Discussion: Data-Driven Transportation Statistics

Michael P. Cohen[1]

[1]American Institutes for Research, 1000 Thomas Jefferson Street NW,
Washington DC 20007-3835

Abstract
This paper summarizes my discussion of the presentations of the session "Data-Driven Transportation Statistics" at the 2012 Joint Statistical Meetings.

Key Words: empirical Bayes, highway safety, multi-state model, travel time

1. Introduction

I was one of two discussants for the session "Data-Driven Transportation Statistics" at the 2012 Joint Statistical Meetings in San Diego. This paper summarizes my discussion. The common thread for this session was the use of innovative data-oriented methods to tackle problems in transportation statistics. Because of the variety of problems and methods, it seems best just to go through the presentations in order, devoting one section to each.

2. "Decision Tree Induction of Driver's Behavior at a Yellow Light"

This presentation by Linda Boyle and Amanda Raven regrettably could not be given at the meetings. The abstract is, however, intriguing. They propose using a method called *decision tree induction,* a data mining technique, to study a driver's behavior at a yellow light. I am eager to find out how this technique may improve upon logistic regression or classification and regression trees (e.g., CART or CHAID).

3. "Multi-State Travel Time Reliability Models with Skewed Component Distributions"

Feng Guo and Qing Li adopt a multi-state approach to modeling travel time reliability. They test their approach on a fixed corridor in San Antonio.

The mixed model makes sense if there are two well-defined states (e.g., congested and free flow) or three (e.g., congested, incident, or free flow).

In the example Feng and Li present, there is a sharp drop-off at 700-799 seconds (or 750-799 seconds when the data are presented in 50 second increments). This seems strange. Is there a physical explanation? Could it just be random variation?

Otherwise, the data seem to fit a gamma model (Figure 1), or perhaps for a more refined analysis, a two-state model to account for the bulge around 2,000 seconds.

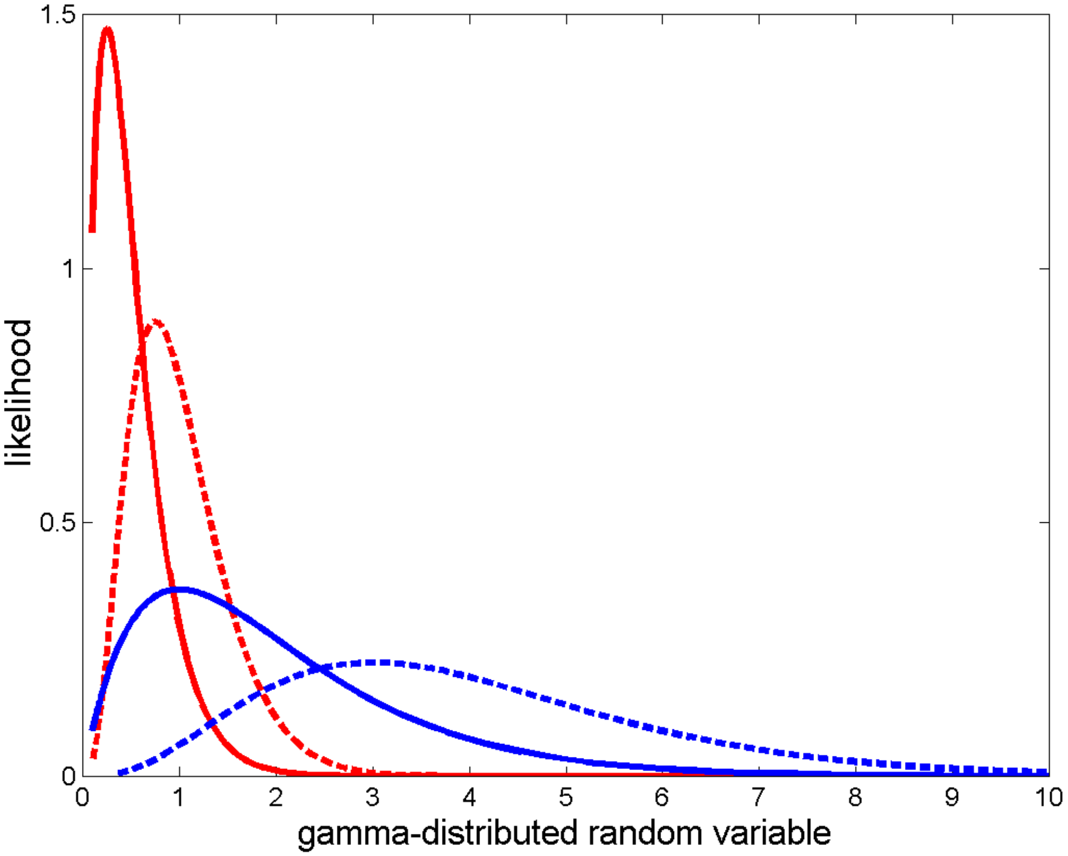

Figure 1: The gamma density for different parameter values.

