Federal Aviation Administration

DOT/FAA/AM-12/15
Office of Aerospace Medicine
Washington, DC 20591

Predicting General Aviation Accident Frequency From Pilot Total Flight Hours

William R. Knecht
FAA Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

October 2012

Final Report

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-12/15		

4. Title and Subtitle	5. Report Date
Predicting General Aviation Accident Frequency From Pilot Total Flight Hours	October 2012
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Knecht WR	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125	11. Contract or Grant No.

12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	14. Sponsoring Agency Code

15. Supplemental Notes

This work was completed under approved FAA Human Factors research task AHRR521

16. Abstract

Craig (2001) hypothesized a "killing zone"—a range of pilot total flight hours (TFH) from about 50-350, over which general aviation (GA) pilots are at greatest risk. The current work tested a number of candidate modeling functions on eight samples of National Transportation Safety Board GA accident data encompassing the years 1983-2011. The goal was largely atheoretical, being merely to show that such data can be modeled.

While log-normal and Weibull probability density functions (pdf) appeared capable of fitting these data, there was some pragmatic advantage to using a gamma pdf. A gamma pdf allows estimation of confidence intervals around the fitting function itself. Log-transformation of TFH proved critical to the success of these data-fits. Untransformed TFH frequently led to catastrophic fit-failure.

Due to the nature of the data, it may be advisable to place the greatest prediction confidence in a middle range of TFH, perhaps from 50-5,000. Fortunately, that is also the range that captures the vast majority of all GA pilots.

With some care, GA accident frequencies appear predictable from TFH, given data parsed by a) pilot instrument rating and b) seriousness of accident. Goodness-of-fit (R^2) tended to be excellent for non-instrument-rated pilot data and good for instrument-rated data. Estimates of median TFH were derived for each dataset, which will be useful to aviation policy makers.

These data suggest that the "killing zone" proposed by Craig may be wider than originally believed.

17. Key Words	18. Distribution Statement
General Aviation, Accidents, Flight Hours, Modeling, Predicting	Document is available to the public through the Internet: www.faa.gov/go/oamtechreports

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	23	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

ACKNOWLEDGMENTS

Deep appreciation is extended to Joe Mooney, FAA (AVP-210), for his help reviewing the NTSB database queries, which provided the data used here, and to Dana Broach, FAA (AAM-500), for valuable commentary on this manuscript.

CONTENTS

Predicting General Aviation Accident Frequency From Pilot Total Flight Hours

Predicting General Aviation Accident Frequency From Pilot Total Flight Hours

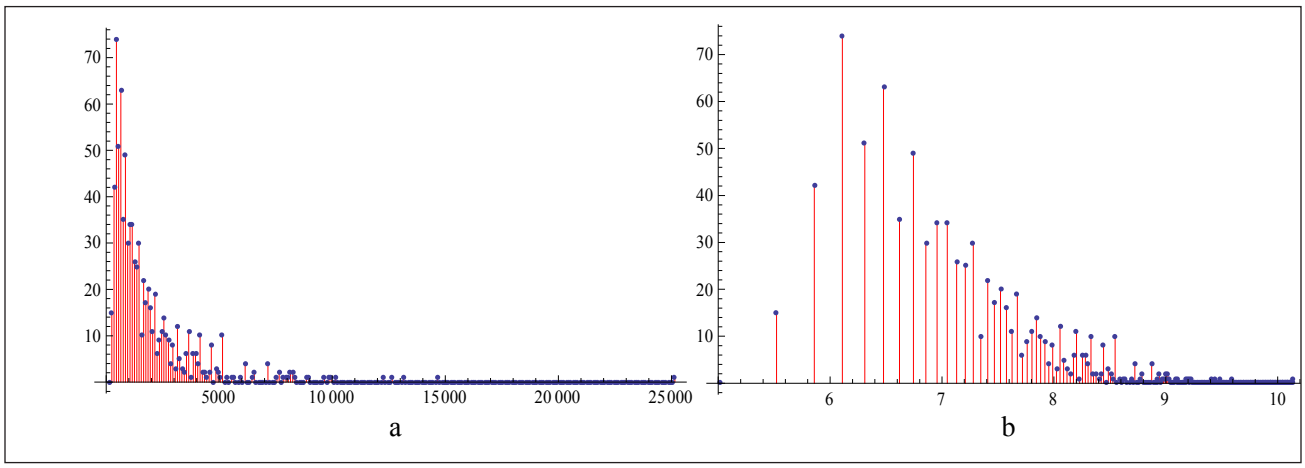


Figure 1. Frequency histogram of a) fatal accident count (y-axis) for instrument-rated GA pilots as a function of total flight hours (x-axis, bin size=100), b) the same data with a natural log-transformed x-axis.

INTRODUCTION

Figure 1a shows an example of one category of data we frequently see in general aviation (GA) accident analysis. This is a histogram of fatal accident counts for instrument-rated pilots of GA aircraft[1] as a function of pilots' total flight hours (TFH).[2] The data reflect actual U.S. National Transportation Safety Board accident data from 1983-2000, inclusive (NTSB, 2011).

It is not hard to appreciate the usefulness of a modeling function here. Such a function would smooth the noise in the data, allowing investigators to better predict how many pilots of a given experience level are likely to be involved in accidents over a given time period. This would be useful, for instance, in allocating resources for pilot training, or as the basis for a statistical covariate of flight risk. Even a casual glance at Figure 1 shows that policy makers would want to focus on pilots having fewer than 5000 TFH, simply because there are far more accidents in that range. The question is how to get beyond the considerable noise in the data to arrive at more precise estimates of this kind.

METHOD

Choosing a modeling function

Ideally, we like modeling functions to be motivated by theory about causal processes inherent to our data. However, in the case of aviation risk, we cannot expect these processes to be few or simple.

Three major processes influence the GA accident rate, two of which have their own set of sub-processes.

