Statistica Sinica **18**(2008), 9-90

MODELING AND PREDICTING EXTRAPOLATED PROBABILITIES WITH OUTLOOKS

Mendel Fygenson

University of Southern California

Abstract: To evaluate the conditional probability of an adverse outcome from a set of covariates, decision makers are often given a limited number of observations and, at times, are required to extrapolate outside the data range. To tackle the extrapolation problem they need to select plausible model(s) and account for various uncertainties in their predictions.

In this paper I propose a framework that provides a pessimistic (optimistic) decision maker with probability models that are consistent with his/her outlook. Viewing the link function in the GLM as a decision weighting function - a key feature of modern choice models in economics - I characterize the outlook of various distributions and order them according to their degree of pessimism (optimism).

A complementary statistical inference procedure is presented for predicting constrained extrapolated probabilities. The statistical inference accounts for two different model uncertainties: model uncertainty in the data range and model uncertainty beyond the data range. The latter cannot be data driven and is dealt with using non-parametric models constrained to capture the decision maker's degree of pessimism (optimism). The proposed methodology is demonstrated by analyzing the 1986 Challenger space shuttle disaster and in assessing the merits of various approaches (e.g., Bayesian, parametric or non-parametric) in handling extrapolation model uncertainty.

Key words and phrases: Challenger disaster, choice models, decision weighting functions, extrapolation, model uncertainty, stochastic ordering.

1. Introduction

In the majority of cases where statistical analysis is called for, one encounters model uncertainty and incomplete data. After all, the true model is rarely known and it is the exception when the data is gathered from a designed randomized experiment. It is therefore not surprising that problems caused by model uncertainty and incomplete data have been hot topics of research in modern statistics resulting in many discussion papers.

Most recently, Copas and Eguchi (2005) considered an asymptotic framework for exploring the bias in likelihood inferences that results from incomplete data in the presence of "local" model uncertainty. Similar issues in the

specific contexts of epidemiological and observational studies were discussed by Greenland (2005) and by Rosenbaum (2004).

In this paper, the focus is on one of the most extreme cases of model uncertainty and incomplete data - that which occurs when extrapolating probabilities to the tail of a distribution. The contextual setup is one in which a decision maker, in the course of a worst-case analysis, must predict the extrapolated probability of an *adverse* outcome from a given set of predictor variables.

Worst-case analyses are a common challenge in many organizations. Government agencies, like the EPA, FDA or DOD, extrapolate probabilities of potential disasters to determine strategies and justify regulations. In the business world, jobs and wealth can be wiped out when the probabilities of future adverse events are projected wrongly. In general, wherever computer networks are used there is a reliance on complex algorithms to predict the probability of system failures. In all of these applications, risk managers face two major challenges: one is *selecting* a plausible model and the other is *accounting* for uncertainties in their predictions. Often, these challenges are made especially difficult because data is sparse.

The simplest setup for extrapolating probabilities is the classic binary regression, where inferences regarding the probability of a binary variable Y given the covariates X are captured via the model

$$P(Y_{ij} = 1|X_i) = F(\beta^T X_i)\,, \tag{1.1}$$

where $F(\cdot)$, the cdf of a latent variable, formulates the *structural* assumption and β is a vector of unknown parameters. Within the GLM framework, (1.1) is expressed as

$$F^{-1}(\theta_i) = \eta_i\,, \tag{1.2}$$

where $F^{-1}(\cdot)$ is the link function, $\eta_i = \beta^T X_i$ and $\theta_i = P(Y_{ij} = 1|X_i)$.

To make inferences in (1.1) or (1.2), the usual approach is to select a single model, $F^*(\beta^T X)$, that fits the sample data "best", and then proceed with inferences on the parameters β as if (i.e., conditionally) $F^*(\cdot)$ is the *true* structure. It is standard practice (recommended in most textbooks) to fit one of the following distributions to the data: logistic, normal, or extreme value – commonly referred to as the Logit, Probit and Clog-log models, respectively. In other words, the approach regularly used in a majority of applications addresses the *parameter uncertainty* of β (e.g., via confidence intervals) but not the *structure uncertainty* of $F^*(\cdot)$.

This practice of neglecting structure uncertainty was first criticized by (mainly) Bayesian researchers (e.g., Draper (1995), Chatfield (1995)) and led to a flurry of publications that use the Bayesian approach (see Clyde and George

(2004) for a good account). Chatfield (1995) asked "*why has mainstream statistics been ignoring* [*structure*] *uncertainty*?" This he found especially strange since the errors resulting from structure uncertainty sometimes account for a large proportion of the total uncertainty. He offered two explanations: (1) we have failed to ask the fundamental question: "*how do you choose the models to be considered in the first place*?"; and, (2) the *frequentist* approach does not adapt easily to cope with model uncertainty.

The issue of model uncertainty is most pronounced and challenging in extrapolation problems. This is due, in part, to a practical dilemma (PD): often, different structures $F^*(\cdot)$ fit the data equally well, yet lead to significantly different predictions *outside* the data range (*e.g.*, Chambers and Cox (1967) or Section 4.2).

In this paper I present a new approach to this challenging problem. For simplicity, the presentation is in the context of a binary regression with one numerical factor X, observed in the range (X^L, X^U), and the aim is to extrapolate the probability of $P(Y = 1|X = X^{Extp})$ where $X^{Extp} > X^U$.

When considering the extrapolation problem above, researchers often begin with model (1.1). This model implies that a single structure can capture all kinds of available information. The model thus makes it natural to assume that the observable relationship continues to hold outside the data range. However, this is an *unverifiable* assumption. It is therefore prudent to first replace (1.1) with the following general model:

$$P(Y_i = 1|X_i = x) = F_0(\alpha + \beta x)I(X^L \le x \le X^U) + F_1(x)I(x > X^U)\,. \quad (1.3)$$

In (1.3), two very different modeling problems are made explicit, highlighting two different types of structure uncertainty. One (the kind considered by Draper (1995) and other Bayseians) is uncertainty with respect to F_0. The associated uncertainty is data-driven and can be dealt with, for example, by the *Bayesian model averaging* approach (e.g., Raftery, Madigan and Hoeting (1997)) or its frequentist version (e.g., Buckland, Burnham and Augustin (1997)). The other (unique to extrapolation problems) is uncertainty with respect to F_1. In this case, the modeling cannot be data-driven and thus requires a different approach.

When the objective is to find a well calibrated fit within the data range, the Bayesian criticism is compelling: an inference procedure that recognizes variability in the observations when estimating the parameters of a model but not when estimating its structure F_0 is logically flawed. This flaw may lead "... to inaccurate scientific summaries and over confident decisions that do not incorporate sufficient *hedging* against uncertainty" (Draper (1995), emphasis added).

However, when the objective is to extrapolate probabilities in F_1, any data-based approach, such as model averaging, is equally indefensible.

The philosophy behind my approach was aptly expressed by P. J. Diggle (2005): 'So how do we make progress with the plethora of challenging scientific problems which are not amenable to investigation through randomized experiments? We build models. And, in so doing, 'we buy information with assumptions' (Coombs (1964)). In general, assumptions which are based on judgments by subject-matter scientists have a high rate of exchange against assumptions which statisticians adopt as a matter of convenience."

The approach I propose is based on viewing $F_1^{-1}(\cdot)$ in the same way behavioral economists view a *decision weighting function*, that is, as a function designed to capture the outlook of a decision-maker. It is well known that decision makers sometimes rely on considerations external to the data when selecting F_1. For example, in matters of life-or-death, a decision maker will often adopt a *pessimistic* outlook to perform a *worst-case* analysis. It is worth noting that the notions of a *pessimistic* or *optimistic outlook* are completely different from, and should not be confused with, the notion of a *risk averse* or *risk seeking* strategy, as commonly used in the economics literature. The latter refer to properties of utility functions while the former are transformations of objective or subjective probabilities (see Diecidue and Wakker (2001)).

To capture a particular outlook in the selection of a distribution function F, scientific context is indispensable. The ideas proposed herein therefore cannot be couched purely in terms of mathematical constraints. Instead, I follow the long-standing tradition in statistics in which functionals of F that have contextual meaning are used in developing non-parametric or semi-parametric frameworks. Notable examples include the Increasing (Decreasing) Failure Rate distributions in reliability studies (Barlow and Proschan (1981)), the Increasing (Decreasing) Odds Ratio models in binary regression (Fygenson (1997)) and the Proportional Hazards model in survival analysis (Cox (1972)). The functionals of F that are meaningful in our context are developed in Section 2.

This paper has three main objectives: first, to propose a constrained non-parametric framework for modeling probabilities with various outlooks; second, to develop a statistical inference procedure that compliments this framework and accounts for both structure and parameter uncertainties; and, third, to investigate, in light of the practical dilemma (PD) above, the merits of various approaches - Bayesian, parametric and non-parametric - in accounting for model uncertainty in extrapolation problems.

To avoid controversy over whether models are true, throughout the paper the word 'model' is used to mean a reduced and parsimonious mathematical representation of a system with relevance to a *specific* objective. The non-parametric

models developed here are designed to capture outlooks with various degrees of pessimism, thereby allowing one to *hedge* against understating the probability of a catastrophe.

In the next section, a (non-parametric) framework for modeling cumulative probabilities with a given outlook is introduced. The framework is general but will be motivated for modeling F_1 within the extrapolation setup of (1.3). Extensions to other statistical models are deferred to Section 5. In Section 3, the statistical methodology for predicting extrapolated probabilities is presented. In Section 4, the Challenger space shuttle disaster is used to illustrate the underlying ideas of the paper. The proposed methodology is used to analyze the pre-launch data and the merits of various approaches for handling model uncertainty (i.e., Bayesian, parametric or non-parametric) are examined. The presentation is such that this section can be read independently of the more technical material. Section 5 ends the paper with thoughts on important issues and open problems. Technical theorems and proofs are relegated to Appendices A and B, respectively.

2. A Framework for Selecting Distributions in Extrapolation Problems

Consider a binary regression with one numerical factor X, observed in the range (X^L, X^U). The aim is to extrapolate $P(Y = 1|X = X^{Extp})$ for $X^{Extp} > X^U$ or, at (1.3), to model F_1 where data is absent.

In general, the fewer restrictions imposed on F_1, the smaller the structure uncertainty. In the statistical literature, F_1 is taken to be equal to F_0 at (1.3) and one encounters two extreme approaches to its modeling. The most restrictive uses a single parametric model (e.g., logistic or normal) and the least restrictive uses a (monotone) non-parametric model. The former maximizes model uncertainty while the latter completely eliminates it. However, in many applications (e.g., Section 4.5) the latter is neither powerful enough nor efficient enough to resolve practical matters.

It is therefore reasonable to explore alternatives between these two extremes. One alternative is to depart from a single parametric model and incorporate other parametric models (e.g., Bayesian Model Averaging). Another is to depart from a non-parametric model and impose additional qualitative constraints. The latter promises minimal structure uncertainty - provided the additional constraints can be justified conceptually.

In the following subsections I identify analytical constraints that capture a decision-maker's outlook (e.g., some degree of pessimism). Since constraints are only meaningful with respect to a family of functions, I first introduce a class of *outlook-revealing transformations* (ORT) that are based on measures of association commonly applied in contingency table analysis.

2.1. Contingency table analysis

To get to a contingency table analysis from a binary regression setup, consider an idealized scenario in which an experiment can be run on k (possibly infinite) discrete values of the factor X in the interval $[X^U, X^{Extp}]$ (i.e., $X^U \leq X_1 < \cdots < X_k \leq X^{Extp}$). The outcome of the experiment (m_i) and the true probabilities (π_i) can be summarized in a $2 \times k$ contingency table:

	X_1	X_2	X_3	$\cdots$	X_k
$Y = 1$	$m_1(\pi_1)$	$m_2(\pi_2)$	$m_3(\pi_3)$	$\cdots$	$m_k(\pi_k)$
$Y = 0$					
Total	N_1	N_2	N_3	$\cdots$	N_k

Note that π_i in the above table is equal to $F_1(X_i)$. Therefore, considering different relationships among the π_i is equivalent to identifying various *qualitative* constraints of $F_1(\cdot)$. Contingency table analysis often starts with testing the following hypothesis:

$$\pi_1 = \pi_2 = \cdots = \pi_k \Leftrightarrow \textit{Column homogeneity} \Leftrightarrow \text{no covariate effect.}$$

Once the null hypothesis of *Column homogeneity* is rejected, one traditionally proceeds to estimate the *strength* and *direction* of the relationship between the response, Y, and the factor, X, via *measures of association*. (A measure of association, $\rho(x, y)$, between two variables is a population parameter that characterizes their joint variation.)

Although different measures of association have been proposed for categorical variables, the most commonly used are *attributable risk* (AR), *relative risk* (RR) and *odds ratio* (OR). I focus on these three because they have been widely applied as measures of *extra-risk* by practitioners in different fields. Since, by construction, the risk of an adverse outcome is a non-decreasing function of the exposure level, a given outlook cannot be represented by a monotone constraint on the risk itself. Measures that describe extra-risk mechanisms therefore provide natural platforms for capturing decision-maker's outlooks (i.e., *outlook-revealing transformations*).

2.2. Outlook-revealing transformations

To return to a binary regression setup, I re-define the three measures of association as functionals of a cdf (F):

$$\begin{aligned}
&\text{(i)} \quad AR_F(x^*, x) = F(x^*) - F(x);\\
&\text{(ii)} \quad RR_F(x^*, x) = \frac{F(x^*)}{F(x)};\\
&\text{(iii)} \quad OR_F(x^*, x) = \frac{F(x^*)}{1 - F(x^*)} \Big/ \frac{F(x)}{1 - F(x)};
\end{aligned} \tag{2.1}$$

where x^* and x are any two fixed values such that $x<x^*$, $F(x)>0$ and $F(x^*)<1$.

The functionals in (2.1) are particularly appropriate for developing a constrained non-parametric framework for extrapolation problems because they capture the strength and direction of the relationship between Y and X *without* requiring explicit knowledge of the formula of F.

To see why these functionals qualify as outlook revealing transformations, consider the case where $Y = 1$ when an adverse outcome occurs, and where x and $x^* = x+\Delta$, with $\Delta > 0$, represent the levels of a risk factor. The *attributable risk*, for example, measures the *excess-risk* from the additional exposure Δ,

$$AR_F(x+\Delta, x) = (P(Y = 1|X = x+\Delta) - P(Y = 1|X = x)) > 0 \text{ for all } x,$$

and a plot of $AR_F(x + \Delta, x)$ vs. x captures changes in the *excess-risk* with increasing exposure x. A *neutral* outlook would suppose that there is *no* change in the excess-risk for more exposed $(x + \Delta)$ individuals as compared to less exposed ones (x). That is, the increase in the risk factor level is inconsequential (i.e., neutral) to the excess-risk mechanism. A *non-decreasing* pattern (at least in some of the x) is intuitively *pessimistic* because it implies that the excess risk of more exposed individuals is non-decreasing in the risk factor levels. That is, an increase in the level of a risk factor has a compounding effect on more exposed individuals. Similarly, if the changes in the excess-risk were *non-increasing* (i.e., at higher exposure levels, the risk factor is "protective" in the sense that more exposed individuals accrue less additional risk than those that are less exposed), this would amount to an *optimistic* outlook on the excess-risk mechanism. The same intuition applies for monotone patterns in the two other ORT since they also reflect the excess-risk mechanism, albeit on a relative scale (see Section 4.3).

2.3. Pessimistic, neutral or optimistic regions of distributions

Given that the ORT in (2.1) are functionals of a cdf, the intuition developed above can be carried over to characterize the outlook(s) inherent in a distribution. Note that a single distribution can reflect different outlooks over different intervals of its support. Focusing on only an interval (and not the entire support) departs from common practice in the study of stochastic orders, but leads to more general results. Moreover, it increases the likelihood that the derived constraints are contextually meaningful and, at the same time, minimizes structure uncertainty.

Definition 2.1. A distribution function F is inherently *pessimistic* (*optimistic*) on an interval J if its ORT is *non-decreasing* (*non-increasing*) in x for all $x^* \in J$. F is said to be inherently *neutral* on J if its ORT is *constant* in x for all $x^* \in J$.

In the above definition, it must be understood that, *on some scale*, x^* must be larger than x by a fixed positive constant in the neighborhood of its boundary value (e.g., $x^* = x+\Delta$, $\Delta > 0$ or, $x^* = \alpha x$, $\alpha > 1$). When applying the definition to characterize a certain family of distributions, an ORT (i.e. a functional that depicts an extra-risk mechanism) and an appropriate scale must be selected. A scale would be appropriate if it allows the ORT to discriminate among the distributions in the considered family such that some of the distributions are classified as pessimistic and others are classified as neutral or optimistic.

Throughout the rest of the paper, the family $\mathcal{F}$ of distribution functions is considered. This family includes all distributions that are absolutely continuous with densities that are strictly positive and have at least one continuous derivative on $(-\infty, \infty)$. In accordance with the ORT in (2.1) and Definition 2.1, the focus here is on distributions that can be labeled as *AR-*, *RR-* or *OR-pessimistic* (*optimistic*) on an interval J of their support.

In the definition below, Δ denotes any positive constant. To make the comparisons meaningful however, Δ should either be taken in the neighborhood of zero or, at least, be small relative to the length of the interval J.

Definition 2.2. A distribution function $F \in \mathcal{F}$ is inherently *AR-*, *RR-* or *OR-pessimistic* (*optimistic*) on an interval J if $AR_F(x+\Delta, x)$, $RR_F(x+\Delta, x)$ or $OR_F(x+\Delta, x)$ is *non-decreasing* (*non-increasing*) in x, respectively, for all $x+\Delta \in J$. F is said to be inherently *AR-*, *RR-* or *OR-neutral* on J if $AR_F(x+\Delta, x)$, $RR_F(x+\Delta, x)$ or $OR_F(x+\Delta, x)$ are *constant* in x, for all $x+\Delta \in J$.

The above definition provides an *ordinal* classification of candidate distributions from pessimism to optimism. However, it offers no guidance as to how to classify distributions within the same category. To make an informed choice between, for example, two or more pessimistic distributions, it is useful to order the distributions according to their *degree* of pessimism (optimism). This is done in Appendix A. The theorems that appear therein are the foundation for much of the statistical methodology in Sections 3 and 4. What follows is a summary of some of the important results derived in Appendix A.

Summary of Results:

1. On the same interval J, when F is *RR*-pessimistic it is also *OR*-pessimistic and *AR*-pessimistic (i.e., *RR*-pessimism is the most pessimistic outlook). Other implications are not possible without further restrictions on F or J.
2. In applications where the subinterval of interest is $J_{interest} \subset (-\infty, \text{Median}]$, F is *RR*-pessimistic $\Rightarrow$ F is *OR*-pessimistic $\Rightarrow$ F is *AR*-pessimistic.

3. The three most commonly used models in binary regressions (*i.e.*, the logit, the probit and the Clog-log models) often fit the same data equally well (e.g., the application in 4.2), but actually convey very different, or even opposite, outlooks regarding the likelihood of adverse outcomes. In particular, the probit is *OR-optimistic*, the logit is *OR-neutral* and the Clog-log is *OR-pessimistic*. Therefore, choosing one of these models over another for F_1 amounts to taking a different outlook position when making decisions in worst-case analyses.

3. Modeling and Predicting Extrapolated Probabilities

To develop the statistical methodology for the probabilistic framework proposed in Section 2, the extrapolation model in (1.3) must first be operationalized. This requires connecting and matching the structures F_0 and F_1 in a single coherent model. For this purpose, I introduce a *transition* structure F_{01}, defined on a small interval at the extreme of the observation range (i.e., an interval $[X^U - \epsilon, X^U]$, where ϵ is some positive scalar and X^U denotes the largest observation). This leads to the following inference model for predicting $P(Y_i = 1|X_i = x)$:

$$F_0(x)I(x < X^{U^-}) + F_{01}(x)I(X^{U^-} \leq x \leq X^U) + F_1(x)I(X^U < x \leq X^* \leq X^{Extp}), \tag{3.1}$$

where is $I(\cdot)$ is the indicator function and $(X^U - \epsilon = X^{U^-} < X^U < X^* \leq X^{Extp})$. The variable X^{U^-} is used as an anchor point in deriving confidence bounds (see Theorem 3.1 below). The variable X^* is included to shorten the length of the interval on which F_1 is required to satisfy the qualitative constraint, thereby minimizing structure uncertainty. In many applications (e.g., Section 4.3), a decision can be made by extrapolating to $X^* << X^{Extp}$.

The particular constraints considered next are imposed on the ORT in (2.1) and come in a variety of strengths (see Appendix A). These qualitative constraints are imposed on F_1 through non-parametric models that are inherently *pessimistic* with respect to one of the three ORT.

The next theorem is essential for conducting a constrained statistical analysis. It provides sharp lower bounds for $P(Y = 1|X = x)$, $x \in J^* = (X^U, X^*]$ using two percentiles $(x_p < x_q)$ of F_{01} (i.e., $F_{01}(x_p) = p$ and $F_{01}(x_q) = q$). To present the theorem, the following definition is required.

Definition 3.1. *Stochastic ordering on J.* Consider the random variables W and X with distribution functions G and F, respectively. W is said to be stochastically larger than X on $J = [a, b]$ if $F(a) = G(a)$ and $F(u) \geq G(u)$ for all $u \in j = [a, b]$.

Theorem 3.1. *Suppose that on $J = [X^{U^-}, X^*]$, F_1 is pessimistic with respect to an ORT and on $J_0 = [X^{U^-}, X^U]$, F_{01} is stochastically larger than F_1, then*

$$g(F_1(x)) \geq \sup_{p<q}\{Ag(q) - (A-1)g(p)\}, \quad X^{U^-} = x_p < x_q \leq X^U < x \leq X^*, \tag{3.2}$$

where $0 < p < q < 1$, $A = (x - x_p)/(x_q - x_p)$, and the function $g(\cdot)$ is the identity, the log or the logit function according to whether F_1 is AR-, RR- or OR-pessimistic, respectively.

Corollary 3.1. *Under the conditions of Theorem* 3.1*, the following are sharp lower bounds for any percentile $X_\gamma \in J^* = (X^U, X^*]$:*

$$X_\gamma \leq \inf_{x_p<x_q}\{Bx_q - (B-1)x_p\}, \quad X^{U^-} = x_p < x_q \leq X^U < X_\gamma \leq X^*, \tag{3.3}$$

where $B = (g(\gamma) - g(p))/(g(q) - g(p))$ and the function $g(\cdot)$ is the identity, the log or the logit function according to whether F_1 is AR-, RR- or OR-pessimistic, respectively.

Remarks.

1. For the inequality in (3.2) or (3.3) to be valid, F_{01} does not have to be equal to or even of the same pessimistic type as F_1. However, if F_{01} is as pessimistic as F_1, one can obtain tighter bounds in (3.2) and (3.3).
2. In relation to the three most commonly used binary regression models (i.e., probit, logit and Clog-log) the probit model will lead to the smallest lower bound in (3.2) and the Clog-log model to the largest.
3. From the inequalities in (3.2) and (3.3), it is clear that structure uncertainty due to F_0 only comes into play via the percentiles (x_p, x_q). If various parametric models fit the data equally well (the practical dilemma (PD) mentioned in Section (1), then it is highly probable that they yield similar estimates for (x_p, x_q). Therefore, uncertainty due to F_0 is likely to be inconsequential to the extrapolation problem (e.g., Section 4.3).

In the next section I derive statistical bounds for any extrapolated probability $F_1(x)$, $x \in J^*$ or percentile $X_\gamma \in J^*$. These are approximate $100(1-\rho)\%$ lower confidence bounds (LCB) or upper confidence bounds (UCB), respectively, that account for the variability inherent in the data.

3.1. Lower and upper confidence bounds for extrapolation

To input data into the estimation of $P(Y = 1|X^{Extp})$ or $X_\gamma \in J^*$ requires modeling F_0 and accounting for data uncertainty. The inequalities in (3.2) and (3.3) allow for any approach in modeling F_0 - parametric, semi-parametric or

non-parametric. As will be shown in Section 4, the non-parametric approach can lead to uninformative summaries of the data. Therefore, in what follows, statistical bounds are derived for $F_1(x)$, $x \in J^*$ or $X_\gamma \in J^*$, where F_0 is modeled parametrically.

The problem of constructing approximate confidence bounds for percentiles in a binary regression setup has been considered by many authors, usually under the standard assumption of a logit or a probit model for F at (1.1).