If one does decide to fit a mixed (multi-state) model, there is still the question of what component distribution to use. The normal, lognormal, and gamma distributions each have two parameters per component and their performance is similar but with the lognormal performing a bit better.

Mixed models can be difficult to fit. Did the authors have difficulty fitting the parameters? One will be tempted to add more and more states to the travel time reliability model as this form of modeling develops so parameter fitting promises to be an important issue.

My questions and comments notwithstanding, it seems clear that multistate models will be having an increasing role in travel time reliability modeling. Guo and Li are to be congratulated for their pioneering work on this topic.

4. "Empirical Bayes Application in Highway Safety Research"

This presentation by Roya Amjadi and Kim Eccles applies empirical Bayes methods to the study of highway crash fatalities at different sites.

I thought I would give a little more background on the empirical Bayes method than Amjadi and Eccles had time to present. Empirical Bayes was first developed by Herbert Robbins (1955, 1964) but he considered a nonparametric version that only works in certain situations. Parametric empirical Bayes was developed by Bradley Efron and Carl Morris in the 1970s (e.g., Efron and Morris 1973, 1975; see also Morris 1983). Their approach became very popular and has been applied in diverse fields.

To give a simple example, let's assume normality. Y_i given τ_i are independent, each with the normal distribution $N(\tau_i, V_i)$. If the variances V_i are small, just use Y_i to estimate τ_i. A Bayesian might assume that τ_i is a random variable with distribution, say, $N(z_i'b, A_i)$ where $z_i'b$ and A_i are determined by prior belief. An *empirical* Bayesian estimates $z_i'b$ and A_i from the data, usually in a non-Bayesian fashion. Thus empirical Bayes is generally not Bayes, but does follow the Bayesian approach up until the last step.

The empirical Bayes estimator will have the form

$$\tilde{\tau}_i = (1 - C_i)Y_i + C_i\, \hat{\tau}_i \text{ where } 0 \leq C_i \leq 1.$$

Here Y_i is the *direct* estimator, that is, it depends only on site i. If it has small variance, then C_i is near 0. If not, then we "borrow strength" from $\hat{\tau}_i$ which depends on all the Y_j.

Specializing our discussion to the traffic crash situation, if there were many pre-treatment crashes at the site, then the estimate for without treatment will be close to the pre-treatment direct estimate. Otherwise, we "borrow strength" from other sites to improve the estimate.

After their important contribution in applying empirical Bayes concepts to traffic crash fatality estimation, Amjadi and Eccles make some thought-provoking comments of a more "philosophical" nature. They argue (I think persuasively) that transportation is in need of its own statistical discipline. They call it *TranStatistics*, analogous to biostatistics. (Alternatively, one could have *TranoMetrics*, analogous to econometrics and psychometrics.) In the early 2000s, Dr. Ashish Sen was Director of the Bureau of Transportation Statistics and he sought to develop a transportation statistics discipline, but for various reasons he had only very limited success. Let us hope that the times are now more auspicious for this much needed development to succeed.

Acknowledgements

A special "thank you" to Dr. Promod Chandhok for organizing the session and for inviting me to be a discussant. I also thank Dr. K. P. Srinath for ably chairing the session and the speakers for their thoughtful presentations.

References

Efron, B., and Morris, C. (1973). "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach," *Journal of the American Statistical Association*, **68**, 117-130.

———— (1975). "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, **70**, 311-319.

Morris, C. (1983). "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, **78**, 47-55.

Robbins, H. (1955). "An Empirical Bayes Approach to Statistics," *Proceedings of the Third Berkeley Symposium on Mathematics, Statistics, and Probability*, Vol. 1, 157-164.

Robbins, H. (1964). "An Empirical Bayes Approach to Statistical Decision Problems," *The Annals of Mathematical Statistics*, **35**, 1-20.