1. Processes that affect the *number of pilots* at different values of TFH.
   a. GA flight is expensive, both in time and money, plus, it takes time to accumulate flight hours. Hence, we expect that many pilots will have relatively few TFH, with ever-diminishing numbers of pilots as TFH increase.
   b. Pilots accumulate TFH at different rates, which may, themselves change, depending on the season of the year, employment situation, the pilot's economic and social circumstances, and whim.
   c. Commercial pilots (those who fly as paid professionals) who also fly GA aircraft have their commercial hours included in their TFH. U.S. National Transportation Safety Board (NTSB) data confirm that commercial flying is statistically safer than GA flying.[3] So, for these pilots, high TFH does not imply proportionately higher risk.

[1] "GA aircraft" are defined here as "all N-tail-numbered aircraft operating in the U.S. under all Federal Aviation Regulations (FAR) Parts except 121 and 135, regardless of airframe type or weight."
[2] Total flight hours includes all flight time logged at the controls of all aircraft, regardless of aircraft class or category.
[3] Several hundred people are regularly killed each year in GA, a fatal accident rate about 40 times greater than large commercial passenger air carriers such as United, Delta, and American Airlines (source: www.ntsb.gov).

d. Pilots constantly enter and leave the pilot population at independent rates, due to economic factors and/or old age.
e. Pilots leave one data collection category and enter another, when obtaining a new category of pilot license or certification (e.g., getting an instrument rating).

2. Processes that affect *each individual pilot's flight risk*.
   a. Pilots differ in innate, average skill.
   b. Flight risk tends to be low when pilots are students, to increase when they first begin to fly solo, then to decrease after they gain experience (Craig, 2001).
   c. Some aspects of flight tend to be more dangerous, for example takeoffs, landings, night flights, and flights in or near severe weather. The type of flight a particular pilot typically engages may differ from another pilot, which will have an affect on the exposure to these more dangerous maneuvers. For example, a pilot who typically makes shorter flights will have more takeoffs and landings per unit time period than a pilot who typically flies longer cross-country flights. Therefore, TFH will never be a perfect proxy for risk.
3. Finally, in some particular data (e.g., those of Figure 1), pilots are included who were passengers having little or nothing to do with the cause of the accident itself.[4]

Given such complexity, we might despair at trying to model these kinds of data. Then again, as George Box observed "All models are wrong, but some are useful" (Box & Draper, 1987, p. 424). Perhaps we can begin with seeing if *any* modeling function can fit these data. If so, then we can at least make useful predictions about expected accident rates, even though these may not be purely theoretic.

In the present work, we proceed with this restricted goal, using a standard technique of minimizing least-squares residuals between actual data and a simple model involving just two component functions. The first component will be a simple log-transform. The second will involve that class of particularly useful modeling functions, the probability density functions (pdfs) based on the natural logarithm *e*. These enclose an area of 1.0 under their curves, making them useful in statistical analysis (Spanier & Oldham, 1987). That class includes the Gaussian (normal), Poisson, log-normal, Weibull, beta, and gamma pdfs.

[4]Data source: NTSB downloadable aviation accident database www.ntsb.gov/avdata/. Obviously, some readers may object to including pilots who were not directly responsible for the accident. However, bear in mind that we are merely describing an analytical method here, not trying to support specific theoretical statements about accident causation. Moreover, a previous study conducted by the author indicate that about 90% of GA accidents involve single-pilot flights, where there is no dispute over responsibility.

To illustrate such functions, let us single out the gamma pdf (Γ_{pdf}). This is a 2-parameter function with a *shape* parameter $\alpha > 0$ and a *scale* parameter $\beta > 0$. Gamma pdfs have been used to model a wide variety of processes, including the size of insurance claims (Hogg & Klugman, 1984), amounts of rainfall (Chiew, Srikanthan, Frost, & Payne, 2005), waiting times and mean-time-to-failure (where it represents time until the αth event in a constant-hazard model), and distributions of microburst wind velocity (Mackey, 1998). Since the GA pilot population arguably consists of a number of sub-populations, some having an accident rate being a function of pilot experience *and* numbers of pilots—both perhaps inversely related to TFH—Γ_{pdf} is a logical function to test.

The basic Γ_{pdf} is represented as a function of x (TFH), α (alpha), and β (beta)

$$\Gamma_{pdf}(x;\alpha,\beta) = \frac{e^{-x/\beta} x^{\alpha-1} \beta^{-\alpha}}{\Gamma(\alpha)} \qquad (1)$$

with the gamma function $\Gamma(\alpha)$ itself described as the Euler integral of the second kind, defined for $\alpha > 0$

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt \qquad (2)$$

Figure 2 shows the behavior of Γ_{pdf} with several different parameter values. We can easily imagine that such a function could be fitted to our kinds of data.

Given our data, a number of practical issues arise:

1. The characteristically long tail of the raw data results in a very poor fit to any kind of pdf.
2. A location (shift) parameter will be required, since TFH does not start at zero for instrument-rated pilots.
3. An amplitude term will be required to scale the unit pdf, whose area under the curve is 1.0, to the binned data, whose area under the curve is much greater.
4. Γ_{pdf} should rightfully be tested alongside other plausible candidate distributions.
5. Once we arrive at a preferred fitting function, methods should be described regarding:
   a. Estimation of starting values for constrained parameter search
   b. Goodness of fit with the original data
   c. Calculation of selected confidence intervals
   d. Quantizing the area under the function

Issue 1 suggests a very simple approach, namely, a compressive transform of the x-axis. Indeed, a natural-log compression might prove suitable for "shrinking the tail," to make the raw data of Figure 1a look more like something found in Figure 2.