With the exception of one case (where F_0 is a logistic and F_1 is OR-pessimistic), the derivation of the approximate $100(1-\rho)\%$ LCB for $F_1(x)$, $x \in J^*$ or UCB for $X_\gamma \in J^*$ is *non-standard* and requires a constrained maximum likelihood approach.

While other asymptotic approaches exist (e.g., the delta method), the likelihood ratio method has been found to have good theoretical properties. In particular, it is invariant under parameter transformations (Cox and Hinkley (1974)) and, in the standard models, it yields coverage probabilities close to their nominal values (e.g., Alho and Valtonen (1995) and Huang (2001)).

In the proposition below, $LL(\alpha, \beta)$ denotes the log likelihood function (of the parameters in F_0), $LL(\hat{\alpha}, \hat{\beta})$ denotes its maximum under the parameterization of (α, β),

$$H(\alpha, \beta;\, x, x_q, x_p) = \frac{x - x_p}{x_q - x_p} g(F_0(\alpha + \beta x_q)) - \frac{x - x_q}{x_q - x_p} g(F_0(\alpha + \beta x_p))$$

and

$$G(\alpha, \beta;\, \gamma, x_q, x_p) = \frac{(x_q - x_p) g(\gamma) + g(F_0(\alpha + \beta x_q)) x_p - g(F_0(\alpha + \beta x_p)) x_q}{g(F_0(\alpha + \beta x_q)) - g(F_0(\alpha + \beta x_p))} \,.$$

Proposition 3.1. *Suppose that on $J = [X^{U^-}, X^*]$, F_1 is pessimistic with respect to an ORT and on $J_0 = [X^{U^-}, X^U]$, F_{01} is stochastically larger than F_1. Then the approximate $100(1-\rho)\%$ LCB for $F_1(x)$, $x \in J^*$ is given by*

$$L_x = \inf_{(\alpha,\beta)} \{H(\alpha, \beta;\, x, x_q, x_p) : 2(LL(\hat{\alpha}, \hat{\beta}) - LL(\alpha, \beta)) \le \chi^2_{1,1-2\rho}\} \,,$$

and the approximate $100(1-\rho)\%$ UCB for $X_\gamma \in J^*$ is given by

$$U_\gamma = \sup_{(\alpha,\beta)} \{X_\gamma = G(\alpha, \beta;\, \gamma, x_q, x_p) : 2(LL(\hat{\alpha}, \hat{\beta}) - LL(\alpha, \beta)) \le \chi^2_{1,1-2\rho}\} \,.$$

Finding U_γ, for example, is a nonlinear programming problem with a concave objective function and nonlinear constraint functions. After some algebra, it

can be shown that a solution must satisfy the following Kuhn-Tucker conditions (Kuhn and Tucker (1951)):

$$\frac{\partial G(\alpha,\beta;\gamma,x_q,x_p)}{\partial\alpha} - \lambda\frac{-\partial LL(\alpha,\beta)}{\partial\alpha} = 0,$$
$$\frac{\partial G(\alpha,\beta;\gamma,x_q,x_p)}{\partial\beta} - \lambda\frac{-\partial LL(\alpha,\beta)}{\partial\beta} = 0,$$
$$-LL(\alpha,\beta) + \frac{1}{2}(2LL(\hat{\alpha},\hat{\beta}) - \chi^2_{2,1-2\rho}) = 0,$$

where $\lambda > 0$ is the Lagrange multiplier.

4. Application: The 1986 Challenger Disaster

The decision process that took place prior to the 1986 Challenger disaster remains typical of contemporary debates in *worst case* analyses. Data from previous space shuttles launches has been used several times over the past two decades to illustrate and assess statistical methodologies for extrapolating the probability of a catastrophe. Revisiting the data here therefore offers both a historical perspective on how statistical methodologies have progressed and aptly motivates modeling probabilities with a given (pessimistic) outlook.

4.1. Background

On the morning of January 28, 1986, shortly after take-off, the space shuttle Challenger exploded, killing all seven crewmembers. The catastrophe was due to a breach in the field joint of a booster rocket, caused by the malfunctioning of two primary O-rings in the joint.

The decision to launch the Challenger was reached the night before after a three-hour teleconference between *engineers* and *management*. A central question in the discussion was whether the anticipated launch temperature, 31 degrees, would allow normal O-ring function. One year earlier, similar O-rings in the booster rocket of Flight 51-C, which was launched at 53 degrees - the lowest temperature of any launch up to that time - were found damaged and breached. This led the engineers to regard low temperatures as dangerous to proper functioning of the O-rings.

The engineers' position was that the flight should be delayed until the Challenger's O-rings' temperature reached 53 degrees. They argued that, taking all "bench" test data and flight data into account, the chances of O-ring failure could only increase with decreasing temperature below 53 degrees.

Management countered that the data did not support the engineers recommendation for delay. They cited a subset of data from bench tests conducted

below 53 degrees that had shown no O-ring damage, and the successful launch of flight 61-A, at 75 degrees, in which the O-rings had shown damage. Their position was that the data failed to show that temperature was a key factor in O-ring performance.

At the time, a formal statistical analysis of the data was not performed. Nevertheless, the following information relevant to any such analysis was available.

(a) The engineers judgment that the probability of erosion may rise monotonically as temperature decreases.
(b) The number of primary O-ring failures (out of 6) as a function of launch temperature for each of the previous 23 space shuttles launches.
(c) All previous launches took place between 53 and 81 degrees. Thus, to predict the probability of failure at the anticipated launch temperature of 31 degrees requires a sizable extrapolation.

4.2. Review of statistical analyses

The first published analysis (in the statistical literature) on the Challenger disaster is by Dalal, Fowlkes and Hoadley (1989), who set out to show "... how statistical science could have provided valuable input to the launch decision process". This paper prompted two more landmark analyses, one by Lavine (1991) and the other by Draper (1995).

All three considered the statistical task of modeling the probability of O-ring failure ($Y = 1$) as a function of launch temperature (X), based on a total of 6*23 = 138 independent binary observations, and their associated temperature values (Figure 1). Since the anticipated launch temperature was 31 degrees, the analyses focused on providing the best possible prediction for$P(Y = 1|X = 31) \equiv p(31)$.

All three analyses started with the model at (1.1). Dalal, Fowlkes and Hoadley (1989) adopted a *linear logistic* model for their inferences. They first verified, via goodness-of-fit tests and other diagnostics, that the model fit the data extremely well and that there is no quadratic or other nonlinear relationship between the logit of the probabilities and temperature. (The MLE of the logistic regression curve is plotted in Figure 1.) To account for *parameter uncertainty*, they constructed a 90% confidence interval using a parametric bootstrap procedure, which reduces the variability but, as they noted, is highly model dependent. Their reported 90% confidence interval for $p(31)$ is (0.5, 1).

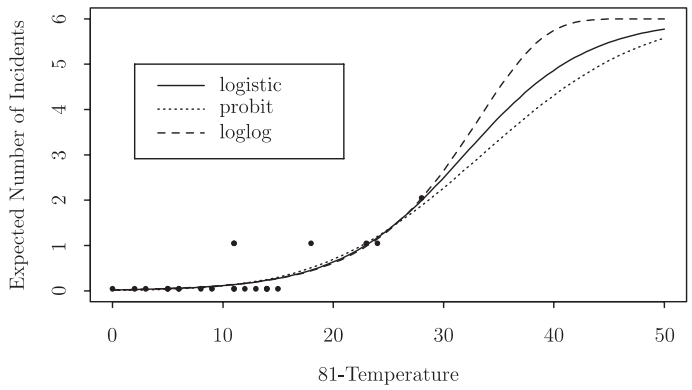

Figure 4.1. 1986 Challenger data fit with common parametric models. The rescaled variable Z = (81-Temperature) is plotted along the x-axis so that the risk is non-decreasing rather than non-increasing.

Lavine (1991) rejected their approach by showing that there are other models besides the logit that fit the data equally well (see Figure 1), yet give very different predictions for very low temperatures. This practical dilemma (see PD, Section 1), he noted, is common to other extrapolation problems, where "... the usual procedures of model selection, model fitting, and diagnostics do not tell the whole story about the probability of failure at 31 degrees". Lavine concluded that a reliable answer could only be based on the engineers' input that the probability of O-ring failure may rise monotonically as temperature decreases. Accordingly, Lavine used a constrained non-parametric framework where $p(31)$ was estimated to be at least 1/3. He acknowledged that the estimate is highly dependent on the one flight at 53 degrees and that the range of 2/3 is rather too large. It is important to note that this range $[1/3, 1)$ does not account for parameter uncertainty and therefore underestimates the actual range size. When parameter uncertainty is taken into account (Section 4.5), the range is about $[1/10, 1)$ - much too wide to be of practical use.

Draper (1995) presented another solution for predicting $p(31)$. It appears as a demonstration of the discrete model expansion (DME) approach - better known nowadays as the *Bayesian model averaging* approach - to account for *structure* uncertainty. The DME approach requires that we choose a finite set L of plausible competing structures and attach priors to each of them. Draper stated that the sensitivity of the analysis of Dalal, Fowlkes and Hoadley (1989) indicates each of the following S_i structures as a plausible candidate for inclusion with *equal discrete* prior probability:

$$L = \{\text{Cloglog}(x), \text{logit}(x), \text{probit}(x), \text{logit}(x, s), \text{logit}(x, x^2), \ \textit{notemp.effect}\},$$

where x is temperature and s is leak check pressure. Draper's analysis highlights the followings points.

(a) The posterior distribution differs considerably from the point mass on the logit(x) model – the implicit framework of Dalal, Fowlkes and Hoadley (1989).
(b) While the model with *no temperature effect* is not supported, all other models are plausible.
(c) The total variance, (i.e., $\text{variance}_{\text{within model}} + \text{variance}_{\text{between models}}$) is $(0.0338 + 0.0135) = 0.0473$, is more than twice the variance of the logit model while the variance between models is about 1/3 of the total variance.
(d) The 90% confidence interval for $p(31)$ is $(0.33, 1)$.

Table 4.0 summarizes the predictions of $p(31)$ in the above three papers and the various uncertainties they (try to) account for.

Table 4.0. Summary of the three published solutions for $p(31)$.

Method of	Estimate of $p(31)$	Uncertainty Accounted for?			90% CI for $p(31)$
		Parameter	(F_0) Structure	(F_1) Structure	
Dalal et al. (1989)	0.90	Yes	No	No	(1/2, 1)
Lavine (1991)	$\geq 1/3$	No	Yes	Yes	NA
Draper (1995)	0.88	Yes	Yes	No	(1/3, 1)

Note: Both Dalal, Fowlkes and Hoadley (1989) and Lavine (1991) run their analyses with two erosions at 75 degrees where there should be none. This has little (Dalal, Fowlkes and Hoadley (1989)) or no (Lavine (1991)) effect on the estimates. Point estimates for Dalal, Fowlkes and Hoadley (1989) and Draper (1995) are means of the posterior distribution for $p(31)$.

4.3. Justification for a pessimistic outlook

After reviewing the three analyses above, one must wonder whether any of them provide a solution that can be considered useful to decision-makers? Lavine did not think so in his 1991 paper, or in 1995 when he discussed Draper's 1995 paper. He suggested that "...any reduction of the range must come from modeling, rather than data consideration".

Section 2 follows up on Lavine's suggestion and provides a modeling approach that capitalizes on factors external to the data. In the Challenger case, based on management position on the eve of the launch, such factors include:
(i) awareness of the potentially disastrous consequences of O-ring failure;
(ii) the belief that the probability for O-ring failure at 31 degrees was very small;
(iii) the belief that $p(31)$ would not be significantly higher than $p(53)$.

Given the high stakes, their limited knowledge, and the engineers position, management would have wanted to make a *worst case* analysis and model $p(31)$ with a *pessimistic* outlook.

A pessimistic outlook cannot be conveyed by simply requiring the risk of catastrophe to be non-decreasing on an interval because this must hold by default for any interval. Therefore, to convey a pessimistic outlook one needs to consider non-decreasing constraints on some functional of the risk, such as the *extra risk* described by the ORT given in Section 2. This notion of pessimism becomes even more compelling in situations like the Challenger disaster where there is the belief that, on the extrapolation interval $J^* = [X^U, X^{Extp}]$, either *(i)* the absolute probability of a catastrophe is very small, or *(ii)* the increase in that probability, relative to $F_1(X^U)$, is insignificant. In this case, $OR_F(x+\Delta, x)$, for example, measures the "*relative-risk*" (of a more exposed $(x+\Delta)$ subject relative to a less exposed (x) subject) and the plot of $OR_F(x+\Delta, x)$ vs. $x \in J^*$ depicts changes in the "*relative-risk*" mechanism. In general, the smaller the interval J^* on which the monotone constraint is required to hold, the easier it is for a decision-maker to conceive of such a requirement as pessimistic and, at the same time, the structure uncertainty of F_1 is minimized.

The analysis of the pre-launch data using pessimistic structures for F_1 in (3.1) is presented next. The analysis is based on the methodology given in Section 3, where F_0 in (3.1) is modeled parametrically. This methodology takes full advantage of all the data but introduces maximal structure uncertainty with respect to F_0. Section 4.4 is therefore the opportune place in which to question the importance, if any, of the structure uncertainty of F_0 in the extrapolation problem. The question is considered with and without accounting for parameter uncertainty (see Tables 4.2 and 4.3, respectively).

To facilitate the presentation of the analyses in Sections 4.4 - 4.6, I use the rescaled covariate $Z = (81 - temperature)$. Note that on factor Z, the probability of erosion $p_z(z)$ is increasing rather than decreasing and the focus on estimating $p_z(50)$ is equivalent to estimating $p(31)$.

4.4. Analysis with pessimistic outlooks

Using the factor Z, the model in ((3.1) can be rewritten as

$$F_0(\alpha+\beta z)I(z < Z^{U^-}) + F_{01}((\alpha+\beta z)I(Z^{U^-} \leq z \leq 28) + F_1(z)I(28 < z \leq Z^* \leq 50)\,, \tag{4.1}$$

where $(28 - \epsilon \leq Z^{U^-} \leq 28)$. Note that 28 corresponds to 53 degrees, the coldest flight temperature observed, and 50 corresponds to 31 degrees, the anticipated launch temperature for the Challenger.

The main goal of the present analysis is to obtain statistical lower bounds on $p_z(50)$, or statistical upper bounds on some γ-percentile $Z_\gamma > 28$, in a way that minimizes the structure uncertainty of F_1 (SU_{F_1}), and accounts for parameter uncertainty as well as structure uncertainty in F_0 (SU_{F_0}).

4.4.1. The Importance of SU_{F_0}

In the statistical framework of Section 3, when parameter uncertainty is ignored, SU_{F_0} can be a factor only if the candidate models yield different estimates for (z_p, z_q) in *Theorem* 3.1. The fact that different parametric models fit F_0 equally well (e.g., Lavine (1991)) suggests that this may not be the case. It is prudent, therefore, to begin by investigating just how important SU_{F_0} is in the extrapolation of $p_z(50)$ or $Z_\gamma > 28$. If it is unimportant, then all the observations can be used to achieve the main goal, which is not to model F_0 but to (pessimistically) extrapolate predictions in the range of F_1.

Table 4.1 reports the smallest Z^* such that $p_z(Z^*) \geq 0.5$ or $p_z(Z^*) \geq 0.9$. The case where F_0 is modeled non-parametrically (e.g., Lavine (1991)) is included for comparison. Including such a model, which is free from structure uncertainty, limits the choices for the percentiles (z_p, z_q) used in (3.2) or (3.3). Like Lavine (1991), I used the two most extreme percentiles, corresponding to the second coldest and coldest launch temperature of 57 and 53 degrees, respectively.

All Z^* in Table 4.1 are strictly smaller than 50, regardless of the model used for F_0. This means that the *upper* bounds on the 50th and the 90th percentiles correspond to a warmer temperature than the expected launch temperature of 31 degrees. Therefore, when parameter uncertainty is ignored, the decision not to launch at 31 degrees is strongly advocated by *all* combinations of F_1 and F_0. Moreover, since the values of Z^* are very similar within an ORT, one can conclude that SU_{F_0} is inconsequential to the launch decision. That the various parametric models for F_0 provide Z^* values similar to those provided by the non-parametric model further reinforces this conclusion.

Table 4.1. The smallest Z^* for which $p_z(Z^*) \geq 0.50$ ($p_z(Z^*) \geq 0.90$).

F_0 and F_{01}	$p_z(24)$ ($p_x(57)$)	$p_z(28)$ ($p_x(53)$)	Z^* (F_1 is AR-pessimistic)	Z^* (F_1 is OR-pessimistic)	Z^* (F_1 is RR-pessimistic)
Increasing[1]	0.1666	0.3333	32.00 (41.60)	31.03 (40.65)	30.34 (33.74)
Clog-log	0.1908	0.3405	32.37 (42.95)	31.38 (42.59)	30.66 † (34.72)
Logistic[2]	0.1956	0.3326	32.89 (44.56)	31.88 (44.15)	31.06 † (35.50)
Probit	0.2003	0.3136	34.60 (48.70)	33.22 † (47.84)	32.17 † (37.41)

[1]from Lavine (1991). [2]from Dalal, Fowlkes and Hoadley (1989). † indicates that, on the interval [24, 28], F_{01} (=F_0) is not as pessimistic as F_1.

Note that parameter uncertainty is not accounted for in this table.

4.4.2. Upper confidence bounds for the median failure temperature

To achieve the main goal, I account for parameter uncertainty by obtaining the approximate 90% upper confidence bounds (UCB) for $Z_{0.5}$ - the *median* failure temperature (Table 4.2). These are based on *Proposition* 3.1, and were obtained using the *FindRoot* subroutine in *Mathematica*.

Readers may be appalled at my choice of such a high percentile. It suggests that one would tolerate a risk for O-ring failure (and its likely disastrous consequences) that is as likely as getting Tails when flipping a balanced coin! While the statistical methodology in Section 3 can be applied to any percentile, I chose the median so that the UCB might lead to an indisputable launch decision.

The results in Table 4.2 can be summarized and interpreted as follows.

1. SU_{F_0} is unimportant.
2. SU_{F_1} and the parameter uncertainty are important.
3. UCB in the last two columns are very similar and are strictly less than 50.
4. The case where F_1 is constrained to be AR-pessimistic and F_0 is modeled using the normal distribution (the most optimistic combination considered) the 90% UCB for $Z_{0.4}$ is 49.8812.

Thus there could have been only one reasonable decision – not to launch at 31 degrees.

Knowing that SU_{F_0} is inconsequential to the launch decision justifies the use of a mixed modeling approach (i.e., modeling F_1 in (3.1) non-parametrically but pessimistically, and F_0 parametrically) in obtaining the LCB or UCB. This approach, which is not semi-parametric in the usual sense, should be particularly advantageous in extrapolation problems with sparse data because it makes use of all the observations.

Table 4.2. 90% approximate UCB for the upper bound on $Z_{0.5}$.

F_0 and F_1	F_1 *AR*-pessimistic		F_1 *OR*-pessimistic	F_1 *RR*-pessimistic
	$z_p = 24$, $z_q = 28$	$z_p = 27$, $z_q = 28$	$z_p = 24$, $z_q = 28$	$z_p = 24$, $z_q = 28$
Probit	**LR:** 54.7126 **(49.8812)***	**LR:** 52.5553	**LR:** 43.8385 **(49.5924)** †	**LR:** 40.5974 †
Logistics	**LR:** 50.9487	**LR:** 48.4431	**LR:** 41.6636 **(47.2307)**	**LR:** 38.8322 †
Clog-log	**LR:** 50.0172	**LR:** 48.2864	**LR:** 41.0748 **(46.6518)**	**LR:** 38.3524 †

Note: The UCB for the cases where F_1 is *OR*- or *RR*- pessimistic using $z_p = 27$ and $z_q = 28$ resulted in differences of less than a degree from that when $z_p = 24$ and $z_q = 28$ were used and is therefore omitted. In the third column the numbers in parentheses are the 95% UCB. *indicates the value given is for the 90% UCB for $Z_{0.4}$. † indicates that on the interval [24, 28], F_{01} (=F_0) is not as pessimistic as F_1.

Might there nevertheless be merit in using a non-parametric approach for modeling F_0 and avoiding its structure uncertainty altogether? In Section 4.5, the standard non-parametric analysis of Lavine (1991) is considered in a manner which accounts for parameter uncertainty. The final results do not bode well for such an approach and, in general, cannot be recommended for extrapolation problems in applications with sparse data.

In Section 4.6, I investigate the possible improvement of using the non-parametric but pessimistic structure for F_1 (and the standard non-parametric model for F_0). To make it more interesting, the merit of the engineers' recommendation that the flight be delayed until the Challenger's O-rings temperature reached 53 degrees is analyzed for the first time. While the analysis reveals that using the standard non-parametric model for F_0 is once again fruitless, using the non-parametric but pessimistic structure for both F_1 and F_0 is productive.

4.5. The standard non-parametric analysis

Another way of bypassing the structure uncertainty controversy is to use a non-parametric model for F in (1.1). Recall that Lavine (1991) estimated $p_z(50)$ to be at least 1/3. This (monotone) non-parametric MLE is based on *one* binomial experiment at $z = 28$ (i.e., 53 degrees) where two failures were observed among six "independent" trials. This estimate, however, does *not* account for parameter uncertainty, which could be problematic in this case.

The main difficulty stems from the fact that not all the data can be used in deriving the confidence intervals, leading to wide confidence intervals or, equivalently, lower confidence levels. A related problem is the discreteness of the distributions of the statistics used in deriving the confidence intervals. This often leads to confidence levels more strict than the targeted nominal level. Finally, there is an ongoing dispute as to the best way of dealing with nuisance parameters - the conditional approach vs. the unconditional one (Agresti (2001)).

To see how these difficulties play out in the Challenger example, I derive a lower confidence bound (LCB) for $p_z(50)$ where both F_0 and F_1 are only constrained to be monotone (Lavine (1991)). Using the pool-adjacent-violators algorithm (Ayer, Brunk, Ewing, Reid and Silverman (1955)) the constrained non-parametric MLE are

$$\begin{aligned}
&p_z(0) = p_z(2) = p_z(3) = p_z(5) = p_z(6) = p_z(8) = p_z(9) = 0,\\
&p_z(11) = p_z(12) = p_z(13) = p_z(14) = p_z(15) = \frac{1}{30},\\
&p_z(18) = p_z(23) = p_z(24) = \frac{1}{6} \text{ and } p_z(28) = \frac{1}{3}.
\end{aligned}$$

With only three jumps in the range of the data, it makes most sense to use the estimated $p_z(28)$ as the lower bound for $p_z(50)$. To derive the $(1-\rho)\%$ LCB for $p_z(50)$, I consider three methods that have been researched extensively.

1. Clopper-Pearson's (1934) Exact method where the LCB, P_L, is a solution to the equation

$$\sum_{k=2}^{6} \binom{6}{k} P_L^k (1 - P_L)^{n-k} = \rho.$$

2. Mid-P Exact method (Berry and Armitage (1995)), where P_L is a solution to the equation

$$0.5 P_L^2 (1 - P_L)^4 + \sum_{k=2}^{6} \binom{6}{k} P_L^k (1 - P_L)^{n-k} = \rho.$$

3. Wilson's (1927) Score method with continuity correction (CC) where

$$P_L = \frac{2np + z^2 - 1z\{z^2 - 2 - \frac{1}{n} + 4npq + 4p\}^{0.5}}{2(n + z^2)},$$

where $q = (1 - p)$ and z is the $(1 - \rho)$ percentile of the standard normal distribution.

With only six observations at 28, the preference for exact methods rather than asymptotic approximation methods is obvious. We bring the Score method for comparison since it has a closed form and was found to perform very well in extensive evaluation studies (e.g., Newcombe (1998)). The results appear in Table 4.3 where $n = 6$, $p = 1/3$, and $z = 1.282$.

Table 4.3. 90% LCB for $p_z(50)$ in Lavine's Framework.

Method	**90% LCB**
Clopper-Pearson Exact	0.0925953
Mid-P Exact	0.1265290
Wilson Score with CC	0.0958944

Without further restrictions on F_0 or F_1, very low values for the LCB should be expected since only a small fraction of the data (6/138) can be used and no consideration for a pessimistic outlook (e.g., worst-case scenario) can be included. Therefore, in general, such an approach cannot be recommended for applications with sparse data.

