks and Systems* consists of three entries each striking a unique and different chord. Boland and Samaniego introduce the notion of the "signature" of a system. The term signature (or imprint), resonates well with engineering wherein it is used to describe the characteristics of rotating machinery vis a vis its vibration. Boland and Samaniego use their notion to characterize the manner in which a system is put together, irrespective of the inherent quality of each member of the system. They make connections between their notion and the notions used in computer science. Their treatment of the topic is exhaustive; it promises to generate added interest in the notion of signatures. The second paper by Kuo and Prasad is in some sense unique among all other entries because it brings into the picture the role of optimization in reliability. Since mathematical optimization is a core discipline of operations research, Kuo and Prasad's entry is noteworthy on two counts. It exposes reliability theorists to the relevance of optimization in system design, and it makes this volume's inclusion in a series in Operations Research and Management Science germane. The third paper by Swift summarizes some of the more recent work in assessing the reliability of systems from a statistical point of view. Such work, motivated by the more recent concerns of infrastructure protection, entails aspects of hierarchical modeling, computations via the Markov chain Monte Carlo, notions of interdependence (causal and cascading failures) and the use of neural nets for reliability assessment. To whet the appetite of probability theorists, Swift caps his entry by including in it the pitfalls of not paying

attention to Borel's paradox which can arise naturally in the context of system reliability assessment.

Part II on *Recurrent Events* consists of three entries, one emphasizing an engineering scenario, another the biomedical scenario and the third a matter of foundations. Arjas and Bhattacharjee demonstrate the importance of hierarchical modeling as a way to borrow strength when dealing with heterogeneous data. They motivate their work by starting with a real life example involving valve failures in a nuclear plant and analyze the ensuing data using the Markov chain Monte Carlo approach in a Bayesian context. Their analyses show how modern statistical techniques when coupled with sophisticated computational approaches can lead to useful practical insights. The second paper by Doksum and James pertains to the use of a class of priors originally proposed by Doksum. These priors are called neutral to the right and they have gained popularity in Bayesian inference. Doksum and James do a Bayesian analysis of Barlow's total time on test transform and we are fortunate to receive this contribution. The third paper by Pena and Hollander is both archival and state-of-the-art. The authors introduce a general class of models for the treatment of recurrent event data that arises in a variety of contexts: health sciences, engineering, economics and sociology. The models are able to incorporate the effects of interventions, accumulations and concomitance. The list of references is exhaustive and the material is expository enough for any novice to benefit. It offers the Bayesians a new window of opportunity for research in an area of investigation that is very general.

Part III on *Information and Design* consists of three entries two of which share a common theme. The aim of failure data analysis, irrespective of whether the data arises from a designed life-testing experiment or retrospectively from the field, is to gain information or knowledge. The latter enables one to make meaningful predictions about future lifetimes. Discrimination, entropy, and information are the three legs on which the notion of "quantified knowledge" rests. In the first entry, Ebrahimi and Soofi provide an authoritative synopsis of the above triage with a focus on how it relates to reliability and life-testing. The entry is rich in examples and almost complete vis a vis coverage; an exception is the topic of how to design experiments for extracting the maximum amount of information that one possibly can. All the same, Ebrahimi and Soofi 's entry should motivate researchers in reliability to consider incorporating information theoretic ideas in reliability analysis; this entry provides a valuable service. The second entry by Nair, Escobar and Hamada pertains to the design of experiments for gathering performance data, with a view towards enhancing reliability. This point of view, popularized by Taguchi, advocates an active philosophy in the sense that the aim of reliability analysis should be to improve performance, not to merely report observed performance - the passive view. Notions of accelerated testing, degradation analysis, robustness and censoring are embodied in

the context of the theme of the entry. The third entry by Wilson, Reese, Hamada and Martz has a futuristic motif. It pertains to the fusion of information about lifetimes that arises from two different sources: physical experiments and computer simulations. The latter is necessitated by either cost and time constraints or by the impossibility of conducting physical tests. For example, the inability to test nuclear weapons due to test ban treaties. The authors' aim is achieved by three modern technologies: hierarchical modeling, Bayesian pooling and Markov chain Monte Carlo. The entry is both state-of-the-art and futuristic; it reinforces the idea that new research and new paradigms are often driven by new problems.

Part IV on the *Failure Rate Function and Burn-in* consists of two entries the first being a prelude to the second. The entry by Block and Savits addresses the fundamental question upon whose answer depends the need for the second entry by Jensen and Spizzichino. The notion of the failure rate function is perhaps unique to reliability and survival analysis. Indeed statistical reliability can be said to owe its existence to the notion of failure rate. Engineers often claim that components and systems exhibit a failure rate function whose shape is like that of a bath-tub. The decreasing form of the failure rate function is intriguing; specifically, is the decrease of the failure rate due to some natural phenomenon or is it the manifestation of something else, like a mixture (be it physical or be it psychological). A knowledge of the form of the failure rate function is useful for commissioning an item to service. This is the theme of the second entry by Jensen and Spizzichino. In the first entry, Block and Savits provide an overview of the various forms of the failure rate function that can occur due to mixing - irrespective of what causes the mixture. The treatment of Block and Savits tend to be mathematical (but not necessarily technical); however, their entry here is expository and relaxed. This entry embodies the view that the good mathematics of reliability theory should be driven by a genuine need. The second entry by Jensen and Spizzichino exploits the kind of results that the first entry can produce, in order to address the question of how much one should test an item (i.e. the notion of "burn-in") prior to commissioning it for use. This entry explores several ramifications of the problem and the material - which tends to be technically sophisticated - embodies the notion of utilities (via costs) - Bayesian decision making under uncertainty and sequential control theory; aspects of Operations Research and Management Science.