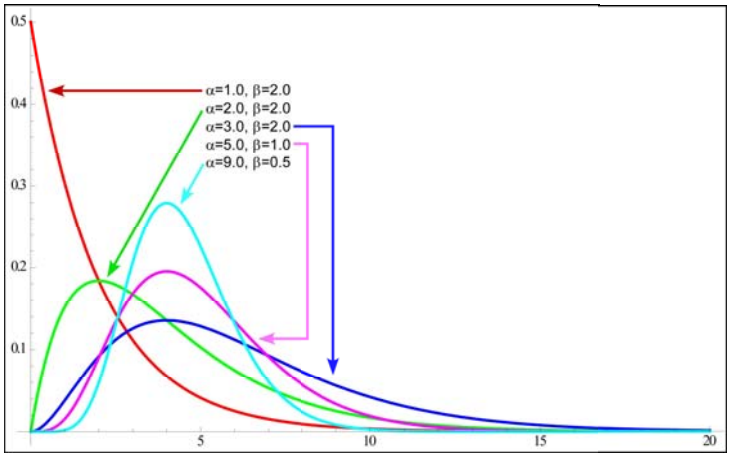


Figure 2. Γ_{pdf} with various values of α and β.

Issues 2 and 3 are purely practical. A location parameter δ (delta) will overcome the problem of non-zero initial data values, while an amplitude term (A) will scale the area under the pdf to our data.

These modifications of Equation 1 lead to the testable form:

$$\Gamma_{pdf}(x;\alpha,\beta) = A\frac{e^{-(\ln(x)-\delta)/\beta}(\ln(x)-\delta)^{\alpha-1}\beta^{-\alpha}}{\Gamma(\alpha)} \quad (3)$$

Issue 4 requires empirical testing of a reasonable set of candidate pdfs (Winkelmann, 2008). Given the nature of the data, we can immediately rule out some candidates[5] For instance, a quick glance at Figure 1a shows that a symmetrical distribution (e.g., Gaussian) obviously cannot fit our data. Additionally, we probably want a function capable of representing overdispersion (variance > mean) or underdispersion (variance < mean). That would rule out the Poisson, where the variance can only equal the mean.

Given these considerations, gamma, log-normal, Weibull, and beta pdfs remain as logical candidates. Because closed solutions do not exist for finding optimal parameter values to such functions, numerical methods must be used. These present their own challenges, as we shall see.

For this study, the *NonlinearFit* function of *Mathematica 7.0* (Wolfram, 2008) was used for parameter estimation. For unconstrained parameters, *NonlinearFit* offers a range of standard numerical methods (e.g., Newton-Gauss, quasi-Newton, Levenberg-Marquardt). For constrained parameters, where starting and/or final parameter values at time t are forced to lie within some range $p_{min}<p_t<p_{max}$, a method such as Karush-Kuhn-Tucker (KKT) is preferable.

An alternative approach might be to use a method like simulated annealing, guaranteed to find a global minimum (Kirkpatrick, Gelatt, Vecchi, 1983; Černý, 1985). However, such methods are computationally intensive and slow, providing little additional benefit, provided we show prudence in our method.

Finally, Issue 5 suggests finding a method for mapping the area under the fitting curve into quantiles, say at 10% intervals. For the time being, let us postpone that discussion until after we settle on a single fitting function and gain some experience in seeing how that function behaves.

The test data

Four candidate model classes were tested: beta, gamma, log-normal, and Weibull pdfs. These were fitted to eight U.S. GA pilot data sets, described in Table 1. The data sets spanned two time periods (1983-2000 and 2001-May 15, 2011), two categories of injury (Serious vs. Fatal),[6] and two categories of pilot instrument rating (Instrument-rated vs. Non-instrument-rated). The data consisted of all pilots involved in U.S. GA accidents during the time period specified, regardless of whether the pilot in question appeared to be legally responsible for the accident.

Table 1. Number of pilots (n) in the ijkth data test set

Time Period (i)	1983-2000		2001-2011	
Accident Category (j)	SERIOUS	FATAL	SERIOUS	FATAL
Instrument-rated pilots (k)	n_{111}= 362	n_{121}= 831	n_{211}=164	n_{221}=465
Non-instrument-rated pilots	n_{112}=1051	n_{122}=1823	n_{212}=328	n_{222}=571

[5]During review of this paper, one reviewer asked about the negative binomial function. This was also tested but eliminated due to frequent optimization failure.

[6]NTSB classifies a "serious" or "fatal" accident as one where at least one person onboard at least one airplane was seriously or fatally injured, respectively.

The raw data were first aggregated into histograms with x-axis bins 100 TFH wide. To aid visual inspection, Appendix A shows one row of data fits using Eq. 3 and a second row using Eq. 1, where the log transform was performed prior to the data fit. The latter is included because that particular method makes it much easier to visualize the underlying functional fit differences.

RESULTS

Goodness of fit

As expected, both the log-transform of TFH and the inclusion of the location parameter δ proved essential. Without the log-transform, parameter estimates failed to converge for nearly all datasets. And, without δ, the same held true for instrument-rated pilot datasets.

Even with the log-transform and δ, however, beta functions rarely produced good data fits. Therefore, beta was eliminated from further consideration as a modeling function, and for the sake of parsimony, results are not shown here. The three remaining fitting functions and eight datasets produced 24 models. Appendix A shows these, with model performance and parameter estimates.

Model goodness-of-fit can be expressed by a variety of metrics. One of the simplest is the coefficient of determination R^2, which varies between 0 and 1, and estimates the proportion of explained variance.

$$R^2 = 1 - \frac{S_{error}}{S_{total}} = 1 - \frac{\sum_{i=1}^{n}(y_x - f_x)^2}{\sum_{i=1}^{n}(y_x - \bar{y})^2} \quad (4)$$

Here, f_x represents the predicted value of y at x, versus the observed value of y_x. Lower R^2s merely reflect noisier data, not necessarily fitting failure.