4.6. Analyzing the engineers position

None of the analyses surveyed in Section 4.2 examined the merits of the engineers' position, which recommended delaying the flight until the Challenger's O-rings' temperature reached 53 degrees. To do so, I now derive the LCB for the excess-risk of launching the Challenger at 31 degrees as opposed to 53 degrees.

It follows from *Theorem* 3.1, that when F_1 is *AR*-, *RR*- or *OR*-pessimistic, one needs first to obtain the LCB for the difference between two proportions, their ratio and the ratio of their odds, respectively. Were it not for the constraint given in (3.2), these would have been standard problems. Consider the case where F_0 is monotone and F_1 is *OR*-pessimistic. It follows from *Theorem* 3.1 that the odds-ratio of interest $OR_{F_1}(50, 28)$ is bounded by

$$\max\left\{\theta^{(A-1)} = \Big(\frac{Odds(28)}{Odds(z_p)}\Big)^{A-1},\ \ 1\right\},$$

where $z_p < 28 < Z^* \leq 50$ and $(A-1) = (50-28)/(28-z_p)$.

To see if and when one can obtain an informative LCB (i.e., LCB > 1), all four possible observed values for z_p should be considered. The exact 90% LCB for the odds-ratios derived via the conditional approach are reported in Table 4.4. Note that the LCB in the first two rows of Table 4.4 are *not* informative. To obtain informative LCB, it was necessary to extend the range of F_{01} to $z_p < 24$ and, as a result, F_0 is no longer just monotone but also constrained to be OR-pessimistic.

Based on the results in Tables 4.3 and 4.4 one cannot, in general, recommend using a non-parametric monotone model for F_0 in applications where extrapolation is the main objective. In many such applications, the number of observations is limited and one cannot afford to ignore other information or use an approach that fails to summarize all the data.

Table 4.4. 90% LCB for the odds-ratio where F_0 is *monotone* and F_1 is *OR-pessimistic*.

z_p	z_q	p	q	θ	LCB for θ	$(A-1)$	$\theta^{(A-1)}$
24	28	1/6	2/6	2.5	Exact: 0.215419 Mid-P: 0.34860	5.5000	Exact: 0.000215 Mid-P: 0.00303
21	28	3/18	2/6	2.5	Exact: 0.354035 Mid-P: 0.53426	3.14286	Exact: 0.038257 Mid-P: 0.13943
13	28	2/60	2/6	14.5	Exact: 1.824160 Mid-P: 2.77624	1.46666	Exact: 2.414852 Mid-P: 4.47098
4.5	28	0/54	2/6	60.555	Exact: 4.66987 Mid-P: 8.28032	0.93617	Exact: 4.232364 Mid-P: 7.23516

Note: The values of z_p in the last three rows are the midpoints of the intervals where no jumps occur in the constrained empirical distribution of F_0. The estimated odds-ratio in the last row is obtained by adding 0.5 to each cell in the 2×2 table.

5. Discussion

5.1. Conservatism and pessimistic outlooks

Worst-case analyses are a common challenge in both public and private organizations. When decision-makers consider worst-case scenarios, a conservative

approach often is, and arguably should be, preferred. But the word conservative means different things to different people. In the statistical literature, a conservative approach can mean modeling without parametric assumptions (e.g., Horowitz and Manski (2000)); using lower bounds in evaluating the estimated variability or efficiency of estimators (e.g., Fuh and Hu (2004)); or choosing the least favorable distribution in testing procedures (e.g., Lehmann and Romano (2005)). In the economics literature, a conservative approach can mean using a risk-averse utility function or adopting a pessimistic outlook, as captured by a transformation on the outcomes probabilities (i.e., decision weighting functions). The outlook-revealing transformations proposed here anchor a probabilistic framework in the latter spirit, but are derived from measures of association commonly used in categorical data analysis.

Conservatism in the face of scientific ignorance raises philosophical issues, such as: *Can humanity advance without taking risk*? *What constitutes a defensible decision when lives are in danger*? A statistical model cannot address such questions directly, but the non-parametric framework introduced here provides a means for formalizing and capturing the sentiments of decision-makers who must grapple with such questions.

Some have argued that the Challenger disaster was the result of bad science because more experiments could have been carried out and a physically based model could have been advanced. Others would defend the decision to go ahead, citing the known costs of delay and the irreducible uncertainty in the data at the time. Using ORT to constrain the analysis, it is now clear that, even without new data, the conservatively predicted probability for O-ring failure was appallingly high — even after accounting for both model and parameter uncertainties. In the future, the use of pessimistic outlooks in reaching a decision can at least provide a defensible position for decision-makers in cases that result in unfortunate outcomes.

5.2. Model Uncertainty and Incomplete Data

The notion of model uncertainty is controversial because it depends on what one presumes to be the meaning of a true (statistical) model and how this model is related to the model(s) one ends up using. Different views on the subject are evident from the different approaches researchers use in dealing with model uncertainty.

The Bayesian approach invokes additional priors on a set of possible structures that can be supported by the data and/or alternative scenarios (e.g., Raftery (1996), Draper (1995), Draper, Saltelli, Tarantola and Prado (2000) and Clyde and George (2004)) and thereby averages out the models' uncertainty. In

this approach, the notion of an objective true model is not necessary. Model uncertainty is really intrinsic uncertainty in the processes that generated the data.

The frequentist approach, by contrast, holds that a true model exists but concedes that one cannot be certain whether the model used is the correct one. To deal with such a reality, frequentists can buy insurance in the form of a non-parametric analysis and pay with power and/or efficiency. Alternatively, they can use semi-parametric models or robust methods. The thinking behind any of these alternatives is that it is difficult to make useful inferences if the models under consideration are grossly misspecified (relative to the true model). Therefore, as a first step, one should quantify any biases that may result from "small" departures from the true model and/or conduct a sensitivity analysis on the extra parameters (e.g., Burnham and Anderson (2002), Huber (2003), Copas and Eguchi (2005), and Greenland (2005)).

Debate as to which approach offers a better solution to the problems caused by model uncertainty is misplaced when the main objective of the analysis is extrapolation. Either approach is equally likely to mishandle the most relevant component of model uncertainty if model (1.1) is considered. The proposed model in (1.3), through the sub-models F_0 and F_1, clearly differentiates between two very different kinds of model uncertainties, each of which requires a different treatment.

With respect to F_0, it is appropriate to ask how to combine model uncertainty with sampling variability. The various ideas - Bayesian or Frequentist - that appear in research on topics such as model misspecification, model sensitivity, model selection, robustness, ignorability assumptions, model bias and over-fitting, all have merit. Although these topics are concerned with different aspects of or causes for the resulting biased summaries, they all consider the presence of incomplete data and model uncertainty as two aspects of one problem.

With respect to F_1, the handling of structure uncertainty cannot be data driven and should rely on qualitative constraints with contextual meaning. The models in (1.3) or (3.1) are designed to take full advantage of all the data and at the same time avoid unverifiable data-related assumptions when selecting F_1. While I used a frequentist approach in my statistical analysis of the model in (3.1), one can view the selection of a pessimistic outlook for modeling F_1 as a *prior* choice justified by the absence of data and/or established scientific mechanisms.

In this paper, the focus has been on one of the most extreme cases of incomplete data- that which occurs when extrapolating probabilities to the tail of a distribution. In such cases all the observations are completely missing in the range of interest and thus the classic work by Rubin (1976) on inference with missing data and many of the well-known methods for incomplete data modeling (e.g., Little and Rubin (2002)) are not applicable.

5.3. The default outlook platform

In considering the Challenger data, and in other applications requiring extrapolation, a good case can be made for using the odds-ratio-transformation (OR_F) as a default in modeling F_1. First, this transformation is characterized through the natural parameter of the Bernoulli distribution (see Theorem A.1). Second, it provides a qualitative discrimination among the three most commonly used models in binary regression (see Theorem A.3). Finally, it identifies a unique boundary distribution, which captures a neutral outlook. A neutral category is useful, perhaps necessary, in any framework that attempts to characterize outlooks toward risk taking.

5.4. Extension to other modeling setups

The proposed framework presents a way for making statistical inferences that account for all major uncertainties if certain qualitative constraints on nonparametric distributions are accepted. The analysis of observational or experimental data from outlook-constrained distributions will have important implications in various fields where the focus is on describing and predicting the complex processes behind human actions or decisions (e.g., economics, sociology, or psychology).

In the economics and literature, for example, a shift to so-called non-expected utility choice models has taken place in the last two decades. In these models the classical independent axiom is usually relaxed by incorporating a decision-weighting function- much like a link function in GLM. Two notable examples are the rank-dependence model of Quiggen (1981) and the Cumulative Prospect Theory developed by Tversky and Kahneman (1992). As these models become increasingly popular, the characterization developed in this paper should have strong implications in decision analyses. Furthermore, the ordering developed in Appendix A will also be useful in analyzing situations in which one must decide when the distribution function F of the prospect X represents a more risky proposition than the distribution function G of prospect W.

In general, any modeling setup in which at least one component of the model is based on a cdf will benefit from the proposed framework. Examples include binary regression, ordinal regression and other setups where GLM are used.

5.5. Topics for future research

5.5.1. Handling multiple predictors

This paper introduces a framework that allows a decision maker to imprint his/her outlook on the analysis of a worst-case scenario with a *single* numerical covariate. The framework is based on ideas requiring that the observed values

of the covariate can be ordered unambiguously. To extend the framework to cases with two or more numerical covariates, one might start with the model at (1.3) and view it as a single-index model. In such a model, the natural ordering requirement would be imposed on the single index $\eta_i = \beta^T X_i$ instead of the single covariate. The ordering of η_i, which is one dimensional, becomes possible after estimating the unknown parameters in F_0. Extension of the model at (1.3) to more than one covariate then requires the use of a constrained non-parametric model in η_i (i.e., $F_1(\eta_i)$) over the extrapolation region of the covariates.

A similar approach is often used in the econometrics literature when testing for the appropriateness of a particular parametric structure F in equation (1.1). There, it is assumed that the single-index stays the same throughout the entire region of the covariates regardless of the appropriate structure (e.g., Horowitz and Hardle (1994)). In worst-case scenarios for which the covariates are risk factors, one can expect a higher estimated η_i for higher values of the covariates, allowing the use of Theorem 3.1 on $\hat{\eta}$ with minor modifications. The complete analysis with multiple covariates that accounts for all uncertainties will be considered elsewhere.

5.5.2. Small-sample performance

The statistical methodology developed in Section 3 for obtaining lower and upper confidence bounds appeals to asymptotic approximation. In large samples, one can expect the LCB to contain $F_1(x)$, $x \in J^*$, and the UCB to contain $X_\gamma \in J^*$, approximately $100(1-\rho)\%$ of the time. Future research should evaluate the small-sample performances of this methodology.

5.5.3. Benchmark analysis

A major field of application for the proposed framework is in quantitative risk analysis. Here the objective is to characterize the probability and severity of damage to humans caused by a chemical or biological agent. Such risk analysis is typically conducted on data obtained from animal bioassays in which rodents are exposed to relatively high doses of the agent. Estimating the risk for common levels of exposure, which are much lower, then requires extrapolation. A prime objective of such analysis is to determine the lower confidence bound for the dose that yields a specific bench-mark risk.

Appendix A.

Throughout what follows, the distribution functions considered are absolutely continuous and their densities are continuous and strictly positive with at least one continuous derivative on $(-\infty, \infty)$. In accordance with the ORT in

(2.1) and Definition 2.1, the focus is on distributions that can be labeled as *AR-*, *RR-* or *OR-pessimistic* (*optimistic*) on an interval J.

The following theorem provides the necessary and sufficient conditions for a distribution to be *AR-*, *RR-* or *OR-pessimistic* (*optimistic*) on an interval J. These conditions are essential for ordering distributions across the ORT in (2.1).

Theorem A.1. *A distribution function F with density f* is inherently

(i) *RR−pessimistic* (*optimistic*) *on J if, and only if,* $\log(F(x))$ *is convex* (*concave*) *for all $x \in J$ or, equivalently, the reverse hazard rate, $f(x)/F(x)$, is increasing* (*decreasing*) *for all $x \in J$;*

(ii) *OR−pessimistic* (*optimistic*) *on J if, and only if,* $\log(F(x)/(1-F(x)))$ (*i.e., the log-odds*)*, is convex* (*concave*) *for all $x \in J$ or, equivalently, the modified hazard rate, $f(x)/(F(x)-(1-F(x)))$, is increasing* (*decreasing*) *for all $x \in J$;*

(iii) *AR−pessimistic* (*optimistic*) *on J if, and only if, $F(x)$ is convex* (*concave*) *for all $x \in J$ or, equivalently, the reverse density function $f(x)$ is increasing* (*decreasing*) *for all $x \in J$.*

Proof. From *Definition* 3.1, if F is *RR−pessimistic* on J, then $F(x+\Delta)/F(x)$ is non-decreasing in x as long as $(x+\Delta) \in J$, for all $\Delta > 0$. This is equivalent to $\log(F(x+\Delta)/F(x))$ being non-decreasing in x as long as $(x+\Delta) \in J$. This implies that $d/dx\{\log(F(x+\Delta)\} - d/dx\{\log(F(x)\} \geq 0 \Leftrightarrow (f(x+\Delta))/(F(x+\Delta)) \geq f(x)/F(x)$ in x as long as $(x+\Delta) \in J \Leftrightarrow f(x)/F(x)$ is non-decreasing in $x \in J$. Thus, $\log F(x)$ is convex for all $x \in J$. The proofs of parts (ii) and (iii) are similar and are therefore omitted.

A.1. Ordering distributions using an ORT

To order two or more distributions on the same interval and with the same ORT, I turn to *univariate directional* orderings. Here the objective is to place, in some sense, one distribution to the right of the other. For a good account, see Shaked and Shanthikumar (1994).

In the discussion, I depart from common practice and require that these orderings hold only on a subinterval, rather than on the entire support of the distributions. This departure may lead to different results despite the usage of similar labels. The following well-known ordering is a good illustration:

Stochastic ordering on J: A random variable W is said to be stochastically larger than X on $J = [a, b]$ and denoted as $X \leq_{SO/J} W$, if $F(a) = G(a)$ and $F(u) \geq G(u)$ for all $u \in J = [a, b]$.
Obviously, in this case F represents a more pessimistic outlook on J than G since bad outcomes are more likely under F than under G.

The example of stochastic ordering demonstrates how directional ordering can be used to label one structure as more pessimistic than another. This direction is pursued and new orderings are proposed based on the ORT in equation

(2.1) of the paper. The relations among the suggested orderings and their relation to the stochastic order are presented below.

Definition A.1. Let X and W be two random variables taking values in R with distribution functions F and G and densities f and g, respectively. Then W is said to be larger than X

(i) in *relative risk order* on J and denoted as $X \leq_{RR/J} W$, if

$$F(a) = G(a) > 0 \text{ and } \frac{f(u)}{F(u)} \geq \frac{g(u)}{G(u)} \text{ for all } u \in J = [a, b];$$

(ii) in *odds ratio order* on J and denoted as $X \leq_{OR/J} W$, if

$$F(a) = G(a) > 0 \text{ and } \frac{f(u)}{F(u)(1-F(u))} \geq \frac{g(u)}{G(u)(1-G(u))} \text{ for all } u \in J = [a, b];$$

(iii) in *attributable risk order* on J and denoted as $X \leq_{AR/J} W$, if

$$F(a) = G(a) > 0 \text{ and } f(u) \geq g(u) \text{ for all } u \in J = [a, b].$$

The following theorem provides necessary and sufficient conditions for the above orderings. The conditions are analogous to the conditions in Theorem A.1.

Theorem A.2. *Let X and W be two random variables taking values in R with distribution functions F and G, respectively, such that $F(a) = G(a) > 0$. Then,*

(i) *$X \leq_{RR/J} W$ if, and only if, $F(u)/G(u)$ is increasing in $u \in J = [a, b]$;* (A.1)

(ii) *$X \leq_{OR/J} W$ if, and only if, $(F(u)/(1-F(u)))/(G(u)/(1-G(u)))$ is increasing in $u \in J = [a, b]$;*

(iii) *$X \leq_{AR/J} W$ if, and only if, $F(u) - G(u)$ is increasing in $u \in J = [a, b]$.*

Proof. From *Definition* A.1, when W is larger than X in *relative risk order* on J, then $F(a) = G(a) > 0$ and $f(u)/F(u) \geq g(u)/G(u)$ for all $u \in [a, b]$. This is equivalent to $d/dx \log F(u)/G(u) \geq 0$ for all $u \in [a, b] \Leftrightarrow F(u)/G(u)$ is non-decreasing in $u \in [a, b]$. The proofs of parts (ii) and (iii) are similar and are therefore omitted.

Remarks.

1. The condition in (A.1) appears in Lehmann and Rojo (1992) as a technical condition. Moreover, the corresponding *relative risk order*, on the *entire* support, appears in Shaked and Shanthikumar (1994) under the name of *reverse hazard order* but with no connection to the relative risk order or the proposed outlooks framework.

2. To the best of my knowledge, the ordering $X \leq_{OR/J} W$ and $X \leq_{AR/J} W$, including the necessary and sufficient conditions given here, do not appear elsewhere.
3. It follows from Theorem A.2 that, when $F(a) = G(a) > 0$, then
   (i) $X \leq_{RR/J} W \Rightarrow X \leq_{OR/J} W \Rightarrow X \leq_{SO/J} W$;
   (ii) $X \leq_{RR/J} W \Rightarrow X \leq_{AR/J} W \Rightarrow X \leq_{SO/J} W$.
   Note, therefore, that each of the orderings introduced here implies a *stronger* ordering than that of the stochastic order and thus conveys a higher degree of pessimism.
4. If, on the interval $J = [a, b]$, F and G also satisfy $F(b) + G(b) \leq 1$, then

$$X \leq_{RR/J} W \Rightarrow X \leq_{OR/J} W \Rightarrow X \leq_{AR/J} W \Rightarrow X \leq_{SO/J} W.$$

This section closes with an important characterization of each of the three most commonly used models in binary regression analyses (i.e., the logit, the probit and the Cloglog). In the theorem below, M_N denotes the median of the normal distribution, which corresponds to the probit model. Note that the results for the logit and the Clog-log models hold for any subinterval $J \subset (-\infty, +\infty)$.

Theorem A.3. *On the subinterval $J \subset (-\infty, M_N]$, we have:*
(i) *the probit is the only model among the three where the underlying distribution is OR-optimistic;*
(ii) *the logit is the only model among the three where the underlying distribution is OR-neutral;*
(iii) *the Clog-log is the only model among the three where the underlying distribution is OR-pessimistic.*

Proof. Part (i) of the theorem is quite involved and appears as *Corollary* 1 in a technical report that can be obtained from the author. Part (ii): for a distribution function F to be *OR-neutral* on R it has to be both *OR-pessimistic* and *OR-optimistic* on R. From *Theorem* A.1 part (ii), $\Psi(x) = d/dx \log(F(x)/(1 - F(x)))$ must be both non-decreasing and non-increasing on R. That is, $\log(F(x)/(1 - F(x)))$ is linear in $x \in R$. It follows from *Theorem* 2.1.5 of Galambos and Kotz (1978, p. 27) that the logistic family is completely characterized by this property. Part (iii): without loss of generality, take $\alpha = 0$ and $\beta = 1$. Note that

$$\frac{f(x)}{F(x)(1 - F(x))} = \frac{e^x}{F(x)} \text{ and } \phi(x) = \frac{d}{dx} log \frac{e^x}{F(x)} = 1 - \frac{f(x)}{F(x)}.$$

When $-\infty < x \leq \infty$, $\phi(x) \geq 0$ since $F(x) - f(x) \geq 0$ and, as $x \to -\infty$, $\phi(x) = 0$ since $\lim_{x \to -\infty} f(x)/F(x) \to 1$.

In summary: The three most commonly used models in binary regressions, which often fit data *equally well* (e.g., the application in 4.4), actually convey

very different and *opposite* outlooks regarding the likelihood of adverse outcomes. Therefore, choosing one model over another for F_1 when extrapolating onto an interval J, amounts to taking a different outlook position when making decisions in worst-case analyses.

Appendix B.

Proof of Theorem 3.1. Consider the case where F_1 is *AR-pessimistic* on J. From *Theorem* A.1 part (iii) it follows that $F_1(x)$ is convex in $x \in J$. Therefore, there exists a line $\alpha + \beta x$ such that $F_1(x) \geq \alpha + \beta x$ for all $x \geq x_q$ where α and β are the solutions of $F_1(x_p) = \alpha + \beta x_p$ and $F_1(x_q) = \alpha + \beta x_q$. Solving these equations gives $\beta = (F_1(x_q) - F_1(x_p))/(x_q - x_p)$ and $\alpha = F_1(x_p) - \beta x_p$. With a little algebra,

$$\alpha + \beta x = \frac{x - x_p}{x_q - x_p} F_1(x_q) - \frac{x - x_q}{x_q - x_p} F_1(x_p) = A F_1(x_q) - (A-1) F_1(x_p)\,.$$

From the assumption that F_{01} is stochastically larger than F_1 on J, comes the inequality

$$A F_1(x_q) - (A-1) F_1(x_p) \geq A F_{01}(x_q) - (A-1) F_{01}(x_p)\,.$$

The proofs for when F_1 is *OR-pessimistic* or *RR-pessimistic* on J are similar and therefore omitted.

Proof of Corollary 3.1. Equation (3.2) is derived by inverting the equation (3.3) of Theorem 3.1.

Proof of Proposition 3.1. Under the conditions that, on $J = [X^{U^-}, X^*]$, F_1 is *pessimistic* with respect to an ORT and, on $J_0 = [X^{U^-}, X^U]$, F_{01} is stochastically larger than F_1, it follows from *Corollary* 3.1 that X_γ is bounded from above by

$$G(\alpha, \beta; \gamma, x_q, x_p) = \frac{(x_q - x_p) g(\gamma) + g(F_0(\alpha + \beta x_q)) x_p - g(F_0(\alpha + \beta x_p)) x_q}{g(F_0(\alpha + \beta x_q)) - g(F_0(\alpha + \beta x_p))}.$$

Let $P_i = P(Y_i = 1 | X = x_i; \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_t)^T \in \Omega$, where the parameter space Ω is an open subset of t-dimensional Euclidean space. Consider the following common regularity conditions.

C1: $\lim_{n\to\infty} n_j/n = c_i$ $(0 < c_i < 1)$ for all $i = 1, 2, \ldots, K$, and,

C2: K must be at least as large as the number of parameters in the model, where K denotes the number of different X values in the data.

C3: The information matrix $\boldsymbol{\Sigma}^{-1} = ((\sigma^{\mathbf{ls}}))$ defined by

$$\sigma^{ls} = \sum_{i=1}^{k} \frac{c_i \dfrac{\partial P_i}{\partial \theta_l} \dfrac{\partial P_i}{\partial \theta_s}}{P_i(1 - P_i)}, \quad (l, s = 1, 2, \ldots, t)$$

is positive definite.

It is not hard to show (Cox and Hinkley (1974)) that, under C1-C3, the likelihood ratio statistic is asymptotically a χ^2_t distribution where t is equal to the number of parameters in the model.

Using standard arguments on a constrained parameter space (Rao (1973, p. 419)), the approximate $(1-\rho)\%$ UCB for $X_\gamma \in J^* = (X^U, X^*]$ is given by

$$L_\gamma = \sup_{(\alpha,\beta)} \{X_\gamma = G(\alpha,\beta;\, \gamma, x_q, x_p) : 2(LL_1(\alpha,\beta) - LL(\alpha,\beta)) \leq \chi^2_{1,1-2\rho}\}\,.$$

Note that $\chi^2_{1,1-2\rho}$ is used instead of $\chi^2_{1,1-\rho}$ because a one-sided rather than a two-sided confidence limit on X_γ is required (Fleiss (1973, pp. 20-21)).

Acknowledgements

The author is deeply grateful to the editors, the associate editors and several anonymous referees for their inputs and suggestions.

References

Agresti, A. (2001). Exact inference foe categorical data: recent advances and continuing controversies. *Statist. Medicine* **20**, 2709-2722.

Alho, J. M. and Valtonen, E. (1995). Interval estimation of inverse dose-response. *Biometrics* **51**, 491-501.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26**, 641-647.

Barlow, R. E. and Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing*. To Begin With, Silver Springs.

Berry, G. and Armitage, P. (1995). Mid-P confidence intervals: a brief review. *The Statistician* **44**, 417-423.

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: a integral part of inference. *Biometrics* **53**, 603-618.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-model Inference*. Springer, New York.