Part V on *Software Reliability and Random Environments* pertains to an issue that is currently important and will continue to be so. As systems become more and more software driven and software dependent, unreliable software is the critical componcnt of a system. The first entry by Chiang and Kuo uses some of the notions and ideas that are useful in reliability, to manage the software development process. This is noteworthy on two counts: the first is that it has often been claimed by experienced software engineers that it is the process that

produces a piece of software that ensures its reliability - not just the innate abilities of programmers to produce error-free codes; the second is that by using system reliability data to manage the process, Chiang and Kuo put into practice Taguchi's philosophy of reliability techniques playing an active role for producing quality software. There is a parallel between this entry and that of Jensen and Spizzichino vis a vis optimum time to "burn-in" and optimum time to release software. Hopefully, these entries will provide some synergy between the said topics. The second entry by Özekici and Soyer pertains to a generic topic in reliability - be it hardware or be it software - namely the manner in which the effects of a random environment can be treated in the context of assessing survivability. The entry, albeit focussed on the context of software, provides an overview of the several modern approaches - mostly based on stochastic process theory linked with Bayesian methodology - that are used in the context mentioned above.

Parts VI and VII pertain to some new and important avenues of application of reliability, namely, *composite materials*, *orthopedics*, *finance* and *forensics*. Of these, finance and orthopedics seem to be most intriguing, and composite materials the most crucial. Lynch and Padgett provide an overview of the recent work on the strength of fibre bundles that they have been doing over the past few years. With the increased emphasis on infrastructure protection and the use of composite materials, this type of research has an added urgency. Their entry pulls together several related topics (such as pooling failure data, interacting systems, Gaussian and inverse Gaussian processes and inferential issues) to develop a coherent package that should appeal to both engineers and statisticians. Their list of references will support this latter claim. Wilson and also Lynn introduce a new frontier for the application of reliability. The former focuses on a specific problem in orthopedics, namely the life-length of hip replacements, and uses a hierarchical approach in the context of a Bayesian analysis to assess lifetimes of such replacements. He illustrates the validity of his approach by considering actual data. This scenario further attests to the growing importance of hierarchical modeling in reliability analysis. Via this work we note the importance of the application of reliability theory to such burgeoning areas as biomedical engineering. Another area of importance for the application of reliability techniques is illustrated in Lynn's entry which is more on the conceptual front than on the practice front. He introduces notions in fixed income instruments - like bonds - and discusses their risk of default (i.e. failure). He then points out opportunities wherein notions of reliability and risk could come into play and discusses some possibilities. He then moves to the notion of "derivatives" and again points out scenarios wherein there could be an interplay between reliability and finance. Lynn's entry is important because it opens a new window of opportunity for the techniques of reliability. The final entry is on warranties. These bring into play the various notions of

probability (objective, logical and personal), utility and game theory, failure models indexed by multiple scales, and forecasting using leading indicators. Violations of warranty are often the cause of litigation - sometimes in millions of dollars - and the role of reliability analyst as an expert witness becomes central. Thus the label reliability in forensics.

Despite the broad coverage that we have endeavored to encompass, we are aware of the fact that there may be other topics that should have been included. In excluding these we take the blame; but then we are also quite delighted with what we have included.

Finally, we would like to take this opportunity to acknowledge the several years of support provided by The Army Research Office, and the Office of Naval Research, for sustaining our work in reliability through the George Washington University's *Institute for Reliability and Risk Analysis*.

The Mathematics of Risk and Reliability: A Select History

risk0485

Nozer D. Singpurwalla
Department of Statistics
The George Washington University
Washington, DC 20052, USA
nozer@gwu.edu

Simon P. Wilson
Department of Statistics
Trinity College Dublin
Dublin 2, Ireland
simon.wilson@tcd.ie
Phone: +353 1 896 1759; Fax: +353 1 677 0711

January 2007

Abstract

This article is a brief description of some landmark advances in the mathematics of risk and reliability, starting with the initial developments of probability theory in the 17th century to the ascendancy of reliability theory during the last 60 years.

Files associated with this contribution:

- history_of_reliability_corrected_.pdf: this document.
- history_of_reliability_corrected_.tex: LaTeX source file (text).
- history_of_reliability_corrected_.bbl: bibliography used by .tex file.