Two aspects of the data affected both the size of R^2 and the stability of model parameter estimates, as measured by the breadth of their confidence intervals. As expected, greater random variation (noise) in the data and smaller sample size (n_{pilots}) both led to lower R^2s. Binning the data, of course, helps dampen noise, but at the expense of lowering the effective sample size, which becomes the number of bins (n_{bins}) rather than n_{pilots}. Figure 3 shows that better results tended to occur with models based on $n_{pilots} > 300$, while little additional advantage resulted from $n_{pilots} > 1000$.

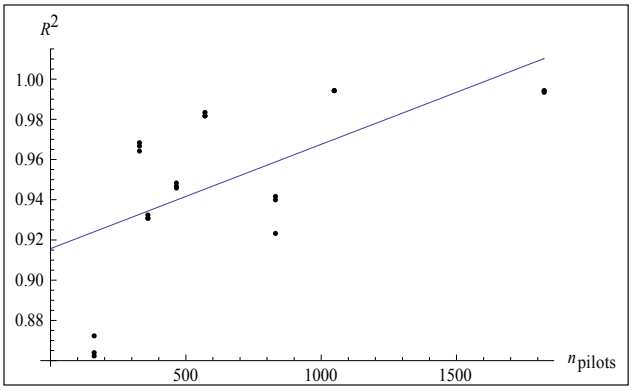


Figure 3. Plot of n_{pilots} (x-axis) versus resulting R^2 (y-axis), with least-squares trend line.

Confidence intervals for the underlying correlation r=Sqrt(R^2) can be estimated using Fisher's method (Glass & Hopkins, 1984, pp. 304-307).

$$r \pm \tanh(Z \pm z_{CI}\sigma_z) = r \pm \tanh\left(0.5\ln\left(\frac{1+|r|}{1-|r|}\right) \pm \frac{z_{CI}}{\sqrt{n-3}}\right) \quad (5)$$

where Z is Fisher's Z-transform of r, σ_z (sigma sub-z) is the standard error of Z, and z_{CI} is the normal z–value corresponding to the desired confidence interval (e.g., for .95CI, z_{CI} =1.96).

Based on R^2, the gamma, log-normal, and Weibull pdfs all appeared reasonable candidates for use with this broad range of accident categories. Appendix A shows R^2s ranging from .862-.994, considered good-to-excellent.

While it was possible to get Poisson models to converge, the R^2s were uniformly and significantly lower than, for instance, those of Gamma ($p_{Wilcoxon}$=.012), supporting exclusion of the Poisson from further consideration. Table 2 compares the two.

The three remaining model classes can be statistically compared by setting up R^2s as if each of the eight datasets were an "individual," and the three models were repeated measures experienced by each "individual." The three "Raw data" rows in Table 3 show the setup of this comparison.

Table 2. Comparing Poisson and Gamma R^2s

Dataset	1	2	3	4	5	6	7	8	Mean	Median
Gamma	.994	.940	.994	.932	.983	.946	.967	.864	.953	.957
Poisson	.890	.770	.880	.748	.929	.873	.892	.840	.853	.877

Table 3. Comparing Gamma, Log-normal, and Weibull R^2s

	Dataset	1	2	3	4	5	6	7	8	Mean	Median
Raw data	Gamma	0.994	0.940	0.994	0.932	0.983	0.946	0.967	0.864	0.953	0.957
	Log-normal	0.994	0.942	0.994	0.931	0.982	0.947	0.964	0.862	0.952	0.956
	Weibull	0.993	0.923	0.994	0.931	0.982	0.948	0.968	0.872	0.951	0.958
Z-transformed	Gamma	2.903	1.738	2.903	1.673	2.380	1.792	2.044	1.309	2.093	1.918
	Log-normal	2.903	1.756	2.903	1.666	2.351	1.802	2.000	1.301	2.085	1.901
	Weibull	2.826	1.609	2.903	1.666	2.351	1.812	2.060	1.341	2.071	1.936

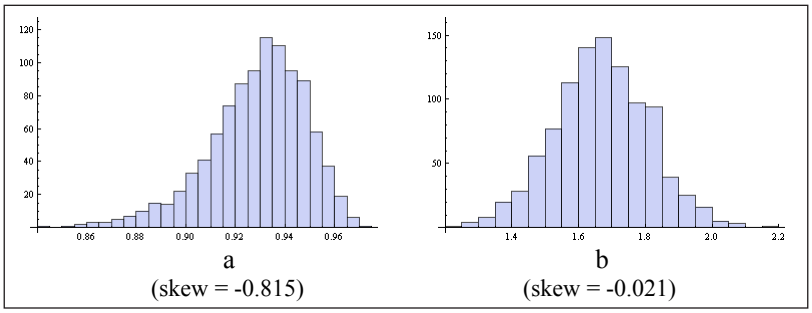


Figure 4. a) Frequency histograms for 1000 Monte Carlo-simulated R^2s randomly generated by one of our Γ_{pdf} models. The y-axis shows the frequency count of R^2s in each x-axis bin. Note the negative skew; b) Z-transform corrects this negative skew.

Quick inspection of these raw R^2s shows no prominent differences between modeling functions. Analysis of variance (ANOVA) can confirm that. But, first we have to consider that our theoretical distributions of of R^2 are not expected to be normal, since R^2 is range-restricted to $0<R^2<1.0$. ANOVA is famously tolerant of some deviation from normality, but these R^2s are high, and might be a problem. Indeed, Monte Carlo simulation reveals considerable negative skewness, as Figure 4a illustrates.

We can correct that skewness using Fisher's Z-transform, a variant of Equation 5:

$$R^2_{corrected} = 0.5 \mathrm{h} \left(\frac{1+|R^2|}{1-|R^2|} \right) \qquad (6)$$

Figure 4b illustrates the resulting improvement in normality. Applying that method to all our R^2s produces the bottom, "Z-transformed" half of Table 3, which we can now more legitimately analyze.

Subsequent ANOVA reveals no significant differences between the gamma, log-normal, and Weibull R^2s (p =.429, NS, Greenhouse-Geisser-corrected for non-sphericity).