Chambers, R. and Cox, D. R. (1967). Discrimination between alternative binary response models. *Biometrika* **54**, 573-578.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *J. Roy. Statist. Soc. Ser. A* **158**, 419-466.

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404-413.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statist. Sci.* **19**, 81-94.

Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete-data bias. *J. Roy. Statist. Soc. Ser. B* **67**, 459-513.

Coombs, C. H. (1964). *A Theory of Data*. Wiley, New York.

Cox, D. R. (1972). Regression models and life tables (with discussion) *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *J. Amer. Statist. Assoc.* **84**, 945-957.

Diecidue, E. and Wakker, P. P. (2001). On the intuition of rank-dependent utility. *J. Risk Uncert.* **23**, 281-298.

Diggle, P. J. (2005). in the discussion of "Local model uncertainty and incomplete-data bias" by Copas and Eguchi. *J. Roy. Statist. Soc. Ser. B* **67**, 495-496.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 45-97.

Draper, D., Saltelli, A, Tarantola, S. and Prado, P. (2000). Scenario and parametric sensitivity and uncertainty analysis in nuclear waste disposal risk assessment: the case of GESAMAC. Ch. 13 in *Mathemtaical and Statistical Methods for Sensitivity Analysis*, Wiley, New York.

Fleiss, J. L. (1973). *Statistical Methods for Rates and Proportions*. Wiley, New York.

Fuh, C. and Hu, I. (2004). Efficient importance sampling for events of moderate deviations with applications. *Biometrika* **91**, 471-490.

Fygenson, M. (1997). A new approach in modeling a categorical response. Part I: The binary case. *J. Amer. Statist. Assoc.* **92**, 322-332.

Galambos, J. and Kotz, S. (1978). *Characterizations of Probability Distributions (Lecture Notes In Mathematics* 675). Springer-Verlag, New York.

Greenland, S. (2005). Multiple-bias modeling for analysis of observational data (with discussion). *J. Roy. Statist. Soc. Ser. A* **168**, 267-306.

Horowitz, J. L. and Hardle, W. (1994). Testing a Parametric Model against a Semiparametric Alternative. *Econom. Theory* **10**, 821-848.

Horowitz, J. L. and Manski, C. (2000). Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. *J. Amer. Statist. Assoc.* **95**, 77-84.

Huang, Y. (2001). Various methods for interval estimation of the median effective dose. *Comm. Statist. Simulation Comput.* **30**, 99-112.

Huber, P. J. (2003). *Robust Statistics*. John Wiley & Sons, Hoboken.

Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programing. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (Edited by J. Neyman), University of California Press, Berkeley.

Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle O-ring data. *J. Amer. Statist. Assoc.* **86**, 919-922.

Lehmann, E. L. and Rojo, J. (1992). Invariant directional ordering. *Ann. Statist.* **20**, 2100-2110.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, second edition*. Wiley, New York.

Lehmann, E. L. and Romano, J. (2005). *Testing Statistical Hypotheses*. Springer, New York.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statist. Medicine* **17**, 857-872.

Quiggen, J. (1981). Risk perception and risk aversion among Australian farmers. *Austr. J. Agri. and Econ.* **25**, 160-169.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251-266.

Raftery, A. E., Madigan, D. and Hoeting, J. (1997). Accounting for model uncertainty in linear regression. *J. Amer. Statist. Assoc.* **92**, 179-191.

Rao, C. R. (1973). *Linear Statistical Inferences and its Applications*. Wiley, New York.

Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91**, 153-164.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* **63**, 581-592.

Shaked, M. and Shanthikumar, J. G. (1994). *Stochastic Orders and Their Applications*. Academic Press, San Diego.

Tversky, A. and Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *J. Risk Uncert.* **5**, 297-323.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.* **22**, 209-212.

Marshall School of Business, University of Southern California, Los Angeles, CA 90089, U.S.A.
E-mail: mfygenson@marshall.usc.edu

(Received May 2006; accepted February 2007)

COMMENTS

I-Shou Chang[1,2], Li-Chu Chien[2] and Chao A. Hsiung[2]

[1]*Institute of Cancer Research and* [2]*National Health Research Institutes, Taiwan*

Fygenson presents a new approach to extrapolation problems in the face of model uncertainty; the presentation is made in the context of a binary regression. Let Y denote the binary variable and X a numerical covariate. Assume the observed covariate values are contained in (X^L, X^U), the problem is to estimate $P(Y = 1|X = X^{Extp})$, where $X^{Extp} > X^U$.

Instead of using (1.3) in Fygenson, which is at the crux of his approach, we would like to propose an alternative approach that considers a Bayesian binary regression model with a prior induced by Bernstein polynomials. Bernstein polynomials have been shown to be a useful modeling tool in Bayesian shape-restricted inference by Chang, Hsiung, Wu and Yang (2005) and Chang, Chien, Hsiung, Wen and Wu (2007), because priors introduced by Bernstein polynomials can have large support, select only smooth functions, and easily incorporate geometric information into the prior.

Let $\varphi_{i,n}(t) = C_i^n t^i (1-t)^{n-i}$ for $t \in [0,1]$. Let $F_{b_n} = \mathbf{F}(n, b_{0,n}, \ldots, b_{n,n}, t) = \sum_{i=0}^n b_{i,n}\varphi_{i,n}(t/\tau)$, with $b_n = (b_{0,n}, \ldots, b_{n,n}), (n, b_n) \in \mathcal{B} = \bigcup_{n=1}^\infty (\{n\} \times \mathbb{R}^{n+1})$, and $0 \le t \le \tau$. We note that F_{b_n} is a Bernstein polynomial with coefficients $b_{0,n}, \ldots, b_{n,n}$. It is readily seen that if $b_{0,n} \le b_{1,n} \le \cdots \le b_{n,n}$, then $F_{b_n}(\cdot)$ is an

increasing function with $b_{0,n} = F_{b_n}(0) \le F_{b_n}(\tau) = b_{n,n}$. Let $I_n = \{F_{b_n} \mid b_n \in \triangle^{n+1}\}$, where $\triangle^{n+1} = \{0 \le b_{0,n} \le \cdots \le b_{n,n} \le 1\}$. We will work with $\bigcup_{n=1}^{\infty} I_n$.

We now assume that

$$P(Y_{jk} = 1 | X_k = x) = F(x), \tag{7.1}$$

for some F in $\mathcal{I}$, the set of all increasing continuous functions on [0,τ] with values in [0,1]. Here $k = 1, \ldots, K$ and $j = 1, \ldots, m_k$. By considering a probability measure on $\mathcal{I}$, we have a Bayesian binary regression model. It is clear that $\mathcal{I} \supset \bigcup_{n=1}^{\infty} I_n$. A probability measure π can be introduced on $\mathcal{I}$ as follows. Let π_n be a conditional density on $\triangle^{n+1}$ and p a probability mass function on $\{1, 2, \ldots\}$; define $\pi(n, b_n) = p(n)\pi_n(b_n)$, which is a probability measure on $\bigcup_{n=1}^{\infty}(\{n\} \times \triangle^{n+1})$. Identifying a Bernstein polynomial with its order and coefficients, we can regard π as a probability on $\bigcup_{n=1}^{\infty} I_n$, hence on $\mathcal{I}$. Priors of this form are referred to as Bernstein priors. We note that the conditional density $\pi_n(\cdot)$ is a critical part of the prior.

For inference purpose, we can use MCMC to sample the posterior density ν of the parameter (n, b_n); it is proportional to

$$\prod_{k=1}^{K} \prod_{j=1}^{m_k} (F_{b_n}(X_k))^{Y_{jk}} (1 - F_{b_n}(X_k))^{1-Y_{jk}} \pi_n(b_n) p(n).$$

We now consider the problem of extrapolation under (1) in two situations. In the Bayesian decision theory framework, one starts with a loss function that summarizes the outlooks of the decision maker, and uses the corresponding Bayes estimate to extrapolate the probability for decision making. In this situation, we will see in the following that Bernstein priors offer the flexibility to model a suitable tail behavior for extrapolation.

In the more vague non-decision-theoretic Bayesian approach, one might ignore the loss function and consider the posterior mean or median as the estimate. In this case, we will see that a Bernstein prior represents a convenient tool to incorporate one's subjective belief or subject-matter knowledge in the model for inference. We note that relevant discussions on statistical inferences and Bayesian decision theory can be found, for example, in Berger (1985).

We now illustrate the use of a Bernstein prior in the context of the 1986 Challenger data. According to Fygenson, the following information can be used to conduct statistical analysis before the decision to launch the Challenger on January 28, 1986.

(a) The engineering judgment that the probability of erosion rises monotonically as temperature decreases.

(b) The number of primary O-ring failures (out of 6) as a function of launch temperature for each of the previous 23 space shuttles launches.
(c) All previous launches took place between 53 and 81 degrees. Thus, to predict the probability of failure at the anticipated launch temperature of 31 degrees requires a sizable extrapolation.

Let X denote the temperature and $\tilde{X} = (81 - X)/50$. Let $Y = 1$ indicate O-ring failure. Before January 28, 1986, the $6 \times 23 = 138$ independent binary observations are taken at 16 different temperature levels specified by $(\tilde{X}_0, \tilde{X}_1, \cdots, \tilde{X}_{15})$=(0.00, 0.04, 0.06, 0.10, 0.12, 0.16, 0.18, 0.22, 0.24, 0.26, 0.28, 0.30, 0.36, 0.46, 0.48, 0.56). The task is to estimate $P(Y = 1|X = 31) = P(Y = 1|\tilde{X} = 1)$.

The Bernstein prior can be specified as follows. Let $p(1) = e^{-\alpha} + \alpha e^{-\alpha}$, $p(n) = \alpha^n e^{-\alpha}/n!$ for $n = 2, \ldots, n_0 - 1$, and $p(n_0) = 1 - \sum_{n=1}^{n_0-1} p(n)$. We note that larger α and n_0 make the prior less informative. Let q_1 be $Uniform(q_{11}, q_{12})$ with support containing $F(0)$, and q_2 be $Uniform(q_{21}, q_{22})$ with support containing $F(1)$; generate a_0 from q_1 and a_n from q_2. Let $a_1 \leq a_2 \leq \cdots \leq a_{n-1}$ be the order statistics of a random sample from $Uniform(a_0, a_n)$; the conditional distribution of π on $\triangle^{n+1}$ is defined to be that of $(a_0, a_1, \ldots, a_n)$. Here q_{11}, q_{12}, q_{21} and q_{22} are defined as follows.

Since $P(Y = 1|\tilde{X}) \leq 1$, we let $q_{22} = 1$. Since the empirical probabilities $P(Y = 1|\tilde{X} = \mathrm{x}) = 0$ or $1/6$ for $\mathrm{x} = 0.00, 0.04, \ldots, 0.22$, we set $q_{11} = 0$ and $q_{12} = 1/6$.

We now specify q_{21}. Consider the largest, say five, $\tilde{X}$'s that have shuttle launches, 0.30, 0.36, 0.46, 0.48, and 0.56, and the corresponding empirical probabilities $P(Y = 1|\tilde{X})$, respectively 0, 1/6, 1/6, 1/6, 1/3. These give the five data points $(0.30, 0)$, $(0.36, 1/6)$, $(0.46, 1/6)$, $(0.48, 1/6)$, $(0.56, 1/3)$, denoted $(\mathrm{x}_1, \mathrm{p}_1), \ldots, (\mathrm{x}_5, \mathrm{p}_5)$. For $i > j$, let $S_{ij} = (\mathrm{p}_i - \mathrm{p}_j)/(\mathrm{x}_i - \mathrm{x}_j)$ denote the slope between $(\mathrm{x}_i, \mathrm{p}_i)$ and $(\mathrm{x}_j, \mathrm{p}_j)$. Since the belief is that the O-ring failure probability increases with $\tilde{X}$, it is reasonable to define q_{21} in terms of the empirical probability at $\tilde{X} = 0.56$, $P(Y = 1|\tilde{X} = 0.56) = 1/3$, and the S_{ij}. We take $q_{21} = 1/3 + (1 - 0.56) \times S$ for some S in $\{S_{ij} \mid 1 \leq j < i \leq 5\} = \{0.00, 0.83, 0.93, 1.04, 1.28, 1.67, 2.08, 2.78\}$.

This suggests that, in the case of Bayesian decision theory, we may let $q_{21} = 1/3$ to ensure model flexibility. In the case of the non-decision-theoretic Bayesian approach, we take $q_{21} = 1/3$ for an optimistic outlook, $q_{21} = 1/3 + (1 - 0.56) \times 0.83 = 0.72$ for a neutral outlook, and $q_{21} = 1/3 + (1 - 0.56) \times 1.28 = 0.90$ for a pessimistic outlook.

On the basis of these pessimistic, neutral and optimistic outlooks, the prior distributions are generated using the Bernstein prior with $\alpha = 10$, $n_0 = 20$ and the corresponding q_{21}; for each prior, the posterior mean is calculated using

100,000 MCMC iterations with a burn-in period of 10,000 iterations. Based on these means, we can compute the expected numbers of O-ring failures in a given launch. Figure 1 reports the expected number of failures (out of 6) at different temperature with dotted, dashed and solid lines for pessimistic, neutral and optimistic outlooks, respectively; circles are the observed data points.

In summary, we find the Bayesian model with a Bernstein prior to be a useful alternative for this type of problem.

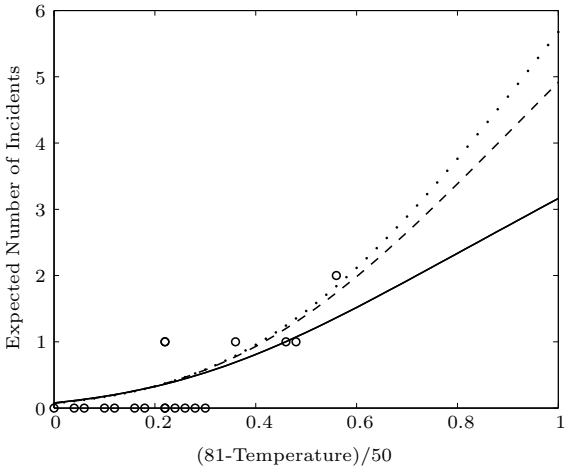

Figure 1. 1986 Challenger data fit with a Bayesian binary regression model with prior induced by Bernstein polynomials.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, second edition. Springer-Verlag, New York.

Chang, I. S., Chien, L. C., Hsiung, C. A., Wen, C. C. and Wu, Y. J. (2007). Shape restricted regression with random Bernstein polynomials. In *IMS Lecture Notes — Monograph Series Complex Datasets and Inverse Problems* **54** (Edited by R. Liu, W. Strawderman and C. H. Zhang), 187-202.

Chang, I. S., Hsiung, C. A., Wu, Y. J. and Yang, C. C. (2005). Bayesian survival analysis using Bernstein polynomials. *Scand. J. Statist.* **32**, 447-466.

Institute of Cancer Research and Division of Biostatistics and Bioinformatics, National Health Research Institutes, 35 Keyan Road, Zhunan Town, Miaoli County 350, Taiwan.
E-mail: ischang@nhri.org.tw

Division of Biostatistics and Bioinformatics, National Health Research Institutes, 35 Keyan Road, Zhunan Town, Miaoli County 350, Taiwan.
E-mail: lcchien@nhri.org.tw

Division of Biostatistics and Bioinformatics, National Health Research Institutes, 35 Keyan Road, Zhunan Town, Miaoli County 350, Taiwan.
E-mail: hsiung@nhri.org.tw

COMMENTS

Cheng-Der Fuh[1,2] and Inchi Hu[3]

[1]*National Central University,* [2]*Academia Sinica*
and [3]*Hong Kong University of Science and Technology*

We would like to thank Professor Fygenson for an important and interesting paper. The proposed approach addresses both parametric and structural uncertainties in statistical models. In order to enrich the topic, we would like to comment on four possible applications and extensions.

High Quantile. In Theorem 3.1, it is assumed that F_1 is convex in the interval J for AR. For RR and OR it is convex after taking log and log-odds transformations, respectively. For commonly used models, such as logit and probit, this implies that J is located before the inflection point of $F_1(\cdot)$. Suppose that the probability we want to extrapolate is after the inflection point, that is, $F_1(\cdot)$ is concave in J. Does this mean that we cannot use logit and probit models anymore? Note that in VaR (value at risk) and pyrotechnics, the probabilities that we would like to predict are all very close to one. Perhaps we can map these probabilities to those very close to zero to circumvent the problem. However, J is then located to the left of the interval in which we have observations. Do the arguments in the proof of Theorem 3.1 still work in the same way? Further, if J straddles the inflection point, we note that no monotone transform of the data can alleviate the problem.

Degradation Analysis. When examining the reliability of high tech products, we are interested in estimating a low quantile in order to provide a proper warranty time. If a manufacturer produces 1,000,000 items and 1% of them fail before the warranty time, then 10,000 customers will buy a product that does not achieve the announced quality and this will tarnish the reputation of the manufacturer.

High tech products are usually very reliable and hence failures are rare. Degradation analysis is one of the methodologies used to overcome the difficulty of assessing the reliability of highly reliable products in limited time. The essential idea is to assess the lifetime T is based on the equation $P(T \geq t) = P(D(t) \leq \tau)$, where $D(t)$ is the actual degradation measure at time t, and τ is the threshold assuming that $D(t)$ is increasing in t.

The physical or chemical property of the degradation measurement is sometimes very complicated and it is not easy to find the functional form of $D(t)$. Empirical models provide an alternative solution. After degradation data is collected we can find an empirical model that fits the data well. However, degradation analysis is basically an extrapolation and the result is highly dependent

on $D(t)$. Corollary 3.1 provides a lower bound for the lower quantile with x_p and x_q estimated from the empirical model. An interesting question is whether we can apply the result in the case of predicting the quantile (median for instance) of T.

Importance Sampling. In importance sampling, a tail event, which corresponds to extrapolation in the paper, becomes a central event under the alternative distribution. It may be possible to view the extrapolation problem as an importance sampling problem, where the value we would like to extrapolate corresponds to an interpolation problem under the alternative distribution. In principle, the variance, and hence the confidence bounds for the tail event, can be worked out via the likelihood ratio under the alternative distribution. Maybe the constrained optimization problem in Theorem 3.1 and Corollary 3.1 can be transformed into one that involves the likelihood ratio under importance sampling.

Sequential Design. In computerized adaptive testing (CAT), when constructing an individualized test, the ability (θ) estimate is updated after the administration of each item, and the next optimal item is selected from an item bank until a prespecified number of items is administered. Items are selected to match the examinee's estimated θ according to an item response theory (IRT) model (probit or logit model) that is assumed to describe an examinee's response behavior. The standard approach has been to next select the item with the maximum Fisher information at the examinee's current estimated ability level. This method is the so-called recursive maximum likelihood estimation (R-MLE) method, cf., Lord (1980, pp.151-153).

Although R-MLE has been studied and used mostly in item response theory, the uncertainties of the likelihood function due to model misspecification and measurement errors, especially at early stages, makes it less than statistically efficient. The paper raises the interesting question of whether there is an analogous theory and method for this sequential design problem.

Acknowledgement

This research was partially supported by NSC of ROC Grant NSC 96-2118-M-008-004 for first author. This research partially supported by Hong Kong RGC Grant HKUST6212/04H for second author.

References

Lord, M. F. (1980). *Applications of Item Response Theory to Practical Testing Problem.* Lawrence Erlbaum, Hillsdale, New Jersey.

National Central University, Jhongli City, Taoyuan County 32001, Taiwan, R.O.C.
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C.
E-mail: stcheng@stat.sinica.edu.tw
Department of Information and Systems Management, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.
E-mail: imichu@ust.hk

COMMENTS

Peter McCullagh

University of Chicago

This paper tackles the classical problem of extrapolation, i.e., of predicting the response Y^* at a point x^* in the covariate space remote from previous experimental or observational values. The problem arises in various areas such as the estimation of risk to individuals exposed to low levels of arsenic in water, the estimation of risk to individuals exposed to pesticides in fruit and vegetables, or the estimation of risk from exposure to second-hand cigarette smoke. The more spectacular area of application is the prediction of major catastrophes or cataclysmic one-time events. It is a truism that all incorrect predictions of the latter type (Y2K, Second coming,...) are made before the non-event and soonest forgotten; the great majority of correct predictions (Challenger, Chernobyl, Katrina, 9/11,...) are made after the event. Despite this sorry record, or perhaps because of it, the prediction of catastrophes has widespread appeal that is hard to resist. And who is better equipped to prognosticate than a statistician?

Without a mathematical model there is no framework to tackle the problem. With a stochastic model p the task is easy: the conditional distribution $p(y^* \mid \text{data}, x^*)$ is used to compute the prognostic probability for any desired event at x^*. With a parametric statistical model $\{p_\theta\}$ and a prior distribution π, the answer $p(y^* \mid \text{data}, x^*)$ is obtained in the same way using the mixture $p = \int p_\theta d\pi(\theta)$. In the absence of a prior distribution, the approximation $p_{\hat{\theta}}(y^* \mid \text{data}, x^*)$ is frequently used, sometimes in a modified form making allowance for errors of estimation.

Where is the problem? The first problem is that all useful models are based on assumptions that are compatible with the data but unverifiable from the data. Furthermore, two models that are equally consistent with the data may give very different predictions on extrapolation. The second problem is that statisticians, Bayesian or otherwise, are seldom sufficiently confident of their models to trust their predictions on extrapolation. The solution offered by Fygenson for binary

extrapolation is to select a few model classes on the basis of their outlook or degree of pessimism on extrapolation. Fygenson's elaborate partial order of pessimism, may be effective for administrative and managerial purposes where risk calculations are important and the costs highly asymmetric. But in such cases, the real problem is to strike a balance between costs and benefits, so both cost and benefit should be quantified and included as key components in the model. For that reason I found Fygenson's outlook less than satisfying.

Classical parametric models for binary regression with a real-valued covariate x are based on three structural assumptions: (i) independence of components; (ii) monotonicity of the probability $\mathrm{pr}(Y = 1; x) = F(\beta x)$ as a function of x; (iii) limits of zero and one as $x \to \pm\infty$. In practice, the choice for F seldom deviates from logit, probit or complementary log-log. Structural uncertainty is a regrettable term because the structure is probabilistic, and no model leaves room for uncertainty in probabilities. Fygenson's use of the terms structure and structural uncertainty refers solely to the choice of F in (ii), the remaining structural components, independence, monotonicity and limits, being ignored or taken for granted. If it were true as claimed that Lavine's (1991) non-parametric model is 'free from structure uncertainty', the model would be useless. In fact, Lavine's model assumes both monotonicity and independence, without which prediction would be impossible. Certainly it is reasonable to use a mixture of plausible response functions for extrapolation, but it does not seem reasonable in the Challenger example for the limit to be exactly zero at very high temperatures, or exactly one at very low temperatures. Furthermore, independence might be questioned if information were made available about the date or method of manufacture of individual O-rings. All of these are part of the model structure whether or not they have an appreciable effect on prediction at 31°C.

Given that the paper starts out with a principled Bayesian tone, the arbitrariness of the subsequent development seems strange. It is important logically to separate the probability of a catastrophic event from the cost of that event. Whether or not it is easy to assess, the cost should have no bearing on the probability, but the cost may have a big bearing on the decision reached. My impression is that the conflation of these themes in the paper leads to confusion.

Acknowledgement

Support for this research was provided in part by NSF grant No. DMS-0305009.

References

Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle O-ring data. *J. Amer. Statist. Assoc.* **86**, 919-922.

Department of Statistics, University of Chicago, 5734 University Ave, Chicago, Il 60637, U.S.A.
E-mail: pmcc@galton.uchicago.edu

COMMENTS

Stephen Portnoy

University of Illinois at Urbana-Champaign

In is indeed a pleasure to be invited to discuss Professor Fygenson's paper. When my colleague, Xuming He, first brought this work to my attention, I can still remember remarking on the courage of someone who would try to tackle the seemingly impossible problem of extrapolation. As I read the paper there were several bright ideas that I found appealing. First, I was familiar with the fact that if one takes a relatively restrictive but still nonparametric class of cdf's and places a small number (say two) linear restrictions on the distributions, then the range of the values of such constrained distributions at a fixed point can be rather small. Furthermore the bounds on the range can often be easily determined by a (simple) finite dimensional optimization. A specific version of this for arbitrary scale mixtures of normal distributions appears in Efron and Olshen (1978), and represents a generalized version of the method of moment spaces that goes back to work in the 1950's (if not earlier; see Karlin and Studden (1966) for example). The earliest statistical application I am familiar with is the optimal regression design result of Elfving (1952). In a joint paper with John Collins (1981), we applied the ideas in some problems in robustness theory, and also generalized the result of Efron and Olshen. Thus, I was not surprised to find that reasonable bounds could be established by defining an appropriately restricted nonparametric family of tail distributions, and then fixing these distributions to agree at two points nearer the center of the data where parametric models might be felt to be adequate. Nonetheless, the development of what seemed to be quite reasonable approaches to defining the family of tail distributions and the narrowness of the interval at the rather extreme extrapolation in the Challenger data were truly remarkable and impressive to me.