Keywords: decision theory, insurance, subjective probability, risk, reliability, utility

1 PREAMBLE

Writing the history about any topic is both challenging and demanding. Demanding because one needs to acquire a broad perspective about the topic, a perspective that generally comes over time and experience. The challenge of writing history comes from the matter of what to include and what to omit. There is the social danger of offending those readers who feel that their work should have been mentioned but was not. But the moral obligation of excluding the works of those who are no more with us is much greater. Writers of history must therefore confront the challenge and draw a delicate line. This task is made easier with the passage of time, because the true impact of a signal contribution is felt only after time has elapsed. By contrast, the impact of work that is incremental or marginal can be judged immediately. It is with the above in mind that the history that follows is crafted. The word "select" in the title of this contribution is deliberate; it reflects the judgement of the authors. Hopefully, the delicate line mentioned before, has been drawn by us in a just and honourable manner. All the same, our apologies to those who may feel otherwise, or whose works we have accidentally overlooked.

2 INTRODUCTION

From a layperson's point of view, a viewpoint that predates history, the term "risk" connotes the possibility that an undesirable outcome will occur. However, the modern technical meaning of the term risk is different. Here, *risk* is the sum of the product of the *probabilities* of all possible outcomes of an action and the *utilities* (or consequences) of each outcome. Utilities are numerical values of consequences on a zero to one scale. Indeed, utilities are probabilities and obey the rules of probability (Lindley, 1985, page 56). They encapsulate one's preferences between consequences. Thus the notion of risk entails the twin notions of probability and utility. Some adverse outcomes are caused by the failure or the malfunc-

tioning of certain entities, biological or physical. For such adverse outcomes, the probability of failure of the entity in question is known as the entity's unreliability; its *reliability* is the probability of non-failure for a specified period of time. In the biomedical contexts, wherein the entity is a biological unit, the term *survivability* is used instead of reliability. Thus assessing reliability (or survivability) is de facto assessing a probability, and *reliability theory* pertains to the methods and techniques for doing such assessments. The linkage between reliability and risk is relatively new (Singpurwalla, 2006). It is brought about by the point of view that the main purpose of doing a reliability analysis is to make sound decisions about preventing failure in the face of uncertainty. To the best of our knowledge, the first document that articulates this position is Barlow et al. (1993). Thus we see that probability, utility, risk, reliability and decision making are linked, with probability playing a central role, indeed the role of a germinator. Our history of risk and reliability must therefore start with a history of probability. Probability is a way to quantify uncertainty. Its origins date back to 16th century Europe and discussions about its meaning and interpretation continue until the present day. For a perspective on these, the review articles by Kolmogorov (1969) and Good (1990) are valuable. The former wholeheartedly subscribes to probability as an objective *chance*, and the latter makes the point that probability and chance are distinct concepts. The founding fathers of probability were not motivated by the need to quantify uncertainty; they were more concerned with action than with interpretation. This enables us to divide the history of probability into three parts: until 1750, 1750–1900, and from 1900. These reflect, in our opinion, three reasonably well-defined periods of development of the mathematics of uncertainty which we label: foundations, maturation and expansion of applicability. Some excellent books on the history of probability are by Hald (1990b,a), Stigler (1990) and von Plato (1994). Since the history of probability is the background for the history of risk and reliability, a reading of these and the exhaustive references therein should provide risk and reliability analysts a deeper appreciation of the foundations of their

subject.

3 TO 1750: THE FOUNDATIONS OF PROBABILITY

Insurance was the first place where the traditional notion of risk had to be quantified. Its use can be traced back 4 millenia to ancient China and Babylonia, where traders took on the risks of the caravan trade by taking out loans that were repaid if the goods arrived. The ancient Greeks and Phoenicians used marine insurance, while the Romans had a form of life insurance that paid for the funeral expenses of the holder. However there is no evidence that insurance was a common practice and indeed it disappeared with the fall of the Roman Empire. It took the growth of towns and trade in Renaissance Europe, where risks such as shipwreck, losses from fire and even kidnap ransom worried the wealthy, for insurance to develop once again. But it was the development of probability in the 17th century that finally saw the foundation for the mathematics of risk, and where our brief history can really begin. We should mention first that the mathematisation of uncertainty can be traced back to Gioralimo Kardano (1501–1575). But it was the short correspondence between Pierre de Fermat (1608–1672) and Blaise Pascal (1623–1662) that began the development of modern probability theory. Their correspondence concerned a gambling question called "The Problem of Points", which is to determine the fair bet for a game of chance where each player has an equal chance of winning, and the bet is won as soon as either player wins the game a pre-determined number of times. The difficulty arises if the number of games to win is different for each player; Fermat's and Pascal's correspondence led to a solution. Meanwhile, a contemporary of both, Christiaan Huygens (1629–1695), was one of the earliest scientists to think mathematically about risk. He was motivated by problems in annuities, which at that time were common means for states and towns to borrow money.

of 100 ['quick conceptions']	
there dies within the first six years	36
The next ten years, or *Decad*	24
The second Decad	15
The third Decad	9
The fourth	6
The next	4
The next	3
The next	2
The next	1
["perhaps but one surviveth 76"]	

Table 1: Reproduction of the table that appears in Graunt (1662).