One salient feature distinguishing the three model classes was the left-hand side of the curve. As Appendix A makes plain in ln(x)-space, Weibull pdfs tended to be "fatter" on the left, whereas gamma and log-normal pdfs tended to be more symmetrical and to resemble each other more closely. The exact nature of the leftmost, lowest-TFH end of these distributions is something that can be investigated more closely in future years, as data accumulate, allowing more reliable statistical analysis.

Final choice of a fitting function

Judging solely by R^2 and $\tilde{x}_{pdf}$, it would be imprudent to recommend one model over another on the grounds of theory. Arguably, though, the gamma pdf Γ_{pdf} is most useful for two reasons. To a lesser extent, experience with these data showed that Γ_{pdf} was the easiest function to fit. More compellingly, Γ_{pdf} allows calculation of confidence bands around the modeling function itself. These confidence bands provide estimates of the net stability of predictions based on each dataset. Appendix A shows how these confidence bands are influenced by dataset size n_{ijk}, as intimated earlier by Figure 3. Smaller datasets are considerably less reliable.

For these pragmatic reasons, we choose to examine Γ_{pdf} in detail for the rest of this report.

Estimating parameter start values for Γ_{pdf}

The raw data show considerable noise. This produces lumpy parameter residual error spaces and the risk of optimization failure due to local minima. Numerical methods for Γ_{pdf} parameter estimates therefore benefit from having start values as close as possible to final values.

Estimates for α and β can be derived by the method of moments (see Appendix B for details).

$$\alpha_{est} = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{7}$$

$$\beta_{est} = \frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n}{\sum_{i=1}^{n} x_i} \tag{8}$$

A start value for the amplitude term (A) can be estimated as

$$A_{est} = \frac{y_{max,}}{y_{Mo,}} = \frac{y_{max}}{\dfrac{e^{-x/\beta}\left((\alpha-1)\beta\right)^{\alpha-1}\beta^{-\alpha}}{\Gamma(\alpha)}} \tag{9}$$

where, of the *binned* data, y_{max} is the maximum function height, and y_{Mo} is the height of the model Γ_{pdf}mode[7](M_o) of

$$M_0 = (\alpha - 1)\beta \tag{10}$$

Finally, we can estimate the location parameter (δ) by forcing the peaks of both the data and Γ_{pdf} to take the same x-value. This leads to

$$\delta_{est} = x_{max,\,binned\ data} - x_{Mo} \tag{11}$$

In practice, the parameter spaces are lumpy enough so that occasional fit-failure can result, even with reasonable starting estimates (with these eight datasets, this happened once). If all else fails, this can be resolved by graphing out the fitting function for the log-transformed data, starting with the estimated parameters, then hand-manipulating them to better values by visual inspection. Figure 5 illustrates this, using *Mathematica*'s *Manipulate* function, which allows function parameters to be adjusted with sliders, and immediately graphs the result.

Parameter confidence intervals for Γ_{pdf}

Parameter confidence intervals can be estimated with methods based on the Student t-distribution

$$p_{iCI} = p_i \pm SE_i t_{(n-p,1-\alpha/2)} \tag{12}$$

where the ith parameter p_i is augmented or diminished by its standard error (SE_i) times the value of t corresponding to n-p degrees of freedom and the desired 2-tailed significance level α (the acceptable Type-1, or false-positive statistical error rate, not to be confused with our α parameter in Γ_{pdf}). With binned data, n will be the number of bins in each dataset, and p=4, the number of parameters in Γ_{pdf} (i.e., A, α, β, δ).

Estimates of the standard error SE_i are beyond the scope of this report. The reader is referred to Ratkowsky (1989, pp. 36-42) for a treatment of parameter confidence intervals. However, a number of common statistical and mathematical software packages (e.g., SPSS, SAS, MATLAB, *Mathematica*) will numerically estimate the appropriate standard errors.

Using Γ_{pdf}

For a given dataset, to estimate the expected accident count for a bin width of 100 TFH centered on a given value of x, simply populate Equation 3's parameters from Appendix A and insert the desired value of x. For example, for non-instrument-rated pilots having fatal accidents from 1983-2000, inclusive, the dataset "1983-2000 NIR FATL," Equation 3 becomes

$$\Gamma_{pdf}(x) = 776.3\frac{e^{-(\ln(x)-2.821)/0.286}(\ln(x)-2.821)^{7.789-1}0.286^{-7.789}}{\Gamma(7.789)} \tag{13}$$

whose net frequency count at median TFH of x=340 is $\Gamma_{pdf}(\tilde{x}_{pdf})$ $\Gamma_{pdf}(340)\approx$193, which we can see is correct from its plot (Figure 6).

Quantizing Γ_{pdf}

Appendix A shows the actual median (0.5 quantile) of Γ_{pdf}. It is also useful to have a method for dividing fitted data into arbitrary quantiles that can be precisely calculated, for instance, into groups containing equal percentages of area under the fitting curve Γ_{pdf}. Equation 14 shows a method for non-log-transformed data, which is based on the definite integral of Γ_{pdf} (Eq. 3) from x=e^{δ} to x_{qn}, the x-value corresponding to a desired quantile q_n (e.g., 0.8), the corresponding area under the curve, represented as

$$q_n = \int_{e^{\delta}}^{x_{qn}} \Gamma_{pdf}(x,\alpha,\beta,\delta)dx \tag{14}$$

[7]Assuming $\alpha \geq 1$, which all αs here are.