Despite my amazement, however, a closer inspection of the paper did suggest some questions, and one set of misgivings. On the applied side, I was surprised at the narrowness of the confidence bounds on the median temperature for the Challenger data. These bounds suggested little more variation at extrapolated values than for observed temperatures. Specifically, I did a quick analysis of the Challenger data using the R-function `glm` and found the confidence intervals

for the response at the two fitted points to be rather wide. This is just what I would have expected in a problem with the somewhat small sample size of the Challenger data (17 binomial(6) observations). It would seem that any reasonable range of these interpolated values would lead to much more additional variation than suggested in the paper. On the purely theoretical side, I wondered whether the word "pessimistic" was really appropriate, and if so was how conservative was the definition. That is, I wondered if there was a more objective way to calibrate the amount of "pessimism" in the various assumptions. Section 1 presents my naive approach to these issues, with mathematical details given in Section 2. Section 3 proffers some closing comments on the approach and the Challenger Data.

1. Estimation Variability and Calibration

My misgivings above suggest the value of a more formal exploration of the size of confidence bands for the extrapolated fitted response probability in the Challenger data. It seemed difficult to me to optimize over the classes of "pessimistic" tail distributions defined in the paper, and so I chose to consider scale mixtures of some some familiar distributions. Specifically, I decided to take three possibilities for the family of tail distributions, all of which are defined on $(0, \infty)$: scale mixtures of a negative exponential distribution, scale mixtures of a logistic distribution for $\log(z)$, and scale mixtures of a normal distribution for $\log(z)$. Here z is the variable (81-temp) (as in the paper). In each case, I fixed the distributions to agree with values of the conditional distribution of the response, given that it exceeds ($z = 24$), given by the `glm` solution for the logistic binary response model at $z = 26$ and $z = 28$. I then sought to extrapolate the probability of a "failure" event at $z = 52$.

Since this does differ from what Professor Fygenson did, some explanations are in order. First, while Professor Fygenson's focus on the median temperature is not at all unreasonable (a "quantile" person like myself can hardly criticize it), the failure probability at $z = 52$ seems more direct and immediately interpretable. The use of the scale mixture families also has some advantages: for fixed parameter values, the extrapolation bounds have closed form expressions in the negative exponential and logistic cases, and for the normal they are the result of a very simple numerical solution for the zero crossing of a one-dimensional monotonic function on a fixed interval. Thus, it is especially easy to find confidence bounds on the fitted value at $z = 52$. So laziness can provide some justification (if nothing else seems convincing). These families are also relatively restrictive but still nonparametric classes (in the sense that the mixing distributions are arbitrary, and thus the families are infinite-dimensional). As a consequence, they

might help to provide some of the calibration I sought in the second question above.

My basic approach is as follows. First, I fit the Challenger data using the "default" logistic regression of the R-function, `glm`, which gave the parameter estimates $\hat{\alpha} = -5.75$ and $\hat{\beta} = 0.170$. From this model, I fit the response at $z = 24, 26, 28$ and considered the conditional distribution of $Z^* \equiv Z - 24 \,|\, Z > 24$. This provided a cdf, F^* for Z^* on the interval $(0, \infty)$ with

$$F^*(2) \equiv r_1 = \frac{fit(26;\, \hat{\alpha},\, \hat{\beta}) - fit(24;\, \hat{\alpha},\, \hat{\beta})}{1 \,-\, fit(24;\, \hat{\alpha},\, \hat{\beta})}, \tag{7.2}$$

$$F^*(4) \equiv r_2 = \frac{fit(28;\, \hat{\alpha},\, \hat{\beta}) - fit(24;\, \hat{\alpha},\, \hat{\beta})}{1 \,-\, fit(24;\, \hat{\alpha},\, \hat{\beta})}. \tag{7.3}$$

For each of the scale-families above, I then considered arbitrary scale mixtures that matched r_1 and r_2 (at $z^* = 2$ and $z^* = 4$) and predicted the probability at the take-off temperature, which corresponds to $z^* = 26$. I then minimized over all scale-mixing distributions matching r_1 and r_2, and thus calculated a (family-wise) lower bound $\hat{L}_i$ for the probability of "failure" at $z^* = 26$. Here, $i = 1, 2, 3$ denotes each of the three families above (negative exponential, logistic-in-log, and normal-in-log). The details for these calculations are provided in Section 2. These families provided lower bounds somewhat below the "pessimistic" bounds of Professor Fygenson (see Table 1).

To assess the statistical variability of these bounds, I considered two methods for sampling the logistic regression parameters (α and β). One was to use the asymptotic normal approximation with mean $(\hat{\alpha}, \hat{\beta})$ and covariance matrix given by the R-function `summary.glm`. The other invoked the bootstrap, based on resampling the 17 (z, Y) pairs, where z is the temperature, and Y is the corresponding binomial observation in the Challenger data set. In each of these two situations, I took 1,000 random samples, and for each sample, recalculated the lower bounds L_i. I then produced a lower confidence bound on the lower bound, either using the lower 0.05 or 0.10 quantile for samples from the asymptotic normal distribution, or using the appropriate bootstrap value: $\hat{L}_i - 2L_i^*(q)$, where $L_i^*(q)$ is the upper $q = 0.95$ or $q = 0.9$ quantile of the bootstrap distribution for the lower bound. These values are listed in Table 1.

Table 1. Lower Bound at z = 26, with confidence bounds.

	L at est.	Samp 0.05	Samp 0.1	Boot 0.05	Boot 0.1
neg exp	0.457	0.208	0.241	0.337	0.365
logis(log)	0.428	0.211	0.247	0.158	0.198
norm(log)	0.430	0.215	0.248	0.158	0.191

From Table 1, it is clear that the lower confidence bounds for these scale-mixture families offer little improvement over naive nonparametric approaches. Any reasonably conservative assessment based on these tail assumptions must conclude that the data does **not** support any informative extrapolation at $z^* = 26$. Of course, these tail models differ from the "pessimistic" classes in the paper. However, I would have expected the scale mixture families to be somewhat smaller than the classes of "pessimistic" distributions introduced by Professor Fygenson. For example, the negative exponential scale mixtures are simply the family of Laplace Transforms that integrate to one. These distributions must be "completely monotone" (see Feller (1966, Sec. XIII.4)), which requires bounds on all derivatives (and not just the first or second).

As a consequence of the results in Table 1 above, either Professor Fygenson's "pessimistic" families are much smaller than the scale mixtures, or the somewhat complex minimizations needed for the lower confidence bounds of the paper may not have been computed with sufficient accuracy. I believe that the latter possibility may be occurring. In this regard, extensive experience with attempts to optimize functions that lack convexity (or concavity) leads me to be cautious about the two-stage optimization in the paper that requires an initial optimization over mixtures to find the pessimistic bound and then constrained optimization over the bivariate confidence set to obtain the confidence bound. In any event, the scale-mixture classes here do seem to calibrate the analysis in the paper: it is very hard to imagine that a pessimistic statistician (or engineer) would consider the scale mixtures as unduly broad alternatives; and thus the variability suggested by these families must be a conservative lower bound on the statistical variability of any approach deserving the name "pessimistic".

Finally, I would like to remark on the relative value of the asymptotic and bootstrap samples for assessing confidence. For discrete models, the naive bootstrap is known to lack any additional accuracy beyond the first-order asymptotics (for example, see Hall (1992, p.90)). The fact that the bootstrap samples from only 17 binomials exaggerates the effect caused by discreteness. Figure 1 gives scatter plots for the asymptotic and bootstrap samples (in the first two plots). The last plot is for the bootstrap samples in the lower right cluster. Given these pictures, the discrepancies in Table 1 are hardly remarkable, and my personal feeling is that the asymptotic sample is providing somewhat more accurate assessments.

2. Minimization over Scale Mixtures

Consider the problem of minimizing $F(x_3)$ subject to constraints, $F(x_1) = q_1$ and $F(x_2) = q_2$ over a convex set of distribution function: specifically over the scale mixture classes of Section 2. The Corollary to Theorem 4 in Collins and

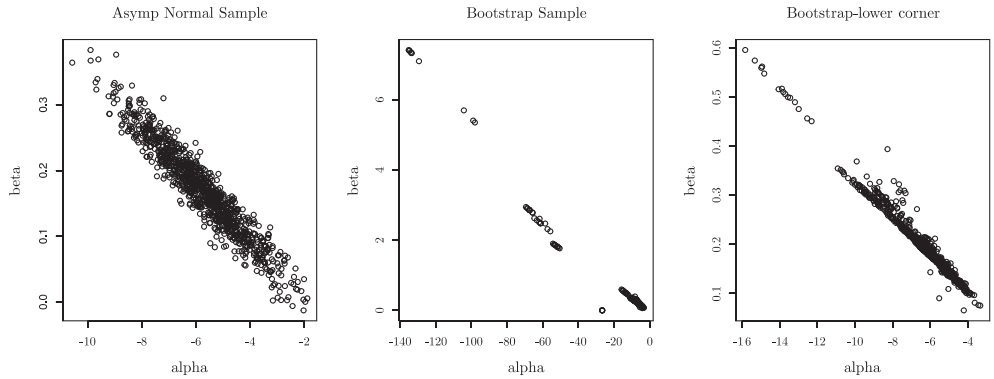

Figure 1. Scatter plots for asymptotic and bootstrap samples.

Portnoy (1981, p.575) provides a straightforward monotonicity condition, under which the optimization is attained at a distribution for which the mixing distribution is a convex combination of two fixed point masses. In the problem here, where x_3 is larger than x_1 and x_2, the distribution minimizing $F(x_3)$ over scale mixtures of the form

$$F(x) = \int G_0(sx)\, dH(s) \tag{7.4}$$

takes H to be a combination of a point mass at infinity and a point mass at s^*; *viz.*,

$$F^*(x) = p + (1-p)\, G_0(s^* x)\,. \tag{7.5}$$

The result is given in Efron and Olshen (1978) for normal scale mixtures, and the condition is shown to hold for this class and for negative exponential scale mixtures (for which (7.4) is just a Laplace transform) in Collins and Portnoy (1981). For scale mixtures of the logistic distribution, the condition can be shown to hold after a rather tedious but straightforward analysis involving the second derivatives of the logistic function. The specific solution is then found by solving the two constraint equations, $F(x_i) = r_i$, for $i = 1, 2$, for p and s^*. Specifically, from (7.5), these equations are

$$(1-p) + pG_0(s^*\, x_1) = r_1 \;\text{ or }\; p(1 - G_0(s^*\, x_1)) = 1 - r_1 \tag{7.6}$$

$$(1-p) + pG_0(s^*\, x_2) = r_2 \;\text{ or }\; p(1 - G_0(s^*\, x_2)) = 1 - r_2\,. \tag{7.7}$$

Eliminating p gives

$$\frac{1 - G_0(s^*\, x_1)}{1 - G_0(s^*\, x_2)} = \frac{1 - r_1}{1 - r_2}\,, \tag{7.8}$$

from which s^* can be found. Either equation in (7.6) or (7.7) can now be used to find p, and the Lower bound (at x_3) can be computed as

$$L \equiv (1-p) + pG_0(s^* x_3) \ . \tag{7.9}$$

For the cases here, denote the constrained probabilities at $s = 2$ and $s = 4$ by r_1 and r_2 given by (7.2). For the negative exponential case, (7.8) and the corresponding p become

$$\frac{1-r_1}{1-r_2} = \frac{\exp(-2\,s^*)}{\exp(-4\,s^*)} = \exp(2\,s^*)\,, \qquad p = \frac{(1-r_1)^2}{1-r_2}\,,$$

and the lower bound at $z^* = 26$ is easily seen to be

$$L_1^* = (1-r_2)\left(\frac{1-r_2}{1-r_1}\right)^{11}\,.$$

For logistic scale mixtures at $\log(z)$, (7.8) and a bit of algebra gives

$$0 = -(r_2 - r_1) \ - \ (1-r_1)\,\exp(s^*\,\log(2)) + (1-r_2)\,\exp(s^*\,\log(2))^2\,,$$

from which we can solve for $a^* \equiv \exp(s^*\,\log(2))$ and p as

$$a^* = \frac{1-r_1 \ + \ \sqrt{(1-r_1)^2 + 4(1-r_2)(r_2-r_1)}}{2(1-r_2)}\,, \quad p = (1-r_1)(1+a^*)\,.$$

It is not hard to see that only the positive square root works. Thus, the lower bound at $z^* = 26$ is

$$L_2^* = 1 - \frac{(1-r_1)(1+a^*)}{1+(a^*)^{13}}\,.$$

For the normal scale mixtures at $\log(z)$, the situation is not quite so simple. Here, (7.8) is

$$\frac{1-\Phi(s\,\log(2))}{1-\Phi(s\,\log(4))} = \frac{1-r_1}{1-r_2}\ . \tag{7.10}$$

It is easy to see that the left hand side of (7.10) is strictly monotonic in s, and can be solved numerically uniquely (and quickly) using the R-function, `uniroot`. Having found s^*, $p^* = (1-r_1)/(1-\Phi(s^*\,\log(2)))$, and the lower bound becomes $L_3^* = 1 - p^*(1-\Phi(s^*\,\log(26)))$.

3. Conclusions

1. I am very impressed by the approach taken by Professor Fygenson to define nonparametric classes of tail distributions that are small enough so that conservative extrapolation can still be used. While I applaud Professor Fygenson's

effort to define "pessimism", I believe the classes of scale mixtures are no less appealing. In general, I believe it would be extremely difficult for the typical scientist (or statistician) to calibrate the degree of pessimism represented by any specific class. I would recommend that a variety of such classes be examined before assessing the efficacy of the extrapolation, and the simplicity of the solutions for scale mixtures would seem to be a valuable selling-point. The results here corroborate Professor Fygenson's conclusion that a range of classes of tail distributions provide somewhat similar results, and that consideration of a relatively small number of such classes might very well provide sufficiently conservative assessments even for "pessimistic" analysts. Nonetheless, the analysis here suggests that attempts to extrapolate rather far into the tail with rather limited data are not likely to be very successful.

2. While the relative small size of the Challenger data appears to preclude little improvement over the naive "monotonic" assesment (*viz.*, that the probability at $z^* = 26$ is not larger than that at the largest observed value, $z^* = 4$), the method would clearly be much more effective for larger sample sizes. Assuming that four times as many observations would halve the difference between the lower bound at the parameter estimates and the lower confidence bounds, the approach would provide some improvement over naive nonparametric bounds. Very large sample sizes must indeed reduce the statistical variability to a negligible value, permitting the "pessimistic" bounds to provide useful assessments of the variability of extrapolations.

3. The extrapolation problem for the Challenger data is in fact rather artificial. The "failures" counted by the binomial responses were in fact not catastrophic ones. The real scientific question is whether greater probability of the kind of event measured in the data had any bearing on the failure of the O-ring at the shuttle launch. In fact, even knowing that the O-ring failed catastrophically at the cold take-off temperature does not establish a connection between the events of the data and the launch event. This connection was, I believe, never firmly established. While the data set is intriguing, one must be careful not to take unwarranted lessons from such analyses. In this sense, it would have been very useful to see some examples of more applicable extrapolation problems for binary response data sets of the sort more typically encountered by statisticians (and for which scientific implications might be clearer).

4. One major question that seems to beg an answer is whether (or how) the approach here can be applied more broadly in statistics. At least, extensions to simple linear regression would seem to be called for. In this regard, Professor Fygenson's assessment that the method only applies when the response function is a distribution function seems overly pessimistic. The results of Collins and

Portnoy (1981) can be applied to the class of bounded monotonic functions (and not just distribution functions that go from 0 to 1); and scale mixtures of such a function would seem to provide some reasonable sets for which the range of variation might be small enough to provide effective extrapolation. The present (rather limited) examples are promising, but application to more realistic extrapolation problems still lies ahead. At least at this point, we finally have some optimistism that pessimism may offer a useful path to progress.

Acknowledgement

This research was partially supported by NSF Grant: DMS06-04229.

References

Collins, J. and Portnoy, S. (1981). Maximizing the variance of M-estimators using the generalized method of moment spaces. *Ann. Statist.* **9**, 567-577.

Efron, B. and Olshen, R. A. (1978). How broad is the class of normal scale mixtures? *Ann. Statist.* **6**, 1159-1164.

Elfving, G. (1952). Optimum allocation in linear regression theory. *Ann. Math. Statist.* **23**, 255-262.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications II.* Wiley, New York.

Hall, Peter (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

Karlin, S. and Studden, W. J. (1966). *Tchebycheff Systems: with Applications in Analysis and Statistics.* Interscience, New York.

Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, Illinois 61820 U.S.A.
E-mail: sportnoy@uiuc.edu

COMMENTS

Vance Berger, Shu Zhang and Yan Yan Zhou

National Cancer Institute, University of Maryland, College Park and California State University, East Bay

Fygenson correctly points out that model uncertainty and prediction uncertainty always accompany statistics research. We select whichever model fits the data best and treat it as the true structure from which the data were drawn

(model uncertainty). The ease with which models are extrapolated makes it tempting to proceed as if this same model applies outside the data range, but this leads to prediction uncertainty. Consider, e.g., a problem that might face math students learning the basics of arithmetic sequences. The first six terms of a sequence are given to be 1, 2, 3, 4, 5, 6; what is the next term? Without even considering the context, the reflex answer would always be 7. Note that this problem may be posed as a special case of the problem considered by Fygenson, as the covariate X would be the place in the sequence, and the outcome Y would be the value. One clearly sees that $Y(X) = X$ for $X = 1, 2, 3, 4, 5, 6$, so one concludes that this pattern will continue when asked to supply a prediction, if you will, of $Y(7)$. But now consider the context. Suppose that we toss a single die six times and observe 1, 2, 3, 4, 5, 6. Clearly, in this context, $Y(7)$ cannot be 7, regardless of what the past may have led us to believe.

The simple, yet profound, tautology governing our ignorance is that one cannot know that which one cannot know. This truth will not change, no matter how much we want to know, or pretend that we can know, or find statisticians (or others) willing to tell us what we want to hear regarding our ability to know. Failure to recognize this profound tautology can potentially produce overly optimistic results in medical research. Consider, e.g., Berger's (2000, Sec. 2.2) corollary to the Heisenberg uncertainty principle, which states that "one cannot observe a patient without altering that patient's response, especially if the patient is aware of being observed". If one binary covariate is the indicator of participation in a medical study, then the covariate takes the value 1 for each patient studied, and the value 0 for each patient to whom the results are to be extrapolated. How do we know how safe we are in this extrapolation, especially without a true random sample?

As another example, consider what it means to validate a surrogate endpoint, regardless of the specific methods used in the validation process. Generally, we conduct a series of trials, collecting both the surrogate endpoint (perhaps a measure of tumor shrinkage in cancer patients) and the clinical endpoint it would replace (perhaps survival time). We note that in all (or most, many, or some) of these trials, it appears possible to predict the true endpoint from the surrogate with the help of some statistical model. Armed with this knowledge, we are now prepared to do without the clinical endpoint in the next study.

The binary response value, Y, represents the extent to which the clinical endpoint can be predicted from the surrogate, and the covariate, X, is the indicator of whether or not the clinical endpoint is collected. Often it will happen that the clinical endpoint will be collected in earlier trials, but not in later trials, as validation studies generally precede studies in which the surrogate endpoint is actually used in place of the clinical endpoint. Because of potential confounding,

this can be problematic. Suppose, e.g., that enthusiasm early in the drug development program causes the investigators to pay special attention to the patients, and that this special attention then translates into additional ancillary care that improves patient health but would not be available once the enthusiasm wears off. Now suppose that the experimental treatment causes an improvement in the surrogate endpoint but has no effect on the clinical endpoint. Then the validation trials might lead one to conclude that the surrogate endpoint is validated. Later, when the clinical endpoint is not collected and the extra ancillary care is gone, only the surrogate would improve, but we would never know that we were being misled. We recognize the need for caution when surrogate endpoints replace clinical endpoints, yet the "usual" method seems rather optimistic.

Optimism is also present when prior estimates of variables are taken as gospel for sample size calculations, or when we fail to reject certain distributional assumptions and then assume them to be true. Fygenson characterizes the uncertainty inherent in extrapolation, and identifies pessimistic distributions. These pessimistic distributions might be quite relevant to several aspects of clinical trial research. A responsible physician may not take a pessimistic outlook when testing a treatment, but would certainly want to rule out the possibility of bad outcomes, and this means considering the worst-case situation. Fygenson proposes a worst-case analysis that can be used as a guideline for assessing the impact of extrapolation beyond the range of the observed data. Of course, we may be dealing with a misnomer, because what is called "worst-case" is actually worst among only a specific class of models. Nevertheless, Fygenson's method, possibly generalized to include a broader set of models, can be quite useful for many aspects of clinical trial research.

References

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials, *Statist. Medicine* **19**, 1319-1328.

National Cancer Institute, 6130 Executive Boulevard, Rockville, MD 20892-7354, U.S.A.

E-mail: bergerv@mail.nih.gov

Department of Mathematics, University of Maryland-College Park, College Park, Maryland, 20742, U.S.A.

E-mail: zhangshu@math.umd.edu

Department of Statistics, California State University, East Bay, 245800 Carols Bee Blvd, Hayward, CA 94542, U.S.A.

E-mail: yanyan.zhou@csueastbay.edu

COMMENTS

Jose M. Bernardo

Universidad de Valencia

In the spirit of the RSS discussion papers, let me begin with a bold statement: I believe your approach is really misguided. You begin by formulating the decision aspects of the problem, but then you simply ignore the basic elements of decision theory.

There are many variations in the axiomatic systems for decision making (see e.g., Bernardo and Smith (1994, Chap. 2) for a review), but they all basically agree in their implications: one should describe the decision makers preferences with a loss function, the available information with a probability distribution, and one should minimize the expected loss. Even if you start form a frequentist viewpoint (the concept of admissibility) you are led, with Wald, to conclude that *any* admissible procedure must be one that minimizes some expected loss.

The possibly pessimistic attitude of the decision maker belongs to the loss function, *not* to the probability model. Thus, if a major disaster can occur, this is assigned a large (but finite) loss and the decision maker's optimal strategy will be to take preventive action, even if the probability of the disaster is small; there is no need to make *ad hoc* modifications of the probability structure. That is supposed to describe the data behaviour, and has nothing to do with a possibly pessimistic attitude of the decision maker. It may well be that you need new models to guarantee a sensible tail behaviour, and the discussion by Professor Hsiung provides an example of how this might be done, but one does *not* mix subjective loss elements into the construction of the probability model.

Some would argue that decision makers often violate the axioms of decision theory in practice. This is surely true, but this does not invalidate the theory. Decision theory is a *normative* theory which dictates how reasonable decision making should be done, not a descriptive theory of how people make decisions. Decision theory should not try to mimic what people do, just as geometry does not try to mimic how people produce (often wrong) surface measures. You may certainly question the axioms themselves, but then you would need to provide a better axiom system. And in any case, I wonder if any serious decision maker would be prepared to claim that he/she could knowingly violate, say, the Savage Sure Thing Principle, and pretend that there are occasions where $a > b$ under H and also $a > b$ under not H, but yet $a < b$.

On a different point, Professor Fygenson justifies his use of maximum-likelihood procedures on their good asymptotic behaviour. This is surprising when

he has focused on problems with typically sparse data. One obviously needs inference procedures that are good for very *small* samples..., but of course a Bayesian analysis is needed to do that and, as Professor McCullagh has earlier mentioned, this paper certainly does *not* fall close to the Bayesian paradigm.

References

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, Chichester UK.

Facultad de Matemáticas, Universidad de Valencia, 46100-Burjassot, Valencia, Spain.
E-mail: jose.m.bernardo@uv.es

COMMENTS

Atanu Biswas

Indian Statistical Institute

Fygenson (2008) essentially proposes estimation of a tail probability in the distribution of failures. Admittedly, tail probability estimation is one of the most difficult problems in statistics, and moreover here the problem is considered in the absence of any data in the tail. Thus, the procedure rests on suitable assumptions on the tail, termed the *outlook* by Fygenson (2008).

