Discussion on Fygenson (2007, *Statistica Sinica*): a DS Perspective

Chuanhai Liu

Purdue University

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 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, 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 tech-

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.

nical 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 2.1 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

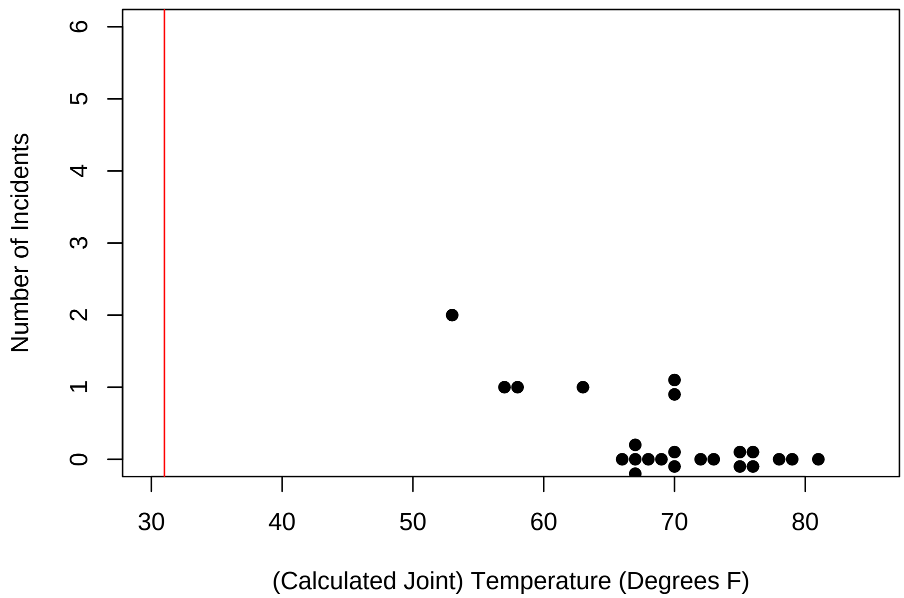

Figure 2.1: The O-ring thermal-distress data (Dalal, Fowlkes, Hoadley (1989)): Field-joint primary O-rings.

°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.

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 $i = 1, ..., 16$. 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 $r = \text{Prob}(L_i \leq P_0 < U_i)$ due to the following fact. 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 2.1 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, $i = 1, ..., 16$, 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

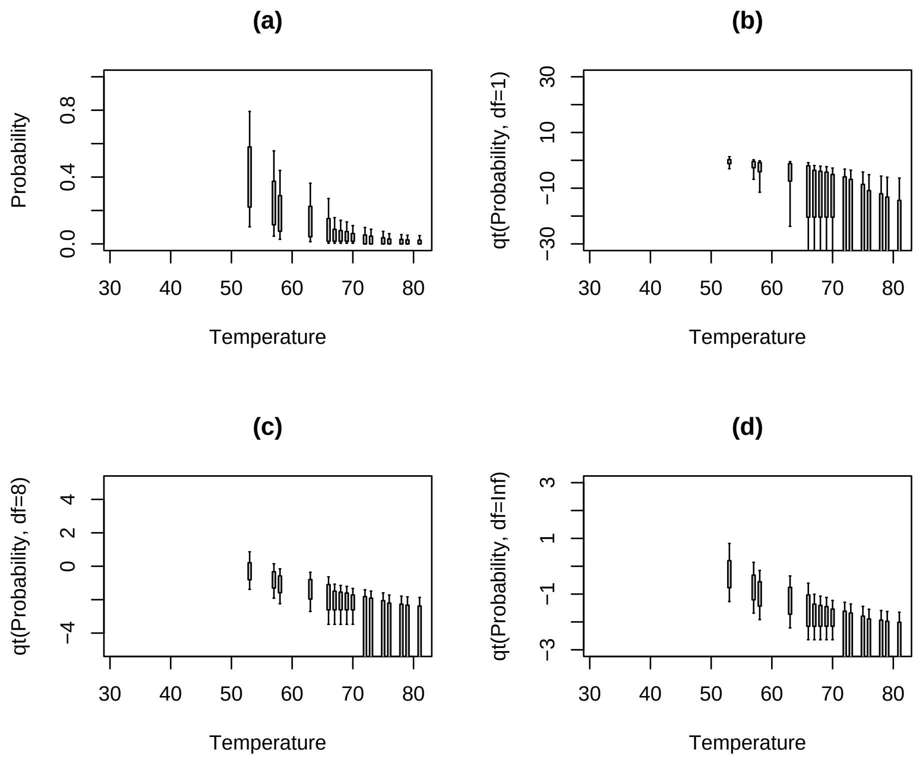

Figure 3.2: DS box-and-whisker plots using the marginal 50% (boxs) and 95% (whiskers) DS intervals obtained based on the monotonicity assumption on P_is: (a) P_i for $i = 1, ..., 16$, (b) the Cauchy-link, (c) the t-link with 8 degrees of freedom, which is approximately logistic, and (d) the probit -link.

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 3.2 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 3.2 (c) and (d) that linear and quadratic logistic and probit regression models are plausible. It is interesting to see that Figure 3.2 (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

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

for $i = 1, ..., 16$, 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.

Additional references

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669-679.

Dempster, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics* **37**, 355-374.

Dempster, A. P. (2007). The Dempster-Schafer calculus for statisticians. *The International Journal of Approximate Reasoning*, 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.

E-mail: chuanhai@stat.purdue.edu