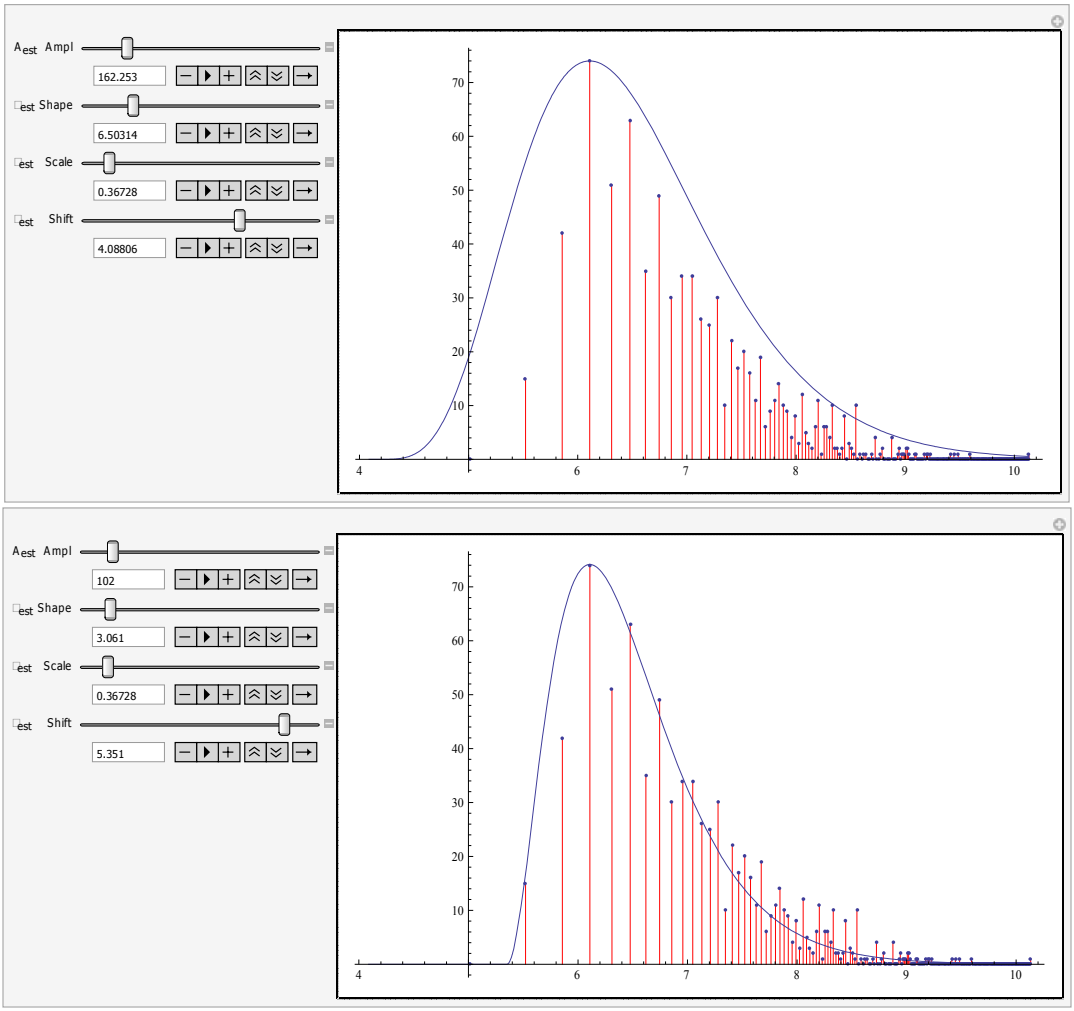


Figure 5. (top) The single dataset where estimated starting parameters (shown) led to fit-failure; bottom) Slight manual adjustment of those initial estimates led to subsequently successful fit.

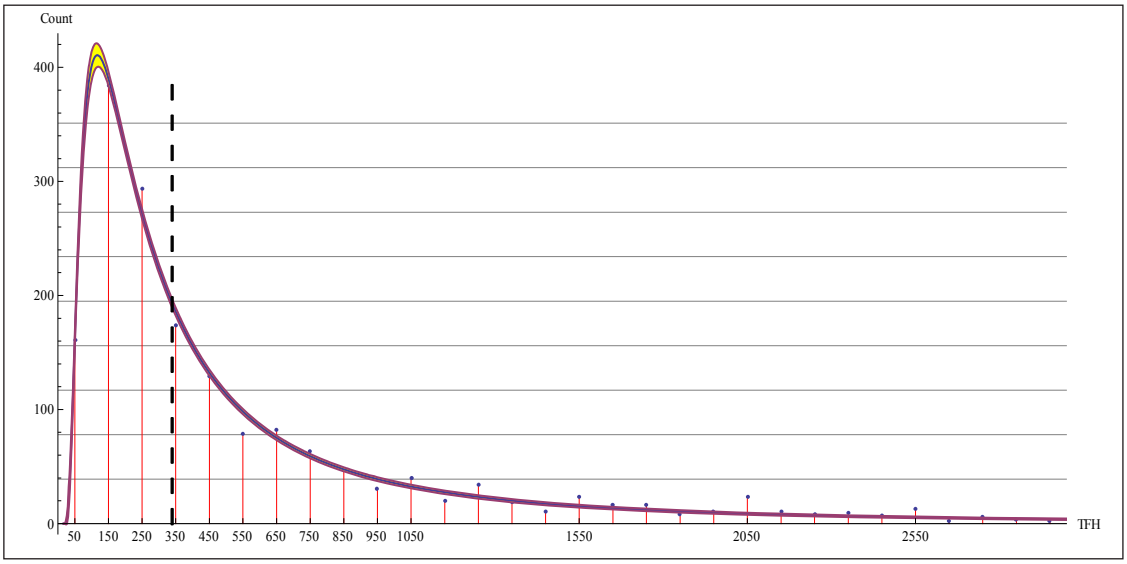


Figure 6. Model Γ_{pdf} for dataset "1983-2000 NIR FATL." The dashed line denotes the median $\tilde{x}_{pdf}$.