Prior to the launch of Challenger, there were 23 launch data, each having 6 O-rings, including 10 failures among $6 \times 23 = 138$ independent observations. The temperature varying between $53-81°$F, the data can be easily fitted by a linear logistic model (Dalal, Fowlkes and Hoadley (1989)) or some other models (Lavine (1991)), but the prediction beyond the above-mentioned range will be different for different models.

Lavine (1995) observed that if the flight data launched at 53°F was excluded from the analysis, given a logistic model, the temperature effect is not statistically significant. Dalal and Hoadley (1991) observed that this data point at 53°F, though influential, could have resulted from the model obtained in an *influential analysis* after deleting that data point. It is then a very important question whether or not a logistic (or clog-log or probit)-type model should be adopted for predicting probability distribution at 31°F, or whether one should ignore the temperature (by considering the data at 53°F as outlier). Here we suggest a data-dependent adaptive approach. One can carry out a standard test procedure for testing whether the data at 53°F is an outlier or not. The P-value (P) is certainly an amount of evidence in favor of "no effect of temperature". We find a probability μ such that (i) $\mu \uparrow P$, (ii) $\mu(0) = 0$, (iii) $\mu(1) = 1$, (iv)

$\mu(0.05) = 0.5$. Then, we obtain two confidence intervals $(p_{TL}(31), p_{TU}(31))$ and $(p_{NTL}(31), p_{NTU}(31))$ by considering temperature (logistic or any such model) or not (that is, 138 independent and identically distributed Bernoulli data). The data-dependent confidence interval is then $(p_L(31), p_U(31))$, where $p_L(31) = \mu \times p_{NTL}(31) + (1-\mu)p_{TL}(31)$ and $p_U(31) = \mu \times p_{NTU}(31) + (1-\mu)p_{TU}(31)$. This is a pre-Challenger analysis.

Suppose, prior of the launch of Challenger, we consider the data at 53°F for modeling. Then any reasonable model − pessimistic, average or optimistic − leads to a large probability of failure at 31°F. For the case of the Challenger, NASA managers had known that contractor Morton Thiokol's design of the solid rocket boosters contained a potentially catastrophic flaw in the O-rings since 1977. Moreover, Thiokol engineers were very concerned that abnormally cold temperatures would affect the O-rings. Maybe the pessimistic outlook is not bad in such a delicate issue, especially in the presence of such engineering concern. Given that the Challenger had an O-ring failure, we contemplate an even more pessimistic model. All the post-Challenger analyses based on the pre-Challenger data are driven by the fact that Challenger failed at 31°F. To prove that the launch of Challenger was a wrong decision, people adopt an outlook that is (Challenger-)data-driven.

People should somehow quantify their outlook or belief. Some sort of (Bayesian) model averaging may be the best approach to such a situation.

In the pre-Challenger era, based on the data of the earlier 23 launches in the temperature range $[53, 81]$, a logit or probit or clog-log is a good fit, and the estimate of $p(53)$ is close to 0.3. Assume that prior to the launch of Challenger, the engineers believed that $p(31) = 0.4$. Then the experimenter could consider a smooth curve within $[31, 53]$ such that $p(53) = 0.3$, $p(31) = 0.4$, $p(t) \downarrow t$, and that there is a smooth transition in the range $[53, 53-\varepsilon]$. Alternately, if the engineers believed in a high value of $p(31)$, say $p(31) = 0.8$, then they could consider a smooth curve accordingly. Thus one has only a non-statistical perception or *outlook* of the engineers, and no data to validate it. Statistics is a *data science* and, in the absence of any data in the range $[31, 53]$, there is not much Statistics here, except possibly setting $p(53) = 0.3$. Either of these pessimistic or optimistic outlooks is acceptable to a statistician in the pre-Challenger era.

In the post-Challenger era, all the analyses of the pre-Challenger data aiming at projecting $p(31)$ is based on the 23 launch data before Challenger and on the fact that Challenger has failed. Extrapolation, outlook or perception, whatever be the appropriate term, is driven by the fact that Challenger failed. Although this is not formally a (Challenger-)data driven outlook in the mathematical sense (as in a conditional model or maybe something like a posterior in the Bayesian sense), it is some sort of ad-hoc (data-driven) outlook.

Consider a hypothetical scenario that it is the post-Challenger era, and that none of the 6 O-rings of the Challenger has failed. I am quite sure that people would now opt for an optimistic outlook (that is $p(31)$ is small), driven by this hypothetical success story of Challenger. Thus, the concept of outlook is somewhat data-driven if there is any data, and mostly ad-hoc if there is no data (where the statistician perhaps has no role to play).

One can as well bring the statistician formally into the business using a Bayesian model as follows. Suppose we have k possible choices of $p(31)$, namely $c_1, \ldots, c_k$, such that $p(53) \leq c_1 \leq \cdots \leq c_k$, with prior probabilities $1/k$ each, which is the pre-Challenger outlook for $p(31)$. The post-Challenger outlook depends on the number of O-rings failures (say m). Then the posterior probability of $p(31) = c_i$ is $c_i^m(1-c_i)^{6-m} / \sum_{j=1}^4 c_j^m(1-c_j)^{6-m}$, which is the new outlook. This is the outlook (with $m = 1$) after January 28, 1986.

References

Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *J. Amer. Statist. Assoc.* **84**, 945-957.

Dalal, S. R. and Hoadley, B. (1991). Problems in extrapolation illustrated with space shuttle O-ring data: comment. *J. Amer. Statist. Assoc.* **86**, 921-922.

Fygenson, M. (2008). Modeling and predicting probabilities with outlooks. *Statist. Sinica* **18**, 9-90.

Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle O-ring data. *J. Amer. Statist. Assoc.* **86**, 919-922.

Lavine, M. (1995). Discussion of the paper by Draper. *J. Roy. Statist. Soc. Ser. B* **57**, 85.

Applied Statistics Unit, Indian Statistical Institute 203 B.T. Road, Kolkata - 700 108, India.
E-mail: atanu@isical.ac.in

COMMENTS

David Draper

University of California, Santa Cruz USA

This interesting and stimulating paper concerns decision-making in the face of uncertainty and takes the *Challenger* space shuttle disaster as its case study. The main technical issue in that applied example is estimating the probability of O-ring failure at a temperature (31°F) that involves a substantial extrapolation away from the bulk of the data available on the night before take-off: previous

shuttle launches had all occurred at temperatures ranging from 53−81°F. Fygenson makes the good point that it may be unwise to rely on the extrapolation uncertainty inherent in a single parametric model, or even in an ensemble of such models (via Bayesian model averaging, for instance), as a good estimate of the full extent of uncertainty about what data would be observed at covariate values far from the observed range, and he develops an elaborate and interesting methodology for coping with the problem in a novel way. (It's particularly interesting to discover, in Fygenson's language, that in the class of GLM link functions the probit is odds ratio (OR)-optimistic, the logit is OR-neutral, and the complementary log-log is OR-pessimistic.) When I wrote about the *Challenger* case study some time ago (Draper (1995)) I focused on inference, not decision-making; this is a welcome opportunity to revisit the example from the latter point of view.

In his Section 5.4 Fygenson states that "In the economics literature ... a shift to so-called non-expected utility choice models has taken place in the last two decades," but good old-fashioned Bayesian maximization-of-expected-utility (MEU) seems to me to be perfectly adequate to the task of deciding whether to launch, as follows. Let the two actions under consideration be $a_1 = \{$launch at 31°F (now)$\}$ and $a_2 = \{$launch at 53°F or higher (later)$\}$, which were the two principal options being discussed the night before the launch. Then the utilities involved can be expressed as in this table, with the quantities u_{11}, u_{12} and u_Δ all taken to be positive:

	Catastrophic Failure	Not
a_1: Launch at 31°F (now)	$-u_{11}$	u_{12}
a_2: Launch at $\geq$ 53°F (later)	$-u_{11} - u_\Delta$	$u_{12} - u_\Delta$

Here $-u_{11}$ represents all the negative consequences of catastrophic failure of the shuttle (lives lost, damage to the space program, and so on), $-u_\Delta$ is the additional dis-utility of delay, and u_{12} represents all of the positive consequences of a successful launch (the scientific value of the mission, the gain in prestige for the space program, and so on). (You could include additional columns representing unknown "states of nature" on the night before the possible launch that are intermediate between catastrophic failure and total success, but this would not change the basic conclusion I'm headed toward.) A moment's reflection is sufficient to notice that, in this real-world context, u_{12} is substantially greater than u_Δ and u_{11} is substantially greater than u_{12}, so let $u_{11} = c_1\, u_\Delta$ and $u_{12} = c_2\, u_\Delta$

for some $c_1, c_2 > 0$; for me (putting myself in the place of the decision-makers on the night before the possible launch), c_1 and c_2 would be on the order of 100 and 10, respectively. Finally, let p_t stand for the probability of catastrophic failure if the launch occurs at temperature t. Then action a_2 would be preferable to a_1 under MEU iff

$$\begin{aligned} E[U(a_2)] &= -p_{\geq 53}(u_{11} + u_\Delta) + (1 - p_{\geq 53})(u_{12} - u_\Delta) \\ &> E[U(a_1)] = -p_{31}\, u_{11} + (1 - p_{31})u_{12}, \end{aligned}$$

and when $u_{1i} = c_i\, u_\Delta$ is substituted in (for $i = 1, 2$), all of the utility values cancel and the optimal decision is to delay the launch if

$$p_{31} - p_{\geq 53} > \frac{1}{c_1 + c_2}.$$

With c_1 and c_2 on the order of 100 and 10, respectively, or any other reasonable values (given the real-world implications of the rocket blowing up), so that $(c_1 + c_2)^{-1}$ is on the order of 0.01, it does not take more than a glance at Figure 4.1 in Fygenson's paper and the observation that the probability of catastrophic failure is a monotone increasing function of the probability of failure of a single O-ring to see that the launch should be delayed. I'm confident that the methodology Fygenson developed in this paper will be highly useful in solving a wide variety of future problems, but I'm less confident that his machinery was needed here to see whether the *Challenger* should have been launched.

References

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion and rejoinder). *J. Roy. Statist. Soc. Ser. B* **57**, 45-97.

Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, 1156 High Street, Santa Cruz CA 95064, U.S.A.

E-mail: draper@ams.ucsc.edu

COMMENTS

Mark S. Kaiser and Daniel J. Nordman

Iowa State University

Professor Fygenson has produced a thought-provoking paper that contains many particulars worthy of consideration. We confine ourselves to an examination of the concept of pessimistic and optimistic outlooks, and the way that this

concept translates into the behavior of response curves for binomial generalized linear models. Such response curves are connected with distribution functions, and we believe the pessimistic/optimistic notion can be formulated in terms of entire distributions, rather than only on specified intervals as done in the paper. This greatly simplifies the conceptual basis for consideration of both desired behavior in extrapolation and uncertainty in model forms.

1. Tolerance Distributions and Link Functions

We wish to cast this discussion in the context of traditional short-term toxicity tests that have a statistical structure identical to the one used in the paper. In particular, short-term toxicity tests produce sets of binomial outcomes at different levels of a covariate, generally a toxic substance. A typical analysis would be conducted using a generalized linear model with binomial random component and a chosen link function (e.g., logit, probit, or complementary log-log, just as in the paper). In this context, there is a correspondence between the link function chosen and the assumed distribution of "tolerances" in the population of organisms being tested. The tolerance of an individual organism is defined to be that value of the covariate such that the organism will not respond (e.g., not die) for any value less than the tolerance, but will respond (e.g., will die) at any value of the covariate greater than or equal to the tolerance. As is well known, link functions for binomial random component models correspond to inverse distribution functions of standardized tolerance distributions. For example, under a (standardized) tolerance distribution F, the probability that an organism dies at a (standardized) exposure level x is $\mu = F(x)$ or, in terms of the link function, $g(\mu) = x$ so that $g = F^{-1}$. We may use F as the parent distribution to define a location-scale family with parameters ξ and σ (i.e., $\mu = F\{(x - \xi)/\sigma\}$) whereby $g(\mu) = -\xi/\sigma + x/\sigma$ is linear in the covariate; note that the location parameter ξ may represent mode rather than expectation in the resultant family of tolerance distributions. For brevity, we consider only standardized forms of tolerance distributions in this discussion. The relation between link function and tolerance distribution is given in the paper as expressions (1.1) and (1.2), so this structure also implicitly forms the starting point for the presentation of Fygenson.

2. Pessimism/Optimism and Distributions

The basic notion of pessimistic versus optimistic outlooks underlies a good deal of the approach to extrapolation taken in the paper. We consider these notions in terms of what the paper calls *attributable risk* (AR) because this is the most direct measure of the probability of an adverse outcome, the central concern in problems of the type considered. We do not find the contrast between pessimistic and optimistic distributions presented in the paper entirely

convincing. Fygenson defines only portions of distribution functions as representing pessimistic or optimistic structures, related to convexity or concavity of the function over certain intervals of the line (Definition 2.1 and Theorem A.1). Any distribution with a continuous unimodal density on the entire line must then contain both pessimistic (convex) and optimistic (concave) portions, so that it seems odd to proclaim a distribution as one or the other. The fact that many distributions contain both types of behavior then motivates construction of a piecewise model as given by expression (3.1) of the paper. While the component distributions of model (3.1), namely F_0, F_{01} and F_1 are all restricted to be absolutely continuous with densities having at least one continuous derivative (see below Definition 2.1), we did not see restrictions imposed that guarantee the entire model determines an absolutely continuous distribution function. This might not be a great issue if all one is concerned with is extrapolation from a regression function for a single value of the covariate, but it would seems lacking as a conceptualization of an overall problem involving tolerance distributions (and densities).

We think of pessimism and optimism as a continuum (with no absolute zero) that may go by either name, and that distributions might be classified only in a relative manner, such as F is more pessimistic (or less optimistic) than G. Thus, the pessimistic *versus* optimistic contrast defined by Fygenson would be replaced entirely by his notion of degree of pessimism (or optimism) as reflected in stochastic orderings, and detailed in Appendix A of the paper. The question becomes, then whether there exist useful classes of distributions that allow stochastic ordering over the entire line, that may be used to define link functions in generalized linear models with binomial random components, and that may be estimated on the basis of observed information. We demonstrate in the next section that at least one such class of distributions can be identified.

The comparison of distributions based on stochastic ordering should be conducted only after the distributions have been "matched" in one or more aspects of their behavior. Without such matching, stochastic ordering may become a relatively uninteresting result of differing parameter values. For example, logistic distribution functions with different location parameters but the same scale parameter are simply translations along the real line. This fact is what underlies the definition of *relative potency* in parallel line bioassays (e.g., Finney (1978)), but is a rather trivial case of stochastic ordering as it relates to the behavior of distributions.

3. Parameterized Link Functions

An alternative approach to accounting for uncertainty in link functions (and hence also tolerance distributions) is to make use of a parameterized family of

links. To our knowledge, the use of families of parameterized link functions was first suggested by Pregibon (1980) who used such families to form alternatives for testing the adequacy of a specific "hypothesized" link. This idea was also used in the analysis of toxicity tests by Aranda-Ordaz (1981). We consider only the simplest among a number of possible families of link functions for binomial responses, given by Pregibon (1980) as

$$g(\mu|\lambda) = \log\left[\frac{(1-\mu)^{-\lambda}-1}{\lambda}\right], \tag{7.11}$$

where μ is the expected value of a binomial random variable (written in terms of proportions), that is, the probability of a response at level x of some covariate. In the paper, this is the combination of expressions (1.1) and (1.2). For the present we consider only $\lambda > 0$ in (1). The family of standardized tolerance distributions implied by this link function are given by

$$F_\lambda(t) = 1 - \frac{1}{\{\lambda\exp(t)+1\}^{\frac{1}{\lambda}}}, \tag{7.12}$$

with corresponding densities

$$f_\lambda(t) = \frac{\exp(t)}{\{\lambda\exp(t)+1\}^{1+\frac{1}{\lambda}}}; \qquad -\infty < t < \infty. \tag{7.13}$$

If $\lambda < 0$ in (1) through (3), the effect is that the support of the density (3) is restricted to $t < -\log(-\lambda)$ and is thus determined by the value of λ.

In the terminology of Fygenson, the distributions (2) are pessimistic for $t < 0$, which is the mode, and optimistic for $t > 0$. But we also have the following.

Result 1. If F_{λ_1} and F_{λ_2} are two distribution functions of the form (2) such that $0 < \lambda_1 < \lambda_2$, then $F_{\lambda_1}(t) > F_{\lambda_2}(t)$ for all $-\infty < t < \infty$.

That is, a random variable with distribution F_{λ_2} is stochastically larger than a random variable with distribution F_{λ_2} over its entire sample space, and both distributions essentially match in the lower portion of the real line (i.e., $\lim_{t\to-\infty} F_{\lambda_j}(t) = 0$).

Proof. Fix t and define a function $h(\lambda) \equiv \log[1+\lambda\exp(t)]$, $\lambda > 0$. It suffices to show $[1+\lambda_2\exp(t)]^{1/\lambda_2} < [1+\lambda_1\exp(t)]^{1/\lambda_1}$ or, equivalently, that $h(\lambda_2)/\lambda_2 < h(\lambda_1)/\lambda_1$. Now, the function $h(\cdot)$ is strictly concave in λ so that, with $\lambda_1 < \lambda_2$, the chord over $(0,\lambda_1)$ must have greater slope than the chord over $(0,\lambda_2)$, that is,

$$\frac{h(\lambda_2)-h(0)}{\lambda_2} \le \frac{h(\lambda_1)-h(0)}{\lambda_1}.$$

Since $h(0) = 0$ the result follows.

This stochastic ordering is illustrated in Figure 1 for six different values of λ ranging from 0.005 to 2.5. Smaller λ correspond to steeper (i.e., leftmost) curves. What is important is that, through the link function (1) and the corresponding distributions (2), we have arrived at an ordering over the *entire* line so that there is no longer any need to restrict functions to particular intervals and construct a response curve as in model (3.1) of the paper. The implication is that the notions of pessimistic versus optimistic distributions as convex or concave over various intervals can be replaced with Fygenson's degrees of pessimism (or optimism) for entire distributions. More pessimistic (or less optimistic) distribution functions pick up probability more rapidly than less pessimistic (or more optimistic) ones, as illustrated in Figure 1.

We point out that the stochastic ordering over the entire line allowed by the distributions of (2) is different than the effect of changes in "slope parameters" for a generalized linear model with a given fixed link. As mentioned previously, the slope corresponds to a scale parameter in tolerance distributions. The effect of scale changes is illustrated for a logit link in Figure 2. Here, changes in scale produce stochastic orderings that differ in direction on either side of the mode of 0 for all of these distributions. Thus, the link function parameter λ in (1), (2), and (3) produces a different effect in terms of Fygenson's pessimism/optimism than does a slope (scale) change in a model with fixed link.

That more pessimistic (less optimistic) distributions accumulate probability more rapidly is clear from Figure 1, but it is illustrative to visualize this phenomenon in terms of density functions for tolerance. Density functions corresponding to the curves of Figure 1 are presented in Figure 3, where smaller values of λ correspond to densities with more rapid declines to the right of the mode.

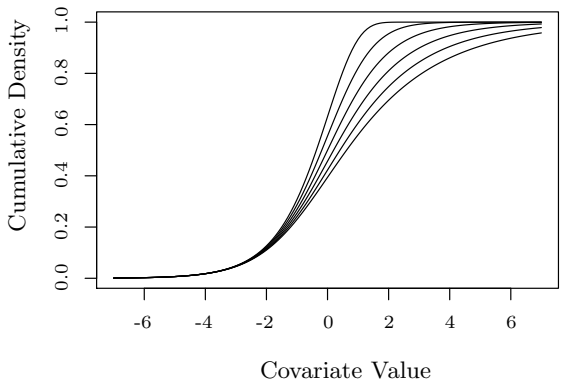

Figure 1. Distribution functions for a variety of λ.

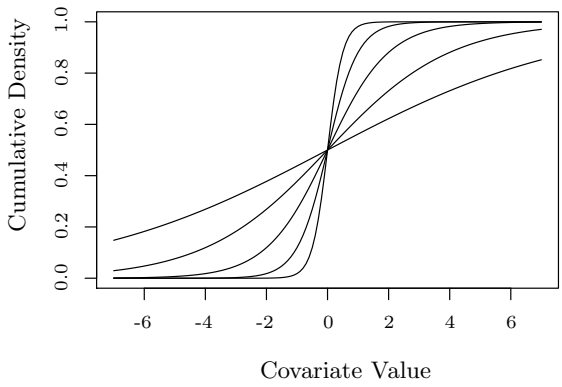

Figure 2. Logistic distribution functions for a variety of scale parameters.

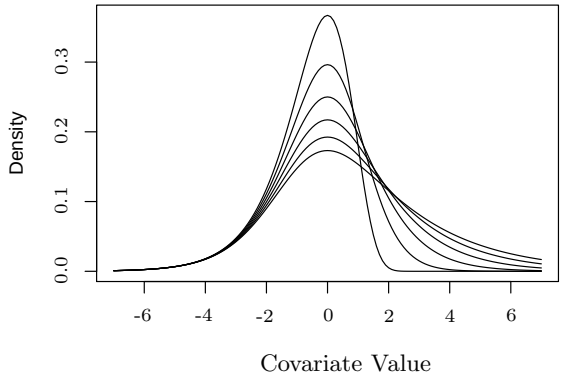

Figure 3. Density functions corresponding to the distributions of Figure 1.

In terms of toxicity tests, or item failure as in the shuttle rocket example, the interpretation of tolerance distributions for different values of λ relative to our modified concept of pessimism/optimism is quite intuitive. All of the distributions in Figures 1 and 3 represent populations of organisms (or test items) with about the same proportion of "highly susceptible" individuals. But more pessimistic distributions represent situations in which "total failure" occurs more rapidly after the mode, while less pessimistic distributions reflect populations with "highly resistant" individuals as well as highly susceptible ones.

Relative to the problem of extrapolation, the value of λ will make little difference in how well data at the low end of the response range can be fitted with a model (this is the practical dilemma of Fygenson), but will give quite different predictions at the upper end of the response range. What is needed to distinguish among the possible values of λ are data at the low end of responses up to about

the modal value. Such data provide the information necessary to determine the "steepness" of the response curve, which corresponds to the rate at which probability is accumulated in the distribution function, and hence the degree of pessimism/optimism the data would suggest. In fact, maximum likelihood estimates of link function parameters such as λ in (1) can be easily computed, simultaneously with all other model parameters, using an algorithm similar to that often employed with traditional generalized linear models (Kaiser (1997)).

The maximum likelihood algorithm just mentioned is a Newton-type algorithm and so also produces the full information matrix for all model parameters, including the link function parameter λ. As a result, one has (i) incorporated uncertainty into the response function through the unknown parameter λ, (ii) estimated that portion of the model form on the basis of observed data, (iii) quantified uncertainty in the estimated value, and (iv) also quantified the effect of that uncertainty on uncertainty about estimates of the other model parameters. Of course, all of this is true only within the class of link functions defined but, in problems for which the response function form is of scientific interest in its own right, this approach might have considerable merit.

References

Aranda-Ordaz, F. L. (1981). On two families of transformations to additivity for binary response data. *Biometrika* **68**, 357-363.

Finney, D. J. (1978). *Statistical Method in Biological Assay.* 3rd edition. Oxford University Press, New York.

Kaiser, M. S. (1997). Maximum likelihood estimation of link function parameters. *Computational Statistics and Data Analysis* **24**, 79-87.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Appl. Statist.* **29**, 15-24.

Department of Statistics, Iowa State University, Ames IA 50011-1210, U.S.A.
E-mail: mskaiser@iastate.edu

Department of Statistics, Iowa State University, Ames IA 50011-1210, U.S.A.
E-mail: dnordman@iastate.edu

COMMENTS

Michael D. Larsen

Iowa State University

1. Introduction

The author wishes to thank Fygenson (2008) for a very interesting and well-developed approach to a hard problem, predicting the *Challenger* disaster. In order to more fully appreciate the functioning of Fygenson's (2008) proposed method it would be interesting to see application to more case studies and simulation from models with known parameter values. It would also be interesting to see simulation with some data under extreme conditions artificially removed.

A Bayesian approach that translates prior opinion and bench test data into prior data points is described in the next section. Such an approach could be used with various models and provide comparisons to methods of Fygenson (2008) and others.