Huygens wrote up the solution of Fermat and Pascal, and is thus credited with publishing the first book on probability theory (Huygens, 1657). Without the benefit of Fermat's and Pascal's theory, Graunt produced the first mortality table by decade (Graunt, 1662), from which he concluded that only 1% of the population survived to 76 years. Table 1 shows this brilliant if unsophisticated effort; see Seal (1980) for a discussion of its use. Graunt's work happened at the time when property insurance as we know it today began. Following the Great Fire of London in 1666, which destroyed about 13,000 houses, Nicholas Barbon opened an office to insure buildings. In 1680, he established England's first fire insurance company, "The Fire Office," to insure brick and frame homes; this also included the first fire brigade. Edmond Halley constructed the first proper mortality table, based on the statistical laws of mortality and compound interest (Halley, 1693). The table was corrected by Joseph Dodson in 1756 and made it possible to scale the premium rate to age; previously the rate had been the same for all ages. The idea of a fair price was linked to probability by Jacob Bernoulli (1654–1705), work that was published posthumously by his nephew Nicholas (Bernoulli, 1713). This work is important because it was the first substantial treatment of

probability, and contained the general theory of permutations and combinations, the weak law of large numbers as well as the binomial theorem. What interested Bernoulli was to apply the Fermat-Pascal idea of a fair bet to other problems where the idea of probability had meaning. He argued that opinions about any event occurring or not were analogous to a game of chance where betting on a certain outcome led to a fair bet. The fair bet then represents the certainty that one attaches to an event occurring. This analogy between games of chance and one's opinions also appears to have been made at the time of Fermat and Pascal (Arnauld and Nicole, 1662). The law of large numbers was particularly important for this argument because Bernoulli realized that, in practical problems, fair prices could not be deduced exactly and approximations would have to be found. This allowed him to justify approximating the probability of an event by its relative frequency. Thus in Bernoulli's ideas we see parts of the two currently dominant interpretations of probability: subjective degree of belief and relative frequency. The relative frequency idea was further developed by de Moivre (1718), who proposed the ideas of independent events, the summation rule, the multiplication rule and the central limit theorem. This connection between fair prices and probability is the basis for insurance pricing. Bernoulli's and de Moivre's work came during a period of rapid development of the insurance market, spurred on by the growth of maritime commerce in the 17th and 18th centuries. We have seen that fire insurance had been available since the Great Fire of London, but up to the 18th century, most insurance was underwritten by individual investors who stated how much of the loss risk they were prepared to accept. This concept continues to this day in Lloyd's of London, beginning in Edward Lloyd's coffechouse around 1688 in Tower Street, London, which was a popular meeting place for the shipping community to discuss insurance deals among themselves. Soon after the publication of Bernoulli's work, corporations began to engage in insurance. They were first chartered in England in 1720, and in 1735, the first insurance company in the American colonies was founded at Charleston, S.C. So, by 1750 all the basic ideas of

probability necessary for quantifying risk — probability distributions, expected values and the idea of fair price, and mortality — were in place, and were in use in insurance.

4 1750–1900: PROBABILITY MATURES

Post 1750, the first notable name is that of Thomas Bayes (1702–1761) and his famous essay on inverse probability (Bayes, 1764). His main contribution was to articulate on the multiplication rule that allows conditional probabilities.to be computed from unconditional ones; vitally, this permitted Laplace (1749–1827) to derive the law of total probability and Bayes' law. In contrast to de Moivre, Laplace thought that probability was a rational belief and the rules of probability and expectation followed naturally from this interpretation (Laplace, 1812, 1814). Poisson (1741–1840) did much work on the technical and practical aspects of probability, and greatly expanded the scope and applications of probability. His main contribution was a generalization of Bernoulli's theorem; his seminal work (Poisson, 1837) also introduced the Poisson distribution. While Poisson agreed with Laplace's rational belief interpretation of probability, criticisms of this view were raised as we move to the second half of the 19th century. John Venn (1834–1923) revived the frequency interpretation of probability, hinted at by Bernoulli, but taken further to state that frequency was the starting point for defining probability (Venn, 1866). We note little attempt so far to quantify the consequences of adverse events through utility and hence to manage risks in a coherent manner. However, we note two developments. First, the idea of utility did arise through Daniel Bernoulli in 1738 and utilitarian philosophers such as Bentham (1748–1832). They proposed rules of rationality that stated individuals desire things that maximise their utility, where positive utility is defined as the tendency to bring pleasure, and negative utility is defined as the tendency to bring pain (Bentham, 1781). Second, the industrial revolution meant that manufacturing and transport carried far graver risks than before, and we do

see the first attempts at risk management through regulation. In the United Kingdom, the Factory Act of 1802 (known as the "Health and Morals of Apprentices Act") started a sequence of such acts that attempted to improve health and safety at work. Following a rail accident that killed 88 people in Armagh, Northern Ireland, the Regulation of Railways Act 1889 made fail-safe brakes mandatory, as well as block signalling. All the main areas of insurance — life, marine and fire insurance — continued to grow throughout this period,. After 1840, with the decline of religious prejudice against the practice, life insurance entered a boom period in the United States. Many friendly or benefit societies were founded to insure the life and health of their members. The close of the 19th century finally allows us to say something about mathematical reliability theory; Pearson (1895) names the *exponential distribution* for the first time.