There is no closed-form solution for arbitrary x_{qn}. Fortunately, an iterative numerical method is easily stated. We note that the minimum value of $\Gamma_{pdf} = 0$ in the linear domain is specified by $x=e^{\delta}$, and that we normalize the area under Γ_{pdf}, given that it computationally represents n pilots put into bins 100 FH wide.

$$x_{qn} \rightarrow \varepsilon = \left| \left(\int_{e^{\delta}}^{x} \Gamma_{pdf} dx \Big/ 100 n_{pilots} \right) - q_n \right| < \varepsilon_c \qquad (15)$$

In other words, start with an arbitrary value for x, integrate Γ_{pdf} (Eq. 3) to find the area under the curve from e^{δ} to x, next calculate the error ε between that area and our desired quantile, and then adjust x in the direction that minimizes ε, halting when ε falls below a critical tolerance value ε_c (epsilon sub-c). This is easily done with software such as *Mathematica*, given a simple statement such as

FindArgMin[Abs[(NIntegrate[G_{pdf}[x],{x,E^d,mu}]/(100*n))-q_n],{mu,Quantile[data,q_n]}][[1]];[8]

The data of Figure 6 produce the plot for ε:

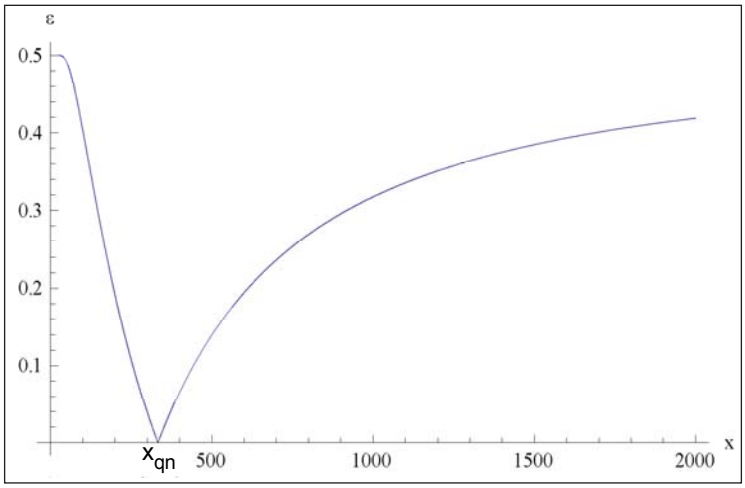


Figure 7. Plot of Eq. 15 for "1983-2000 NIR FATL," here minimizing at the median $x_{qn} = \tilde{x}_{pdf}$.

Naturally, the minimization algorithm will oscillate around the discontinuity, but this poses no practical problem to finding an accurate solution for the median.

A conservative appraisal of model accuracy at extreme x-values

Extensive experience with "extreme" datasets, such as the ones encountered here, teaches us to be wary of what otherwise may seem like high-precision modeling at very low and very high values of x (here being TFH). Often, we find that conservative interpretation works best. In other words, we consciously choose to limit high "logical confidence" in predicted accident frequency to a middle range of, say, 50-5,000 TFH.

The roots of this conservativism are grounded in model construction, sampling, and residual error spaces. Residual error is, of course, the average squared difference between the *y*s our model predicts, given the *x*-values of the data. Total residual error is the quantity we wish to minimize by adjusting our model parameters. Now, what we sometimes see is that this residual error can be minimized *pretty* well by a *range* of models, all of which provide a *pretty good* fit to *most* of the data. This happens when we have a complex residual error landscape with many local hills and valleys (as opposed to a simple, monotonic landscape with just one deep, global minimum).

The exact geometry of the error space is, of course, largely dictated by the data. But, it is also a function of the model and of how the data are set up. For instance, suppose we have two accidents, one at 15,000 TFH, the other at 15,050 TFH. In that event, the error space very much depends *how wide* our sampling bins are and *where each bin starts and finishes* on *x*. In the very long right-hand tail of data like ours, the typical actual accident frequency bin count is going to be 0. But, if our bins are wide along *x*, or spaced in a certain way on the *x*-axis, instead of getting two bins, each 1 unit high, we can end up with one bin 2 units high. Amidst a sea of bins 0 units high, the residual error of this 2-accident bin will be almost $(2\text{-}0)^2 = 4$, as opposed to the $2(1\text{-}0)^2 = 2$ units it would otherwise be.

The point is that the right-hand tail of distributions like these can become "logically unreliable." Most bins in the long right-hand tail will contain zero cases. Consequently, the farther out on the *x*-axis a non-zero bin is,

[8]Here, *Quantile[data,qn]* is just a specification to the function *FindArgMin*, telling it to find the minimum of |(Γ*pdf(x)/100n)-qn*|, starting at x-value *mu*, specified as the desired quantile of the raw data, which Mathematica conveniently has a built-in function to find.

the greater its effect on model parameters. That makes the far end of a long-tailed distribution less trustworthy than we would like it to be.

The left-hand end of this kind of distribution is also somewhat troubled, as we can see from the confidence intervals in Appendix A. Many of these confidence intervals tend to be wide on the left, which also happens to be a region of relatively few accidents. The very lowest TFH bin is often not far removed from the very tallest, in which case a small δ shift in the curve to the right or left, can accompany a large change in the amplitude parameter *A*, and/or a large shift in α and/or β. All this goes to show that highly sloped frequency distributions can sometimes be represented by a multiplicity of *pretty good* models which, nonetheless, may vary widely in parameter estimates, which differ most in their predictions at extreme values of *x*.

Therefore, as stated previously, we are prudent to put our greatest faith in the middle TFH ranges for these models, perhaps in the 50-5,000 TFH range. Fortunately, that is the range most interesting to most of us under most circumstances, because it captures the vast majority of accidents.

DISCUSSION

Figure 8a shows a frequency histogram commonly seen in general aviation (GA) accident analysis, namely accident count as a function of pilots' total flight hours (TFH). A modeling function would be useful to smooth the noise in such data, allowing investigators to better predict how likely pilots of a given experience level are to be involved in accidents. This would be useful, for instance, in allocating resources for pilot training or mentoring, and as the basis for a statistical covariate of flight risk.