2. Prior Information as Prior Observations

Let the observed data from variables (X, Y) be denoted by $(x, y) = \{(x_i, y_i), i = 1, \ldots, n\}$, where X is continuous and Y is binary. Let F be the CDF for Y given X. Imagine that one can express prior belief by adding artificial data to the set of observations (x, y). Let $(x_a, y_a) = \{(x_j, y_j),\ j = 1, \ldots, m\}$ be the artificial observations. The values x_a could be values of interest, or ones for which some opinion can be formulated. The values y_a then are thought to have arisen from the model F given values of $X = x_a$.

Consider the values of X to be fixed. If F is defined by a linear logistic regression model and the data points (observed and artificial) are all mutually independent, then the posterior distribution for logistic regression parameters given the observed data is proportional to the likelihood (involving the observed data) times the prior distribution (involving the artificial data):

$$\begin{aligned} P(\alpha, \beta|y) &\propto P(y|\alpha, \beta)g(\alpha, \beta) \\ &= \prod_{i=1}^{n} \frac{e^{(\alpha+\beta x_i)y_i}}{1 + e^{\alpha+\beta x_i}} \prod_{j=1}^{m} \frac{e^{(\alpha+\beta x_j)y_j}}{1 + e^{\alpha+\beta x_j}} \\ &= \frac{e^{\sum_k (\alpha+\beta x_k)y_k}}{\prod_k (1 + e^{\alpha+\beta x_k})}, \end{aligned}$$

where the index k in the last line extends through both the observed and artificial data ($k = 1, \ldots, n + m$).

Choosing a prior distribution that matches the form of the likelihood simply means choosing a conjugate prior distribution. Thinking of the prior distribution as representing or arising from prior observations is a common approach. Examples can be found in Gelman, Carlin, Stern and Rubin (2004) for Bernoulli data

(p.40), normal data (p.79), multinomial data (p.93), linear regression models (pp. 383-384), generalized linear models (p.421), and mixture models (pp. 465-466).

A sample from the posterior distribution of the logistic regression parameters given the observed data and prior observations can be generated using the Metropolis algorithm. A convenient jumping distribution is a bivariate normal distribution centered at the current value (α, β) with variance given by the estimated variance (inverse of the negative of the matrix of second partial derivatives, or observed Fisher information) from the augmented likelihood. Each pair of sampled parameter values corresponds through the inverse logistic transformation to a function for predicted probability based on temperature.

3. Initial Data Analysis

The *Challenger* data are available online (Asuncion and Newman (2007)) and consist of 23 rows corresponding to the 23 earlier shuttle flights. There were six O-rings measured on each flight and a total of seven failures. Prior information could potentially come from many sources. The engineers thought that the chance of O-ring failure would be at least as high at 31 degrees as at 53 degrees. Management did not think the risk any higher than in previous launches. Imagine having prior data from six O-rings. As a compromise based on these assessments, one could imagine three failing and three not failing at 31 degrees F. There also were bench test data at 50, 75, and 100 degrees and some tests at below 50 degrees. O-ring failure occurred at 50 but not at 75 and 100. Below 50 degrees there apparently was not a failure. Imagine having data on four additional O-rings with a failure at 50 degrees and non-failures at 50, 75, and 100.

Table 1 contains estimates based on the observed and imaginary prior data. The 95% posterior intervals are based on 50,000 draws from the posterior distribution. The three successes at 31 degrees in priors 1 and 3 greatly reduce the predicted probability of failure while greatly increasing the uncertainty at 31 degrees. The one failure at 50 degrees in prior 2 increases the predicted probability of failure while reducing the uncertainty at 31. The one failure at 50 in prior 3 does not compensate for the three successes at 31 when predicting for 31 degrees. In summary, the outcome is very sensitive to prior assumptions and the use of the model with prior observations gives a clear description of the impact of alternative specifications. It is clear that heaping prior observations on a desired outcome will make it likely. The recommendation is to be clear as to the content of prior observations and study sensitivity to reasonable alternatives.

Table 1. Estimates and predictions using logistic regression. Standard errors are in parentheses.

Data	Estimate of β	Prediction at 31°F	Prediction at 53°F	95% Intervals at 31°F	95% Intervals at 53°F
Observed	-0.1795 (0.0582)	0.9628 (0.0655)	0.4607 (0.1962)	(0.4324, 0.9994)	(0.1054, 0.6292)
Observed, Prior 1	-0.0996 (0.0249)	0.6342 (0.1795)	0.1624 (0.0539)	(0.2775, 0.9102)	(0.0760, 0.2970)
Observed, Prior 2	-0.1822 (0.0531)	0.9659 (0.0533)	0.3398 (0.1268)	(0.5724, 0.9991)	(0.1324, 0.6053)
Observed, Prior 3	-0.1028 (0.0249)	0.6705 (0.1697)	0.1751 (0.0554)	(0.3235, 0.9198)	(0.0859, 0.3059)

Prior 1 has 6 data points: 3 failure and 3 success at 31 degrees.
Prior 2 has 4 data points: 1 failure at 50 and 3 successes at 50, 75 and 100.
Prior 3 has 10 data points: 6 from prior 1 and 4 from prior 2.

4. Expanded Mixture Model

As pointed out in Lavine (1991), more data at high temperatures under a parametric model implies increased precision at low temperatures, which might or might not make sense. The work of Fygenson (2008) acknowledges this possibility by allowing different distributions to govern different sections of the range of X. In terms of a parametric model, this suggests using a mixture model (McLachlan and Peel (2000)). Conditional on the x-values, the y_i's could be assumed to have been generated independently from a mixture of logistic regressions:

$$\begin{aligned} &P(y_i = 1|x_i, \alpha_1, \beta_1, \alpha_2, \beta_2, \pi) \\ &\quad = P(y_i = 1|x_i, \alpha_1, \beta_1)\pi + P(y_i = 1|x_i, \alpha_2\beta_2)(1-\pi) \\ &\quad = e^{(\alpha_1+\beta_1 x_i)y_i}(1+e^{\alpha_1+\beta_1 x_i})^{-1}\pi + e^{(\alpha_2+\beta_2 x_i)y_i}(1+e^{\alpha_2+\beta_2 x_i})^{-1}(1-\pi). \quad (14) \end{aligned}$$

It is assumed that x_i is positive, $0 \geq \beta_2 > \beta_1$, and $0 < \pi < 1$. Since $y_i = 1$ indicates failure, and failure is more likely at low temperatures, the second mixture component with larger β has higher probability of failure and is relevant for low temperature launches.

Instead of specifying a prior distribution for the logistic regression coefficients one could consider, as before, expressing prior belief by adding artificial data to the set of observations (x, y).

Further prior information can be included through definition of some observations as coming from mixture component one (α_1, β_1) and others from mixture component two (α_2, β_2). Let $z = (z_i, i = 1, \ldots, n+m)$ be a binary indicator for membership in mixture component 1. The index i pertains to both the observed data (x, y) and the artificial data (x_a, y_a). In the *Challenger* example, one could set the z-value to 1 for the case with the largest x-value (high temperature).

One could also set the z-value to 0 for the artificial data point with the smallest x-value (low temperature).

A common choice for a prior distribution on π is a Beta(ϵ_1, ϵ_2) distribution, with ϵ_1 and ϵ_2 small positive values. In the application, $\epsilon_1 = \epsilon_2 = 1$.

A special case of this model would set $\beta_1 = 0$, which corresponds to constant probability of failure across temperatures. Although not realistic for low temperatures, this assumption might be reasonable at moderate to high temperatures, which should be more likely to be relevant in mixture component one.

5. Second Data Analysis

A model with constant probability for high temperatures and a logistic regression component for low temperatures is fit to the *Challenger* data. Further prior information in the form of a Beta distribution is placed on the probability of failure in mixture component one. The reason for including this extra prior information is that if there are no cases with failures (or no cases with successes) assigned to a mixture component during an iteration of data augmentation, then the conditional distribution of the probability of success would not be a proper distribution; that is, the usual data augmentation algorithm would fail. Since the probability of failure is expected to be very low, a Beta(0.1, 0.9) prior distribution is used. Since there is a mix of successes and failure at low temperatures no extra prior information was placed on the second component's logistic regression parameters.

Values of z_i are assigned for a number of cases. Observations with temperature below 55°F were placed into mixture component two, whereas those above 72°F were assigned to mixture component one. The eighty-four observations between these two points were initially unassigned.

Table 2 contains the results for the three versions of augmented data. Under this model the observations at high temperatures have little impact on the coefficients of the second mixture component. As a result, the intervals in Table 2 for predictions at 31°F are wider than in Table 1. For predictions at 53°F the intervals are much wider as well. Under all the scenarios the decision to launch at 31°F is not supported. Indeed, the chance of distress of an O-ring for a launch at 53°F seems not insignificant.

Without further prior information, the original data set and the augmented data set with very few cases pre-assigned to mixture components encountered problems in estimation. The cause of the problem was allocations of observations to mixture components such that observations in one mixture component had no failures or there was a complete break in the second mixture component in the range of temperature between successes and failures. These problems are caused by the sparsity of data and small number of observed failures in the data.

Table 2. Estimates using a mixture of a Bernoulli model and a logistic regression model.

Data	95% Intervals at		Probability of failure in component one	Estimated size of components	
	31°F	53°F		One	Two
Observed, Prior 1	(0.1814, 0.8709)	(0.1755, 0.5880)	(0.0000, 0.0232)	115.59	28.41
Observed, Prior 2	(0.0729, 0.9999)	(0.1456, 0.7340)	(0.0000, 0.0446)	116.07	25.93
Observed, Prior 3	(0.1945, 0.8497)	(0.1987, 0.6021)	(0.0000, 0.0221)	119.16	28.84

Priors are as in Table 1.

Points over 72°F are assigned to component 1; those under 55°F are assigned to component 2.

6. Summary and Discussion

Fygenson (2008) has addressed the hard problem of expressing uncertainty for extrapolations in an innovative way. A Bayesian approach that expresses prior information through artificial observations also could be considered. It is demonstrated that prior information can be included in parametric models as prior observations. It is further demonstrated that a mixture of distributions can be fit to these data. The results are very sensitive to both the prior assumptions and model choices.

Dalal, Fowlkes, and Hoadley (1989) and others consider the full problem of predicting catastrophic failure instead of only distress of O-rings. In future work it would be interesting to compare procedures under a range of simulated and actual data conditions.

Other work that seriously considers use of information from other rocket programs is Martz and Zimmer (1992). A further approach would involve modeling the amount of erosion due to heat in O-rings as a function of launch temperature and leak check pressure and predicting erosion at 31 degrees. One could further incorporate scientific evidence relating the likelihood of gas blowby to the extent of erosion. Hierarchical models (six O-rings per launch) also could be considered.

References

Asuncion, A. and Newman, D. J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/∼mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science. [http://archive.ics.uci.edu/beta/datasets/Challenger+USA+Space+Shuttle+O-Ring].

Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *J. Amer. Statist. Assoc.* **84**, 945-957.

Fygenson, M. (2008). Modeling and predicting extrapolated probabilities with outlooks. *Statist. Sinica* **18**, 9-90.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, second edition. Chapman & Hall, CKC: Boca Raton, Florida.

Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle O-ring data. *J. Amer. Statist. Assoc.* **86**, 919-922.

Martz, H. F. and Zimmer, W. J. (1992). The risk of catastrophic failure of the solid rocket boosters on the space shuttle. *Amer. Statist.* **46**, 42-47.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley-Interscience.

Department of Statistics, Iowa State University, Ames IA 50011-1210 U.S.A.

E-mail: larsen@iastate.edu

COMMENTS

Chuanhai Liu

Purdue University

Abstract: Fygenson (2007, hereafter F07) proposed an interesting framework that provides a pessimistic (optimistic) decision maker with probability models. From a Dempster-Shafer (DS) perspective (Dempster (2007)), here we provide an invited discussion on F07.

Key words and phrases: The Dempster-Shafer theory.

1. Introduction

In statistical analysis, it is important to discuss both uncertainty due to model choice and uncertainty about parameters that must be estimated given a model. Such a discussion is sorely needed when scientists grow more serious about statistical methods and scientific inference. This has motivated me, as an applied statistician, to take a closer look at the different schools of thoughts on statistical inference. In particular, I was intrigued by Fisher's attitude toward statistical inference, his understanding of the problem of statistical inference, and the innovative idea *behind* the mathematical formulation of his fiducial argument, then by the subsequent Dempster-Shafer (DS) theory (Dempster (1966), Shafer (1976), Dempster (2007) and references therein). Our more detailed overview on Fisher's fiducial argument and the DS theory is in Liu and Zhang (2007).

The *pessimism/optimism* aspect of Fygenson's work is somewhat reminiscent of the DS theory, as the latter typically provides "data-driven" *probabilities/plausibilities* for assertions of scientific interests. In words, the *probabilities*

and *plausibilities* of the DS theory provide useful reference points for further personal adjustment if decision makers see a need for incorporating their *pessimism/optimism* outlooks. Due to technical differences between DS and Fygenson in the use of Bayesian and frequentist methods, the above comments on the connection between *pessimism/optimism* and *probabilities/plausibilities* are necessarily at a non-technical level.

To follow this discussion, the readers will need to read Dempster (2007), which "outlines a contemporary view of the DS theory that is not part of the standard curriculum in statistics, but may become so in a few years" (Arthur P. Dempster, private conversation). Due to the need for omitting many technical details for lack of space, I shall provide in this contribution simple results, which are related directly to *probabilities/plausibilities* and indirectly to *pessimism/optimism*, in a preliminary DS analysis of the Space Shuttle Pre-Challenger data. Following Fygenson, I focus on the case with the calculated temperature at space shuttle launch time as the single covariate. My purpose here is to illustrate that DS analysis can be a conceptually straightforward way to address both parameter uncertainty and model uncertainty, although the analysis presented is not meant to settle the uncertainty surrounding the Challenger disaster (for more discussion on the reliability of the shuttle, see, for example, Feynman (http://www.fotuva.org/feynman/challenger-appendix.html)).

2. Parameter Inference

As formulated in Dempster (2007), DS analysis starts with state spaces of real-world variables of interest and a DS model. The DS output for any assertion has three components (p, q, r) with $p + q + r = 1$, where p is the *probability* for the truth of the assertion, q is the probability against the truth of the assertion, and r is the residual probability that is understood as the probability of "don't know". The combined probability $p + r$ is called the *plausibility* for the assertion. The corresponding Bayesian output would have $p + q = 1$, that is, $r = 0$. The "don't know" component introduced in DS provides a flexible way for the data analyst to realistically quantify his/her uncertainty. The decision maker can use DS results directly or, for example, eliminate the "don't know" by fusing the DS results further with his/her personal prior information and *pessimistic/optimistic* outlook.

For a simple example, we consider inference about the long run probability of success, P, of Bernoulli trials from a sample consisting of X successes and $n - X$ failures. Given the observed data, in a DS analysis the observer's uncertainty about P is regarded as the same as knowing that P lies between the X-th and

$(X+1)$-th among n ordered draws from the uniform on $(0,1)$. This is illustrated below with the O-ring failure counts for each observed temperature value.

The data consists of the observed O-ring failure counts at each of the 16 observed temperature values $t_1 = 53, \ldots, t_{16} = 81$. Figure 1 shows the O-ring failure counts for 23 pre-Challenger space shuttle launches, each involving 6 field-joint primary O-rings. For example, there were four launches made at 70 °F; the corresponding O-ring failure data at 70 °F consists of the observations $(0,6), (0,6), (1,6)$, and $(1,6)$. Under the assumption that the O-ring failure data are independent binomial counts, DS inference about the failure probability at 70 °F can be made based on the binomial count $X = 0+0+1+1 = 2$ with probability P and size (the number of Bernoulli trials) $n = 6+6+6+6 = 24$. We note that an alternative model could start with different P values for different launches.

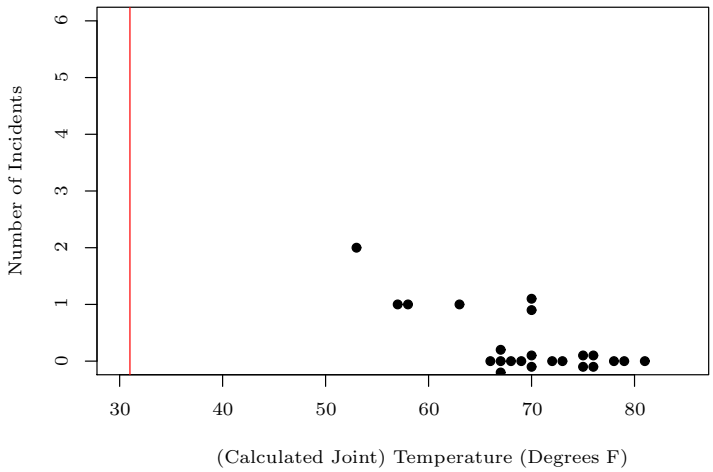

Figure 1. The O-ring thermal-distress data (Dalal, Fowlkes and Hoadley (1989)): Field-joint primary O-rings.

Let n_i be the total number of observed O-rings, and let X_i be the number of observed O-ring failures at t_i for $i = 1, \ldots, 16$. The DS model for the corresponding failure probability P_i is characterized by the *associated* (a)-random interval $[L_i, U_i]$, where L_i and U_i are the X_i-th and (X_i+1)-th order statistics of a sample of size n_i from the standard uniform distribution $U(0,1)$. For example, the (p,q,r) for the assertion $\{P_i \leq P_0\}$ with a fixed P_0 is computed as follows: $p = \text{Prob}(U_i \leq P_0)$, $q = \text{Prob}(L_i > P_0)$, and $r = \text{Prob}(L_i \leq P_0 < U_i)$. Thus given $P_i \in [L_i, U_i]$, the posterior event $\{U_i \leq P_0\}$ is evidence for the assertion

$\{P_i \leq P_0\}$, $\{L_i > P_0\}$ is evidence against the assertion, and $\{L_i \leq P_0 < U_i\}$ provides an instance of "don't know".

It is suggested by the observed data shown in Figure 1 that the failure probability is decreasing in temperature. The observed number of incidents at 70 °F raises the question of the existence of outliers. For a simple DS answer to this question, we aggregate the data over the temperature interval $(60, 70)$ and use a simple binomial model with failure probability $P(70-)$ for the aggregated data. Similarly, we denote by $P(70)$ the failure probability at 70 °F. The assertion of interest is $\mathcal{A} = \{P(70) \leq P(70-)\}$. Based on the two independent binomial counts associated with the O-ring failure probabilities $P(70)$ and $P(70-)$, the DS (p, q, r) output for the assertion $\mathcal{A}$ is $(0.05, 0.70, 0.25)$. Given the fact that the data at $t = 70$ is the extreme case picked up visually, and that the amount of "don't know" component is $r = 0.25$, in what follows we do not treat the data at $t = 70$ as an outlier. We note that one may be more interested in the DS comparison of $P(70)$ based on the data at $t = 70$ alone and the corresponding probability from a sensible DS model based on the data without the observations at $t = 70$. The sensible way along this path would involve the issue of multiple testing. This gets quickly beyond what can be discussed here.

3. Modeling Building and Model Uncertainty

Exploratory data analysis (EDA) has proved to be a useful tool for building statistical models for data. For extrapolation problems, as emphasized by Fygenson, care must be taken because a model fitting the observed data well may hide the uncertainty in both the trend and variability. Thus, instead of a single model, a class of plausible models needs to be considered in such a situation so as to reflect our uncertainty about extrapolated probabilities outside the data range.

To explore plausible parametric models, the Gibbs sampler was implemented to generate 10,000 a-random intervals for P_i, $i = 1, \ldots, 16$, with the assumption that the O-ring failure probability is decreasing in temperature. The details of implementing the Gibbs sampler are omitted here. The marginal 50% and 95% DS intervals, as the DS counterpart of repeated-sampling confidence intervals and Bayesian credible intervals, were computed based on the posterior draws. More specifically, the lower end of the 95% DS interval is the (lower) 2.5% quantile of the lower end of the a-random interval for P_i, while of the upper end of the 95% DS interval is the 97.5% quantile of the upper end of the a-random interval. These intervals are shown in Figure 2 using the "DS box-and-whisker" plots in both the original probability scale and t-link scales with various numbers of degrees of freedom (df). The t-link with 7 or 8 degrees of freedom can be viewed as an approximation to the logistic-link (see, for example, Albert and Chib (1993) and

Liu (2004)). Of course, the t-link with an infinite number of degrees of freedom is the probit link. It is thus seen from Figure 2 (c) and (d) that linear and quadratic logistic and probit regression models are plausible. It is interesting to see that Figure 2 (b) indicates that the possible quadratic trend in temperature in the observed data range can be corrected via (or, more precisely, confounded with) a t-link with small numbers of degrees of freedom, given that the main assertion of interest is about the *lower* O-ring failure probability at 31 °F, since the corresponding upper probability is close to one for almost all sensible models. These DS-box plots also show that the data at 70 °F has certain effects on the *lower* failure probabilities over the interval from 65 to 69 °F.

Based on our EDA, which in a certain sense extends the idea of John Tukey's EDA, we consider the simple sampling model $X_i|(n_i, P_i) \sim \text{Binomial}(n_i, P_i)$, with

$$P_i = \text{pt}_\nu(\alpha + \beta t_i) \qquad ((\nu, \alpha, \beta) \in \Omega_\nu \times \Omega_\alpha \times \Omega_\beta = \{1, 2, 4, 8, 16, \infty\} \times \mathcal{R} \times \mathcal{R})$$

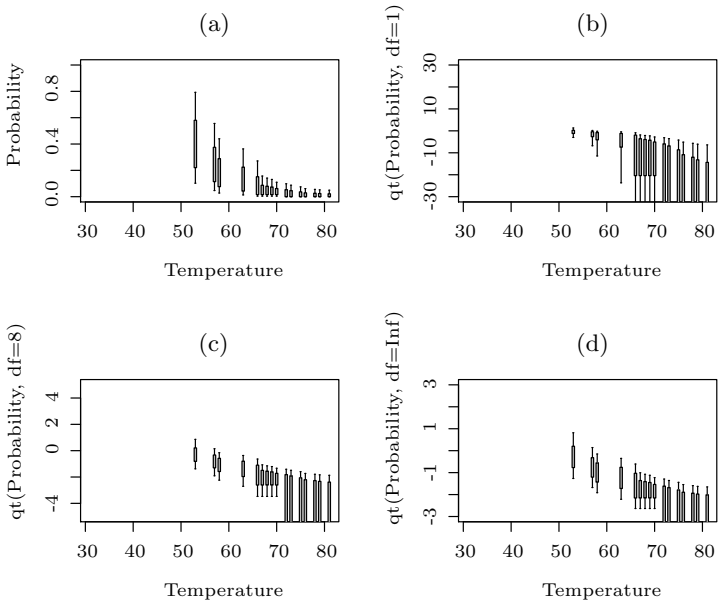

Figure 2. DS box-and-whisker plots using the marginal 50% (boxs) and 95% (whiskers) DS intervals obtained based on the monotonicity assumption on P_is: (a) P_i for $i = 1, \ldots, 16$, (b) the Cauchy-link, (c) the t-link with 8 degrees of freedom, which is approximately logistic, and (d) the probit -link.

for $i = 1, \ldots, 16$, where pt_ν denotes the cdf of the student-t distribution centered at zero with unit scale and ν degrees of freedom. The posterior a-random set for inference about (ν, α, β) is a stack of polygons in the two-dimensional space of (α, β) with ν in a subset of $\{1, 2, 4, 8, 16, \infty\}$ as the stack index.

The posterior a-random set can be simulated using the Gibbs sampler. Preliminary results are promising and are expected to be reported elsewhere.

Acknowledgement

The author is grateful to Arthur P. Dempster for intensive e-mail exchanges on statistical inference in general and the Dempster-Shafer theory in particular, and for helpful comments on earlier drafts of this contribution.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669-679.

Dalal, S. R., Fowlkes, E. B. and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *J. Amer. Statist. Assoc.* **84**, 945-957.

Dempster, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.* **37**, 355-374.

Dempster, A. P. (2007). The Dempster-Schafer calculus for statisticians. *Internat. J. Approx. Reason.*, to appear.

Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (Edited by A. Gelman and X. Meng), 227-238.

Liu, C. and Zhang, J. (2007, in preparation). The minimal belief principle: a new inferential method built on the Dempster-Shafer theory. Technical Report, Purdue University.

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

Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907-2067, U.S.A.