5 FROM 1900 TO THE PRESENT: UTILITY AND RELIABILITY ENTER

The first half of the twentieth century saw the beginning of the modern era of probability; Kolmogorov (1903–1987) axiomized probability and in doing so freed it from the confusions of interpretation (Kolmogorov, 1956). It also saw many developments in the frequency interpretation of probability, and several advances in subjective probability. Von Mises (1883–1953) wrote a paper extolling the virtues of the frequentist interpretation of probability (von Mises, 1919). Together with the work of Karl Pearson (1857–1936) and Fisher (1890–1962), methods of inference under the frequency interpretation of probability became the dominant approaches to data analysis and prediction. However, at about the same time there were breakthrough developments in the subjective approach to statistical inference and decision making. Noteworthy among these were the work of Ramsey (1931) who proposed that subjective belief and utility are the basis of decision making and the non-separability of

probability from utility. Jeffreys' (1891–1986) highly influential book on probability theory combined the logical basis of probability with the use of Bayes' Law as the basis of statistical inference (Jeffreys, 1939). At about this time, de Finetti (1906–1985), unaware of Ramsey's work, adopted the latter's subjectivistic views to produce his seminal work of probability (de Finetti, 1937), later translated into English (de Finetti, 1974). De Finetti is best remembered for the above writings, and his bold statement that "*Probability Does Not Exist*"! The period 1900–1950 also saw the laying of the foundations of modern utility theory, from which a prescription for normative decision making comes about. The mathematical basis of today's quantitative risk analysis is indeed normative decision theory. Impetus for a formal approach to utility came from von Neumann and Morgenstern (1944) with its interest in rational choice, game theory, and the modelling of preferences. This was brought to its definitive conclusion by Savage (1954), who proposed a system of axioms that linked together the ideas of Ramsey, de Finetti, and von Neumann and Morgenstern. Readable accounts of Savage' brilliant work are in DeGroot (1970) and Lindley (1985), two highly influential voices in the Bayesian approach to statistical inference and decision making. Not to be overlooked is the 1950 treatise of Wald (1902–1950) whose approach to statistical inference was decision theoretic. However, unlike that of Savage, Wald's work did not entail the use of subjective prior probabilities on the states of nature. Hardly mentioned up to now is the mathematical and the statistical theory of reliability. This is because it is only in the 1950's and the 1960's that reliability emerged as a distinct field of study. The initial impetus of this field was driven by the demands of the then newer technologies in aviation, electronics, space, and strategic weaponry. Some of the landmark events of this period are: Weibull's (1887–1961) advocacy of the Weibull distribution for metallurgical failure (Weibull, 1939, 1951), the statistical analysis of failure data by Davis (1952), the proposal of Epstein and Sobel (1953) that the exponential distribution be used as a basic tool for reliability analysis, the work of Grenander (1956) on estimating the failure rate function and the book of Gumbel (1958) on

the application of the theory of extreme values for describing failures caused by extremal phenomena such as crack lengths, floods, hurricanes, etc., the approach of Kaplan and Meier (1958) for estimating the survival function under censoring and the introduction in Watson and Wells (1961) of the notion of burn-in. Some, though not all, of this work was described in what we consider to be the very first few books on reliability; Bazovsky (1961), Lloyd and Lipow (1962), and Zelen (1963). Initially the statistical community was slow to embrace the Weibull distribution as a model for describing random failures; indeed the Journal of the American Statistical Association rejected Weibull's 1951 paper. This is despite the fact that the Weibull distribution is a member of the family of extremal distributions (Gnedenko, 1943). Subsequently, however, the popularity of the Weibull grew because of the papers of Lieblien and Zelen (1956), Kao (1958, 1959), and later the inferential work of Mann (1967, 1968, 1969). Today, along with the Gaussian and the exponential distributions, the Weibull is one of the most commonly discussed distributions in statistics. Whereas the emphasis of the works mentioned above has been to the statistical analysis of lifetime data, progress in the mathematical and probabilistic aspects was also made during the 1950's and 1960's. A landmark event is Drenick (1960) on the failure characteristics of a complex system with the replacement of failed units. It started a line of research in reliability that focused on the probabilistic aspects of components and systems; in a similar vein is a book by Cox (1962). The next major milepost was the paper by Birnbaum et al. (1961) on the structural representation of systems of components; inspiration for this work can be traced to the classic paper of Moore and Shannon (1956) on reliable relays. This was followed by the paper of Barlow et al. (1963) on monotone hazard rates. This work was highly influential in the sense that it spawned a generation of researchers who explored the probabilistic and statistical aspects of monotonicity from different perspectives. Much of this work is summarised in the two books of Barlow and Proschan (1965, 1975). There were other notable developments during the late 1960's and mid 1970's, some on the probabilistic aspects, and the others on

the statistical aspects. With regards to the former, Marshall and Olkin (1967) proposed a multivariate distribution with exponential marginals for describing dependent lifetimes. The noteworthy features of this work are that the distribution was motivated using arguments that are physically plausible, and that its properties bring out some subtle aspects of probability models. At about the same time, Esary et al. (1967) proposed a notion of dependence that they called *association*. This notion was motivated by problems of system reliability assessment, and the generality of the idea was powerful enough to attract the attention of mathematical statisticians and probabilists to develop it further. During this period, and perhaps earlier than that, there was important work in reliability also done in the Soviet Union. Indeed, Kolmogorov (1969) in his expository papers on statistics, often used examples from reliability and lifelength studies to motivate his material. The book by Gnedenko et al. (1969), and the more recent review by Ushakov (2000), gives a perspective on the Soviet work in reliability. Some other developments in that period were the papers of Cox (1972), and of Esary et al. (1973) and the book by Mann et al. (1974). Cox's highly influential paper provided a means for relating the failure rate with covariates. A similar strategy was used in Singpurwalla (1971), in the context of *accelerated testing*. The paper by Esary et al. on shock models and wear processes was remarkable in two respects. The first is that it addressed a phenomenon of much interest to engineers and produced some elegant results. Second, it paved the way for using stochastic processes to obtain probability models of failure (Singpurwalla, 1995). The book by Mann et al. integrated the probabilistic and statistical techniques used in reliability that were prevalent at that time, and by doing so it created a template for the subsequent books that followed. The book was also the first of its kind to make a case for using Bayesian methods for reliability assessment. Subsequent to the mid 1970's interest in reliability as an academic discipline took a leap and several books and papers began to appear, and are continuing to appear today. Notable among the former are the books by: Lawless (1982), Martz and Waller (1982), Nelson (1982, 1990), Gertsbakh (1989),

Crowder et al. (1991), Meeker and Escobar (1998), Aven and Jensen (1999), Singpurwalla and Wilson (1999), Hoyland and Rousand (2004) and Saunders (2006). With the exception of Martz and Waller (1982) and Singpurwalla and Wilson (1999), the statistical paradigm guiding the material in the above books has been sample theoretic (i.e. non-Bayesian). In terms of signal developments during the period, two notable ones seem to be Natvig (1982) suggestion to consider multi-state systems, and the consideration of subjective Bayesianism in reliability. The latter was triggered by Barlow's interpretation of decreasing failure rates caused by subjective mixing (Barlow, 1985), and brought to its conclusion by Gurland and Sethuraman (1995); also see the discussion in Lynn and Singpurwalla (1997) of Block and Savits (1997). The book by Spizzichino (2001) is an authoritative treatment of the generation of subjective probability models for lifetimes based on *exchangeability*. Some other developments in reliability have come about from the biostatistical perspective of survival analysis. Notable among these are Ferguson (1973) and its advocacy of the *Dirichlet process* for survival analysis, and Aalen (1978) and its point process perspective and the martingale approach to modelling lifetimes. The former has been exploited by Sethuraman (1994), and the latter by Pena and Hollander (Pena and Hollander) and Hollander and Pena (2004) in a varicty of contexts that are germane to reliability. To conclude, the last sixty years have seen two trends in risk. First of all, the idea of risk has spread to many other fields outside the traditional areas of insurance and actuarial science. It is now an important idea in medicine, public health, law, science and engineering. Secondly, driven by its increasing use and by the growth of computing and data collecting power, increasingly complex quantifications of risk and reliability have been made to make better use of increasing quantities of data; reliability and risk models, inference and prediction with those models, and numerical methods have all advanced enormously. Since the 1960's in particular, the literature on reliability, risk and survival analysis has grown in journals that cover statistics, philosophy, medicine, engineering, law, finance, environment and public policy. Annual conferences on risk in all

these subject areas have been held for the last 30 years. To these two trends we might add that the magnitude of the risks being quantified and managed has increased over the last century; environmental pollution, intensive food production and the nuclear industry being examples. The same trends in reliability theory can be discerned as those in risk: the spread of application into new fields and the impact of increasing computing power and availability of data. It is worth comparing seminal books on statistical reliability of the 1960's such as Bazovsky (1961) and Barlow and Proschan (1965) with that of the current decade (Singpurwalla, 2006) to see how much the field has changed. The debate over the interpretation of probability, and uncertainty quantification more generally, continues. The important work of Savage (1954), DeGroot (1970) and de Finetti (1974) publicized the justifications for the laws of probability through their interpretation as a subjective degree of belief. This, along with the practical development of the necessary numerical tools, has increased the use of subjective probability and Bayesian inference in the last 30 years. The strong link between risk, reliability, and the mathematical tools of probability and decision making, that has existed for 400 years, looks set to continue.

Acknowledgements

The work of Nozer D. Singpurwalla was supported by The Office of Naval Research Grant N00014-06-1-037 b and by The Army Research Office Grant W911NF-05-1-2009.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics 6*, 701–726.

Arnauld, A. and P. Nicole (1662). *L'art de penser.* Paris.

Aven, T. and U. Jensen (1999). *Stochastic models in reliability series: stochastic modeling and applied probability*, Volume 41. New York: Springer.

Barlow, R. and F. Proschan (1965). *Mathematical theory of reliability.* New York: Wiley.

Barlow, R. and F. Proschan (1975). *Statistical theory of reliability and life testing* (First ed.). New York: Holt, Rinehart and Winston, Inc.

Barlow, R. E. (1985). A Bayes explanation of an apparent failure rate paradox. *IEEE Transactions on Reliability 34*, 107–108.

Barlow, R. E., C. A. Clarotti, and F. Spizzichino (1993). *Reliability and decision making.* London: Chapman and Hall.

Barlow, R. E., A. W. Marshall, and F. Proschan (1963). Properties of probability distributions with monotone hazard rate. *Ann. Math. Stat. 34*, 375–389.