E-mail: chuanhai@stat.purdue.edu

REJOINDER

Mendel Fygenson

University of Southern California

D0. Overview

I would like to begin by expressing my delight with the large number of contributions from such a diverse group of researchers. I thank all the discussants for reading the paper and preparing insightful comments. I am especially honored that Professors Portnoy and McCullagh made the long trip to deliver their

critiques in person. Their presentations in the meeting and those of Professors Fuh and Hsiung, were pleasant, astute and pertinent.

I am also deeply grateful to the editorial board of Statistica Sinica and to their trustees for the opportunity to present the first discussion paper in what I am sure will be a distinguished tradition, and for the great hospitality they extended to me in Taiwan. The paper was greatly improved by the constructive critiques of the editors and referees. What shortcomings remain are entirely my own.

As it is not feasible to address all points raised by the discussants, my response focuses on themes common to multiple discussants that touch upon the most fundamental issues in the paper.

First, several discussants, including Portnoy and McCullagh, questioned the notion of "pessimism" in the paper for its apparent arbitrariness. Bernardo found the framework misguided and suggested that pessimism (optimism) can and should be defined within the classical decision framework. Other discussants, particularly Chang, Chien, and Hsiung, offer a Bayesian analysis that captures some form of pessimism using a Bernstein prior. I address these issues in Section D1. I would like to emphasize here, however, that I am not advocating abandonment of the classical decision framework. My intension is rather to provide additional modeling capabilities for cases where catastrophes are likely and extrapolations are required to make decisions. In economics, a movement to similarly extend the classical framework began about 20 years ago and rose to prominence in the last decade.

Several discussants, most notably Portnoy, raise technical issues. Others, especially Fuh and Hu, and Kaiser and Nordman, asked for/proposed ways in which certain results can be extended. These are addressed in Section D2.

Another common theme among many discussants was the Challenger example. I fully agree with those that suggest other applications must be considered to properly test drive the proposed framework. I am grateful to Fuh and Hu for bringing to my attention potential applications in the areas of Value at Risk Models, Pyrotechnics, Degradation Analysis, Importance Sampling and Sequential Design. In Section D3, I describe another important application, that of Low-Dose Extrapolation.

Many discussants (Biswas; Chang, Chien and Hsiung; Draper; Larsen; Liu) propose alternative analyses of the O-ring data in the Challenger example. In the paper, these data were considered mainly to facilitate the comparison with Lavine (1991) and Draper (1995), who use them to illustrate proper handling of various aspects of model uncertainty.

In the end, I concur with Draper's statement that "it does not take more than a glance at Figure 4.1 in Fygensons paper and the observation that the probability

of catastrophic failure is a monotone increasing function of the probability of failure of a single O-ring, to see that the launch should be delayed."

D1. Choice Models and Related Concepts in Economics

To clarify the relationship between my framework and classical decision theory, and to justify the notion of outlooks (i.e., pessimism and optimism) as I use it, a brief summary of the leading choice models and their developments is called for. The interested reader will find a more in-depth treatment (e.g., about the different axioms) in Diecidue and Wakker (2001). In the summary that follows, I focus on choice models that are commonly applied in economics and other social sciences, where maximization of expected utilities plays a key role in decisions.

Choice models are often described within the following generic setup: a decision maker (DM) faces a prospect or a lottery $(p_1, x_1; p_2, x_2; \ldots; p_n, x_n)$ whose real outcomes x_i have probabilities p_i that sum to 1. When the vector of probabilities, $\mathbf{p} \equiv (p_1, \ldots, p_n)$, is known (objectively or subjectively), the decision is labeled a choice under risk. Otherwise it is labeled a choice under uncertainty. To facilitate the presentation, I describe the developments of choice models under risk, although most statistical modeling problems are more closely aligned with choices under uncertainty. To evaluate the prospect $(p_1, x_1; p_2, x_2; \ldots; p_n, x_n)$, it is often assumed that the DM is using the following general weighting model:

$$\sum \pi_i (\mathbf{p}) \cdot U(x_i). \tag{1}$$

Here, U is a utility function that captures the DM's value or preference with respect to the various outcomes, and π is a decision weighting function that captures the DM's belief or outlook with respect to the outcomes' probabilities. Note that utility functions, which represent benefit or satisfaction, are complementary to loss functions. Furthermore, π_i can be seen as the DM's belief about the likely occurrence of x_i, possibly involving a *misperception* of its probability, or his/her assessment of the relative importance of x_i, in which case he/she may *deliberately* choose a weight other than p_i. The latter case is what I had in mind in proposing pessimistic/optimistic models in the paper.

To earn the label of "rational", a DM's reasoning must satisfy, at a minimum, three requirements: Completeness, Transitivity and Monotonicity. Viewed in terms of the utility function, these requirements (roughly) imply that U is a *consistent* and strictly monotone function (i.e., x_i is preferred to x_j if and only if $U(x_i) \geq U(x_j)$). However, over the last 35 years or so, it has been observed that people often violate some of these rational requirements. Accordingly, researchers have begun to explore ways to modify the classical choice model.

To derive the classical expected utility (EU) model from (1), we need to add the *Independence* requirement- also known as the *sure-thing* principle (Savage

(1954)). Under this requirement, the decision weight π_i for outcome x_i depends only on its probability p_i and not on any of the other outcomes or their probabilities. Then, model (1) becomes

$$\sum \phi(p_i) \cdot U(x_i), \tag{2}$$

where $\phi(p_i) \geq 0$ and $\sum \phi(p_i) = 1$. It can be shown that this last condition (i.e., $\sum \phi(p_i) = 1$) implies that in (2) we must have $\phi(p_i) = p_i$, yielding the classical EU model (i.e., $\sum p_i \cdot U(x_i)$). Therefore, to have choice models with non-linear weighting functions $\phi(p_i)$, it is necessary to relax the Independence requirement and allow the weighting function to depend on the probabilities of other outcomes (i.e., $\phi(\mathbf{p})$).

One such example is the rank-dependence model of Quiggin (1982), which has become the most celebrated of the "non-expected" utility models. Rank-dependence was also used by Tversky and Kahneman (1992) to modify their original Prospect Theory (Kahneman and Tversky (1979)) and create the Cumulative Prospect Theory (CPT). (The first author received the 2002 Nobel prize in Economics for the above work and for having integrated insights from psychological research into economic science). In what follows, we see that the word "cumulative" captures well the rank-dependence idea of Quiggin. Interestingly, McCullagh (1980) in the paper "Regression Models for Ordinal Data" uses a similar idea to capture the ordinal scale of a response variable in his proposed models.

D1.1. Rank-dependence for decisions under risk

The *rank-dependence* requirement allows a DM to weigh an outcome not only with respect to its probability, but also with respect to how good the outcome is in comparison to the other possible outcomes. Formally, the *rank-dependence* assumption implies that the decision weight π_i of getting outcome x_i depends only on its probability p_i and its rank. Ranks are given by a distribution function that assigns to each outcome the probability of receiving that outcome or anything worse. It effectively orders the outcomes from worst to best, assigning a value of zero to anything below the worst outcome, and a value of one to anything above the best outcome. Thus, the rank of the i^{th} best outcome $x_{(i)}$ is captured by its value in the distribution function, $F\left(x_{(i)}\right)$.

To get to the Rank-Dependence Utility (RDU) model from (1), consider (without loss of generality) a prospect $(p_1, x_1; p_2, x_2; \ldots; p_n, x_n)$ in which $x_1 \leq x_2 \leq \ldots \leq x_n$, and $\pi_i = \phi(\mathbf{p})$. The probability-weighting function $\phi(\cdot)$ is strictly increasing with $\phi(p_n) = \pi_n$ and

$$\pi_i = \phi(1 - F(x_{i-1})) - \phi(1 - F(x_i)), \quad i \leq (n-1). \tag{3}$$

(Note that the RDU model can be formulated with the following dual probability weighting function $\pi_i^* = \phi\left(F(x_i)\right) - \phi\left(F(x_{i-1})\right)$.)

The Cumulative Prospect Theory model is a *sign-rank-dependence* model. It generalizes the RDU model by introducing two different probability weighting-functions, one for gains and one for losses (i.e., π^+ and π^- , respectively).

Many studies try to derive and estimate decision weights of *ordinary* people in specific situations (e.g., Abdellaoui (2000) and Gonzalez and Wu (1996, 1999)). Empirically, it has been observed that ordinary people use decision-weighting functions that are concave for small probabilities and convex for moderate and high probabilities. This phenomenon is one of two reasons that I defined outlooks inherent in a distribution only on an interval of its support. The other reason was to minimize structural uncertainty - for more on this subject see Section D2.

D1.2. Risk aversion under the EU model

An important aspect of classical decision theory (which revolves around the maximization of expected utility) addresses the question of how two different DMs, with two different utility functions, react to the same prospect. A crucial concept for comparing DMs with different utilities is risk *attitudes*. For example, risk aversion is an *attitude* (not an *outlook*) that causes a DM to avoid uncertainty. Not all DMs share the same attitude and, even if they do, it might not to be to the same extent.

A risk-averse attitude is reflected in the properties of a DM's utility function U. A DM is labeled *risk averse* (*risk tolerant*) if he/she uses a *concave* (*convex*) utility function U. The DM who uses a linear utility function is labeled *risk neutral*. Moreover, a DM with U_1 is said to be more risk averse than a DM with U_2 if $U_1 = g(U_2)$, where $g(\cdot)$ is an increasing and concave function (Pratt (1964) and Arrow (1974)).

D1.3. About outlooks: pessimism vs. optimism

In the economics literature, *pessimism* or *optimism* are distinct from (and not to be confused with) *risk averseness* or *risk tolerance*. In the paper, this distinction is key because the objective is to introduce a statistical framework in which the unknown probabilities are constrained (non-parametrically) to capture outlooks. Moreover, it is important to note that pessimistic or optimistic outlooks *cannot* be incorporated within the classical EU model because they require nonlinear probability-weighting functions in (1). This fact was first noted by Savage at the end of Chapter 4 in his book (Savage (1954)).

To clarify the meaning of a *pessimistic* outlook in economics, refer to the RDU model and consider a prospect where the i^{th} best outcome $x_{(i)}$ occurs with probability p_i, has a rank $F(x_{(i)})$, and a probability-weighting function

$\pi_i(\bar{p}) = \phi(p_i + (1-\bar{p})) - \phi(1-\bar{p})$, $i \leq (n-1)$ where $\bar{p} = F(x_{(i)})$. (Note that $\pi_i(\bar{p})$ is the same as π_i in (3).)

A pessimistic outlook would require that the decision weight $\pi_i(\bar{p})$ decreases when the rank of $x_{(i)}$ increases (i.e., when $\bar{p}$ goes up, the probability of an outcome worse than $x_{(i)}$ also goes up). Therefore, a DM adopts a pessimistic outlook if the decision weight function $\pi_i(\bar{p})$ he/she uses is decreasing in $\bar{p}$. It is straightforward to see that $\pi_i(\bar{p})$ is decreasing in $\bar{p}$ if and only if $\pi_i(\cdot)$ is convex. In a similar fashion, a DM is an optimist (i.e., has an optimistic outlook) if and only if the decision weight function he/she uses is concave. These definitions imply that a DM who evaluates a prospect using the classical EU model (i.e., $\pi_i(\bar{p}) = p_i$) is adopting a *neutral outlook* (but not necessarily a risk neutral *attitude*, which would require a linear utility function U in (1)).

The notions of pessimism and optimism appeared in Quiggin (1982) and have been developed by Yaari (1987) in his dual theory to the classical EU model. In the economics literature, pessimism (optimism) is formally characterized by a convex (concave) probability-weighting function.

For completeness, it bears mentioning that similar definitions for pessimistic and optimistic outlooks were proposed for choices under uncertainty. By relaxing the Independence requirement to that of *rank-dependence* (as we did for decisions under risk), we get Schmeidler's (1989) Choquet Expected Utility (CEU) model. In this framework, a DM is labeled a pessimist (optimist) if the, so-called, capacity $\omega(\cdot)$ he/she uses is convex (concave). (Note that $\omega(\cdot)$ is convex if and only if $\omega(A \cup B) + \omega(A \cap B) \geq \omega(A) + \omega(B)$.)

The simplest way to arrive at the various notions of outlooks as used in the paper is to consider the following binary regression setup with *one* risk factor (X):

$$P(Y = 1|X = x) = F(\alpha + \beta x) \text{ and } \beta > 0. \tag{4}$$

In the paper, the (generic) model (4) was taken as the point of departure (in the same way we started with model (1)). Indeed, neither model is primitive in the sense that both require some initial assumptions. In the GLM framework, model (4) can be written as:

$$F^{-1}(p_x) = \alpha + \beta x, \tag{5}$$

where $p_x \equiv P(Y = 1|x)$.

To answer the question of how should a professional DM pick F^{-1} (or F) to reflect a particular outlook, consider two cases: 1) F is known (choice under risk) or 2) F is unknown (choice under uncertainty). When F is known, a GLM framework that provides for a DM's outlook can be written as

$$\phi\left(F^{-1}(\theta)\right) = \phi(x), \tag{6}$$

where $\phi(\cdot)$ plays the same role in (6) as in (3). In particular, a DM would have a neutral outlook if $\phi(\cdot)$ is the identity function (or a linear function) and a pessimistic (optimistic) outlook if $\phi(\cdot)$ is convex (concave). When F is unknown (and there are no observations), $\phi(F^{-1}(\cdot))$ in (6) can be treated as one unknown transformation. However, then a different approach is required to answer the above question. The paper put forward one such possibility.

To define three categories of outlooks for a distribution function F *without* invoking relativity to other distributions (as was done in Appendix A of the paper) it is important to note that, by definition, F, which captures risk in model (4), is non-decreasing. Therefore, the paper explored the notion of an extra-risk mechanism, which, in a binary regression setup, leads naturally to various measures of association. By referring to the most commonly used measures of association, the hope was that the resulting outlooks would have an intuitive appeal and that it would be easier for researchers to identify increasing (decreasing) patterns than convex (concave) patterns. (Note that Theorem A.1 of Appendix A in the paper provides the equivalent conditions for Definition 2.2 in terms of convexity and concavity.)

Admittedly, there is some degree of arbitrariness in using a particular measure of association. I would argue, however, that a good case was made for using the odds-ratio in a binary regression setup (see Section 5.3 in the paper).

D2. On Some Other Technical Issues

In this section, I begin by addressing some relevant probabilistic issues and then shift focus to some of the statistical issues raised by multiple discussants.

D2.1. Increasing Risk, Ordering of Distributions and Closure

There are situations in which a DM must decide when the distribution function F of the prospect X represents a more risky proposition than the distribution function G of prospect Y. In the economics literature, Rothschild and Stiglitz (1970, 1971) were the first to provide a comparative definition of increasing risk and related properties. In the statistics literature, some forms of ordering were proposed even earlier, the most familiar example being the usual stochastic order. In economics, this ordering is called *first-degree stochastic dominance* and it is less important than so-called *second-degree stochastic dominance*, which compares two distributions with *equal* means so as to determine which is more likely to take on "extreme" values. In statistics this is known as *convex* (*concave*) ordering. Note that if X is smaller than Y in convex order, then the variance of X is smaller than the variance of Y and, as such, prospect X can be viewed as less risky than prospect Y despite yielding the same expectation.

Importantly, in both economics and statistics all orderings are considered with respect to the entire support of the distributions in question. In the paper,

the framework put forward is more general in that one can extend the interval J to include the entire support. Extending the intervals to the entire support in the context of the new types of ordering (e.g., OR-order and AR-order) and determining their relation to other known orderings (see Shaked and Shanthikumar (2006)) is currently under investigation. Initial investigations (with Shaked) suggest that the implications are non-trivial in both directions (i.e., looking on the new ordering with respect to the entire support, or considering the known ordering only on subintervals of the support).

The question of whether the class of scale mixture families is smaller than the class of "pessimistic" distributions put forth in the discussion is intriguing. A partial answer emerges from noting that, in the case of an OR-pessimistic (optimistic) distribution, neither of the equivalent conditions in Theorem A.1 (allowing only symmetric distributions) is invariant under mixture. For example, there exist pairs of logistic distributions such that their 50% mixture is neither OR-pessimistic nor OR-optimistic. But, from Theorem A.3 it follows that the logistic distribution is OR-neutral, (i.e., it is both OR-pessimistic and OR-optimistic at the same time). With respect to the logistic distribution, it is important to note that it is the *only* unimodal distribution that does not have a change in its OR-outlook. All other unimodal distributions (with support on the entire real line) change their outlook category due to the unimodality. This can be seen as another reason for defining the various outlooks only on a subinterval rather than across the entire support.

Another important question concerns the extendability of the results in Theorem 3.1, Corollary 3.1, and Proposition 3.1 to applications that require left extrapolation to extremely low probabilities. Low-dose extrapolation is one area of application requiring such results. Indeed, under the constraints of AR-pessimism and OR-pessimism, I have already derived upper as well as lower bounds for extremely low (e.g., 10^{-6}) probabilities - details are available on request.

D2.2. On model, structure, form and parameter uncertainties

The importance of model uncertainty, which includes (depending on the model) *structure* and/or *form* and/or parameter uncertainties, to statistical inferences needs no further argument. McCullagh finds the use of structural uncertainty in the paper "a regrettable term because the structure is probabilistic, and no model leaves room for uncertainty in probabilities. Fygenson's use of the terms structure and structural uncertainty refers solely to the choice of F".

In Draper's 1995 seminal paper on the topic, two very different uncertainties are considered. Both are attributes of the variability inherent in the observations when used to pick the *form* of F and/or to estimate the parameters of the model.

In this sense, a non-parametric approach to F creates no *form* uncertainty. Moreover, assuming a particular non-parametric constraint "prior" (e.g., pessimistic outlook), or any prior for that matter, also creates no *form* uncertainty - but it does create *structure* uncertainty because it is a probabilistic assumption. Not being a Bayesian statistician, my "priors" require justification and I hope that Section D1 suffices. I am particularly grateful to Draper for his remark that "Fygenson makes the good point that it may be unwise to rely on the extrapolation uncertainty inherent in a single parametric model, or even in an ensemble of such models (via Bayesian model averaging, for instance)". The number of times I have been hit over the head for not using Bayesian model averaging for the extrapolation problem is just too painful to admit! Nevertheless, using a non-parametric pessimistic constraint is still a probabilistic assumption and, in an effort to minimize its impact (i.e., minimizing *structure* uncertainty) on the final analysis or decision, I suggest the following steps.

1. Minimize the length of the interval for which the assumption is required.
2. Decide, *a priori*, for which probability (or quantile) the decision would be acceptable by all.

In keeping with these suggestions, the paper focuses on inference of the median "lethal" temperature (though the results apply equally well to any smaller quantile). In the case of the O-ring data, for all pessimistic models considered, the interval J did *not* have to be stretched out to 31 degrees (i.e., $Z = 50$).

Professor Portnoy wonders why the lower bounds on the probability of $F_Z(50)$ are so high and whether parameter uncertainty was fully and correctly accounted for.

Lower bounds on $F_Z(50)$ deduced from the values in Table 4.1 are indeed much higher than the bounds in Professor Portnoy's Table 1. (I find this surprising but have no explanation to offer.) However, it should be noted that the widths of the lower *confidence* bounds in the two tables are not directly comparable. Table 4.1 displays confidence bounds for a quantile whereas Table 1 reports bounds for a probability.

To facilitate comparison, I present the required bounds in Table D below. These bounds were derived by fitting a logistic regression to the observations (i.e., for F_0) and by constraining F_1 to be AR-pessimistic. When applying Theorem 3.1 and Proposition 3.1 of the paper, I used the same parameter estimates used by Professor Portnoy (i.e., $\hat{\alpha} = -5.75$ and $\hat{\beta} = 0.170$). The results in Table D indicate that the widths of the lower confidence bounds on $F_Z(50)$ from the two approaches are comparable.

We all agree that the right way to account for parameter uncertainty is to use a proper confidence interval procedure. From my reading of the statistics

literature, the likelihood ratio method can be highly recommended, despite being an asymptotic procedure. To justify its use in cases where extrapolation of probabilities is the main objective, it is necessary to revisit model (1.3) of the paper. In this model, parameter uncertainty has little to do with F_1 (in the region of extrapolation) and everything to do with F_0 (in the region of observations). The latter is taken to be one of the common models (e.g., logistic or normal) and its parameters are estimated using *all* the observations. In the Challenger example there are 138 observations to account for parameter uncertainty- is this number too small to justify the use of the likelihood ratio method?

Table D. Lower Bounds and Lower Confidence Bounds on $F_Z(50)$.

	Estimated lower bound	90% (approx.) LCB	95% (approx.) LCB
$z_p = 24$, $z_q = 26$	0.818	0.422	0.319
$z_p = 24$, $z_q = 28$	0.889	0.449	0.336
$z_p = 26$, $z_q = 28$	0.950	0.470	0.350

D3. Concerning the Low-dose Extrapolation Problem

Low-dose extrapolation is common in risk evaluation of carcinogens. Regulatory agencies are often forced to base their risk evaluations on bioassay data because, for many carcinogens, human data on the effects of long-term exposure to very low doses are not available. In bioassays, animals are exposed to much higher doses than humans are likely to be exposed to, and for much shorter time intervals. Thus, reliance on bioassay data poses two fundamental problems. One is the problem of species conversion: effects on animals need to be converted into implications for people. The other is the problem of low-dose extrapolation (within a species): effects of very low doses must be extrapolated from the much higher dose levels used in the bioassay.

To set safety standards, regulatory agencies (e.g., the EPA) traditionally employ the benchmark dose (BMD) method with a default model of low-dose-linearity. They claim that this approach is inherently "conservative", leading to safe doses (SD) that are protective of the publics health. These SD often correspond to doses for which the upper bound on the projected lifetime incremental risk is 1 in 1,000,000. However, for carcinogens that are directly or indirectly beneficial, these SD may be unpractical and/or excessively protective of the public's health.

Applying the framework introduced in the paper, I have written a technical report evaluating just how conservative the current BMD method is and providing, for the first time, a lower bound on the projected lifetime incremental risk from a SD. The lower bound complements the upper bound provided by

the current BMD method and will hopefully enable more productive risk/benefit analyses. The derivation of these results invokes the classes of AR-pessimism and OR-pessimism. It turns out that among the parametric models that are available from the EPA's software (called BMDS) for fitting the dose-response curves, all types of outlooks (i.e., pessimistic, neutral or optimistic) are represented. The methodology is illustrated by analyzing the same renal cancer incidence data used by the EPA in 2001 to evaluate the carcinogenicity of and set the SD for chronic exposure to bromate.

References

Abdellaoui, M. (2000). Parameter-free elicitation of utilities and probability weighting functions. *Management Sci.* **46**, 1497-1512.

Arrow, K. J. (1974). *Essays in the Theory of Risk Bearing*. North-Holland, New York.

Diecidue, E. and Wakker, P. P. (2001). On the intuition of rank-dependent utility. *J. Risk Uncert.* **23**, 281-298.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 45-97.

Gonzalez, R. and Wu, G. (1996). Curvature of the probability weighting function. *Mgmt. Sci.* **42**, 1676-1690.

Gonzalez, R. and Wu, G. (1999). On the shape of the probability weighting function. *Cog. Psych.* **38**, 129-166.

Kahneman, D and Tversky, A. (1979). Prospect Theory: An analysis of decision under risk. *Econometrica* **47**, 263-291.

Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle O-ring data. *J. Amer. Statist. Assoc.* **86**, 919-922.

McCullagh, P. (1980). Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B* **42**, 109-142.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica* **32**, 123-136.

Quiggin, J. (1982). A theory of anticipated utility. *J. Econ. Behaviour and Organization* **3**, 323-343.

Rothschild, M. and Stiglitz, J. E. (1970). Increasing Risk: I. A definition. *J. Econ. Theory* **2**, 225-243.

Rothschild, M. and Stiglitz, J. E. (1971). Increasing Risk: II. Its economic consequences. *J. Econ. Theory* **3**, 66-84.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica* **57**, 571-587.

Shaked, M. and Shanthikumar, J. G. (2006). *Stochastic Orders*. Springer, New York.

Tversky, A. and Kahneman, D. (1992). Advances in Prospect Theory: Cumulative representation of uncertainty. *J. Risk Uncert.* **5**, 297-323.

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica* **55**, 95-115.

Marshall School of Business, University of Southern California, Los Angeles, CA 90089, U.S.A.

E-mail: mfygenson@marshall.usc.edu