Barnard, G. A. (1958). Thomas Bayes' essay towards solving a problem in the doctrine of chances. *Biometrika 45*, 293–315.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc. London 53*, 370–418. Reprinted in 1958; see Barnard (1958).

Bazovsky, I. (1961). *Reliability theory and practice.* Englewood Cliffs, N.J.: Prentice-Hall.

Bentham, J. (1781). *An introduction to the principles of morals and legislation.* Latest edition published by Adamant Media Corporation, 2005.

Bernoulli, J. (1713). *Ars conjectandi.* Basileæ: Thurnisiorum, Fratrum.

Birnbaum, Z. W., J. D. Esary, and S. C. Saunders (1961). Multi-component systems and structures and their reliability. *Technometrics 3*, 55–77.

Block, H. W. and T. H. Savits (1997). Burn-in. *Statistical Science 12*, 1–13.

Cox, D. R. (1962). *Renewal theory.* London: Methuen.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B 34*, 187–220.

Crowder, M. J., A. C. Kimber, R. L. Smith, and T. J. Sweeting (1991). *Statistical analysis of reliability data.* London: Chapman and Hall.

Davis, D. J. (1952). An analysis of some failure data. *J. Amer. Statist. Assoc. 47*, 113–150.

de Finetti, B. (1937). Calcolo delle probabilità. In *Atti dell'XI Convegno, Torino-Aosta*, Associazone per la Mathematica Applicata alle Scienze Economiche Sociali. Typescript for the academic year 1937–38, University of Padua.

de Finetti, B. (1974). *Theory of probability.* New York: Wiley. 2 volumes.

de Moivre, A. (1718). *The Doctrine of Chances.* London.

DeGroot, M. H. (1970). *Optimal statistical decisions.* New York: McGraw-Hill.

Drenick, R. F. (1960). The failure law of complex equipment. *Journal of the Society for Industrial and Applied Mathematics 8*, 680–690.

Epstein, B. and M. Sobel (1953). Life testing. *J. Amer. Statist. Assoc. 48*, 486–502.

Esary, J. D., A. W. Marshall, and F. Proschan (1973). Shock models and wear processes. *Annals of Probability 1*, 627–649.

Esary, J. D., F. Proschan, and D. W. Walkup (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics 38*, 1466–1474.

Ferguson, T. S. (1973). A Bayesian analysis of some non-parametric problems. *Ann. Stat. 1*, 209–230.

Gertsbakh, I. B. (1989). *Statistical reliability theory.* New York: Marcel Dekker, Inc.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math. 44*, 423–453.

Gnedenko, B. V., Y. K. Belyaev, and A. D. Soloyev (1969). *Mathematical models of reliability theory.* New York: Academic Press.

Good, I. J. (1990). Subjective probability. In J. Eatwell, M. Milgate, and P. Newman (Eds.), *The New Palgrave: Utility and Probability.* New York: W. W. Norton and Co.

Graunt, J. (1662). *Natural and political observations...made upon the bills of mortality.* London: T. Roycroft. Reproduced in a more modern format in *Journal of the Institute of Actuaries*, vol. 90, no. 1.

Grenander, U. (1956). On the theory of mortality measurement, I and II. *Skandinavisk Actuarietidskrift 39*, 70–96, 125–153.

Gumbel, E. J. (1958). *Statistics of extremes.* New York: Columbia University Press.

Gurland, J. and J. Sethuraman (1995). How pooling data may reverse increasing failure rates. *Journal of the American Statistical Association 90*, 1416–1423.

Hald, A. (1990a). *A history of mathematical statistics from 1750 to 1930.* New York: Wiley.

Hald, A. (1990b). *A history of probability and statistics and theory applications before 1750.* New York: Wiley.

Halley, E. (1693). An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the City of Breslaw; with an attempt to ascertain the

price of annuities uppon lives. *Phil. Trans. Roy. Soc. 17*, 596–610. Reproduccd in *J. Inst. Actuaries*, 112, 278–301.

Hollander, M. and E. A. Pena (2004). Nonparametric methods in reliability. *Statistical Science 19*, 644–651.

Hoyland, A. and M. Rousand (2004). *System reliability theory: models and statistical methods*. New York: Wiley.

Huygens, C. (1657). Tractatus, de ratiociniis in aleæ ludo. In F. Schooten (Ed.), *Excercitationum Mathematicarum*. Lugd Batav.: Libri Quinique. J. Elsevir.

Jeffreys, H. (1939). *Theory of probability*. Oxford: Oxford University Press.

Kao, J. H. K. (1958). Computer methods for estimating Weibull parameters in reliability studies. *Trans. IRE – Reliability Quality Control 13*, 15–22.

Kao, J. H. K. (1959). A graphical estimation of mixed Weibull parameters in life testing electron tube. *Technometrics 1*, 389–407.

Kaplan, E. L. and P. Meier (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association 53*, 457–481.

Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (Second ed.). New York: Chelsea Publishing Company. Translation edited by Nathan Morrison.

Kolmogorov, A. N. (1969). *The theory of probability in mathematics, its content, methods and meaning*, Volume 2, part 3. Cambrdige, Mass.: MIT Press.

Laplace, P.-S. (1812). *Théorie analytique des probabilitiés*. Paris. Reprinted in *Oeuvres*, 10: 295–338, 1894.

Laplace, P.-S. (1814). *Essai philosophique sur les probabilités.* Paris. 6th edition translated by F. W. Truscott and F. L. Emory as *A philosophical essay on probabilities*, 1902.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data.* New York: Wiley.

Lieblien, J. and M. Zelen (1956). Statistical investigation of the fatigue life of deep-groove ball bearings. *J. Res. Nat. Bur. Standards 57*, 273–316.

Lindley, D. V. (1985). *Making decisions* (Second ed.). London: Wiley.

Lloyd, D. K. and M. Lipow (1962). *Reliability: management, methods and mathematics.* Englewood Cliffs, N.J.: Prentice-Hall.

Lynn, N. J. and N. D. Singpurwalla (1997). Burn-in makes us feel good. *Statistical Science 12*, 13–19.

Mann, N. R. (1967). Tables for obtaining best linear invariant estimates of parameters of Weibull distribution. *Technometrics 9*, 629.

Mann, N. R. (1968). Point and interval estimation procedures for 2-parameter Weibull and extreme-value distributions. *Technometrics 10*, 231.

Mann, N. R. (1969). Optimum estimators for linear functions of location and scale parameters. *Annals of Mathematical Statistics 6*, 2149–2155.

Mann, N. R., R. E. Schafer, and N. D. Singpurwalla (1974). *Methods for statistical analysis of reliability and life data.* New York: Wiley.

Marshall, A. W. and I. Olkin (1967). A multivariate exponential distribution. *Journal of the American Statistical Association 62*, 30–44.

Martz, H. F. and R. A. Waller (1982). *Bayesian reliability analysis.* New York: Wiley.

Meeker, W. Q. and L. A. Escobar (1998). *Statistical methods for reliability data.* New York: Wiley.

Moore, E. F. and C. Shannon (1956). Reliable circuits using less reliable relays I. *J. Franklin Inst. 262*, 191–208.

Natvig, B. (1982). Two suggestions of how to define a multistate coherent system. *Adv. Appl. Prob. 14*, 434–455.

Nelson, W. (1982). *Applied life data analysis.* New York: Wiley.

Nelson, W. (1990). *Accelerated testing.* New York: Wiley.

Pearson, K. (1895). Contributions to the mathematical theory of evolution II: skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London 186*, 343–414.

Pena, E. A. and M. Hollander. Models for recurrents events in reliability and survival analysis. In R. Soyer, T. A. Mazzuchi, and N. D. Singpurwalla (Eds.), *Mathematical Reliability: an Expository Perspective*, pp. 105–123. Boston: Klüwer Academic Publishers.

Poisson, S.-D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et matière civile.*

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays*, pp. 156–198. London: Kegan Paul.

Saunders, S. C. (2006). *Reliability, life testing and prediction of service lives.* New York: Springer.

Savage, L. J. (1954). *The foundations of statistics* (First ed.). New York: Wiley.

Seal, H. L. (1980). Early uses of Graunt's life table. *J. Inst. Actuaries 107*, 507–511.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, 639–650.

Singpurwalla, N. D. (1971). Problem in accelerated life testing. *Journal of the American Statistical Association 66*, 841–845.

Singpurwalla, N. D. (1995). Survival in dynamic environments. *Statistical Science 10*, 86–113.

Singpurwalla, N. D. (2006). *Reliability and risk: a Bayesian perspective*. Chichester: Wiley.

Singpurwalla, N. D. and S. P. Wilson (1999). *Statistical methods in software reliability: reliability and risk*. New York: Springer.

Spizzichino, F. (2001). *Subjective probability models for lifetimes*. Boca Raton: Chapman and Hall/CRC.

Stigler, S. S. (1990). *The history of statistics: the measurement of uncertainty before 1900*. Cambridge: Harvard University Press.

Ushakov, I. (2000). Reliability: past, present and future. In *Recent advances in reliability theory: methodology, practice and inference*, pp. 3–22. Berlin: Birkhäuser.

Venn, J. (1866). *The Logic of Chance*. London: McMillan and Co. Third edition published by McMillan & Co., London, 1888. Fourth edition published by the Chelsea Publishing Company, New York, 1962.

von Mises, R. (1919). Grundlagen der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift 5*, 52–99.

von Neumann, J. and O. Morgenstern (1944). *Theory of games and modern behavior*. Princeton, N.J.: Princeton University Press.

von Plato, J. (1994). *Creating modern probability.* Cambridge: Cambridge University Press.

Wald, A. (1950). *Statistical decision functions.* New York: Wiley.

Watson, G. S. and W. T. Wells (1961). On the possibility of improving the mean useful life of items by eliminating those with short lives. *Technometrics 3*, 281–298.

Weibull, W. (1939). A statistical theory of the strength of materials. *Ingeniörs Vctenskaps Akademiens, Handlingar 151*, 45–55.

Weibull, W. (1951). A statistical distribution function of wide applicability. *ASME J. Appl. Mech. 18*, 293–297.

Zelen, M. (1963). *Statistical theory of reliability.* Madison, Wi.: The University of Wisconsin Press. As editor, Proceedings of an advanced seminar by the Mathematical Research Center, U.S.A. Army.