In this report we tested a number of candidate modeling functions on eight samples of NTSB GA data encompassing the years 1983-2011. Appendix A shows that the gamma, log-normal, and Weibull probability density functions were all able to fit such data, given x-axis data bins 100 TFH wide. Estimates of goodness-of-fit (R^2) ranged from .86-.99 (good-to-excellent) and did not differ significantly across those three model classes.

Log-transformation of TFH proved critical to the success of these data-fits. Untransformed TFH (e.g., Fig. 7a) frequently led to catastrophic fit-failure.

The raw data exhibited an extremely long right-hand tail, due in part to pilot aging and dropout, but also to the confound of relatively few pilots having large numbers of commercial FH rolled into their GA TFH. The log-transform (Fig. 8b) effectively compressed this "long tail," allowing successful data-fit.

Although log-normal and Weibull functions appeared capable of fitting these data, there is some pragmatic advantage to using a gamma pdf (Equation 3). A gamma pdf allows estimation of confidence intervals around the fitting function itself (.95CI, Fig. 8b). The width of these confidence intervals, of course, is both a function of the sample size and inherent noise in the data.

Due to the nature of the data, it may be advisable to place the greatest prediction confidence in a middle range of TFH, perhaps from 50-5,000. Fortunately, that is also the range that captures the vast majority of all GA pilots.

Because more than one function class seems capable of fitting these data, no simple theoretical claims can be made about causation. Causation certainly involves multiple processes, some perhaps embodying sums of independent exponential decay processes (gamma), "failure rate" processes (Weibull), and multiplicative random processes (log-normal). Theorizing is hampered by a host of factors, such as the confounding of relatively safer commercially logged flight hours counted as TFH, the dropout of non-instrument-rated pilots to become instrument-rated, and that some phases of flight are more dangerous than others, meaning that flight risk is not merely linearly proportional to time spent aloft.

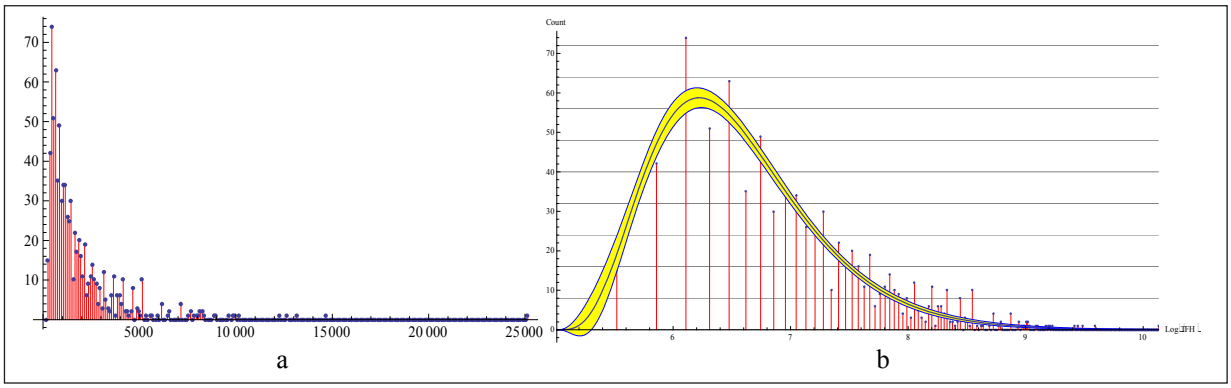


Figure 8. Frequency histograms of GA fatal accident count (y-axis) with a) untransformed, and b) natural log (ln)-transformed x-axis (which eventually proved essential to successful data-fitting).

Therefore, the goal of the current effort is largely atheoretical, being merely to show that such data can be modeled. With some care, GA accident frequencies can be predicted from TFH, given data parsed by a) pilot instrument rating and b) seriousness of accident. Goodness-of-fit (R^2) tended to be excellent for non-instrument-rated pilot data and good for instrument-rated data. Estimates of median TFH were derived for each dataset, which will be useful to aviation policy makers.

REFERENCES

Box, G.E.P., & Draper, N.R. (1987). *Empirical model-building and response surfaces*, New York: Wiley.

Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45: 41–51.

Chiew, F.H.S., Srikanthan, R., Frost, A.J., & Payne, E.G.I. (2005). Reliability of daily and annual stochastic rainfall data generated from different data lengths and data characteristics. In: *MODSIM 2005 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand*, Melbourne, December 2005, pp. 1223-1229.

Craig, P.A. (2001). *The killing zone.* New York: McGraw-Hill.

Glass, G.V., & Hopkins, K.D. (1984). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Hogg, R.V., & Klugman, S.A. (1984). *Loss distributions*. New York: Wiley.

Kirkpatrick, S., Gelatt, C.D., & Vecchi, M.P. (1983). Optimization by simulated annealing. *Science* 220 (4598): 671-80.

Mackey, J.B. (1998). Forecasting wet microbursts associated with summertime airmass thunderstorms over the southeastern United States. Unpublished MS Thesis, Air Force Institute of Technology.

National Transportation Safety Board. (2011). *User-downloadable database*. Downloaded May 26, 2011, from www.ntsb.gov/avdata

Ratkowsky, D.A. (1989). *Handbook of nonlinear regression models.* New York: Marcel Dekker.

Spanier, J. & Oldham, K.B. (1987). *An atlas of functions.* New York: Hemisphere.

Winkelmann, R. (2008). *Econometric analysis of count data.* (5th Ed.). Berlin: Springer-Verlag.

Wolfram Mathematica Documentation Center. *Some notes on internal implementation.* Downloaded April 29, 2011, from http://reference.wolfram.com/mathematica/note/SomeNotesOnInternalImplementation.html#20880

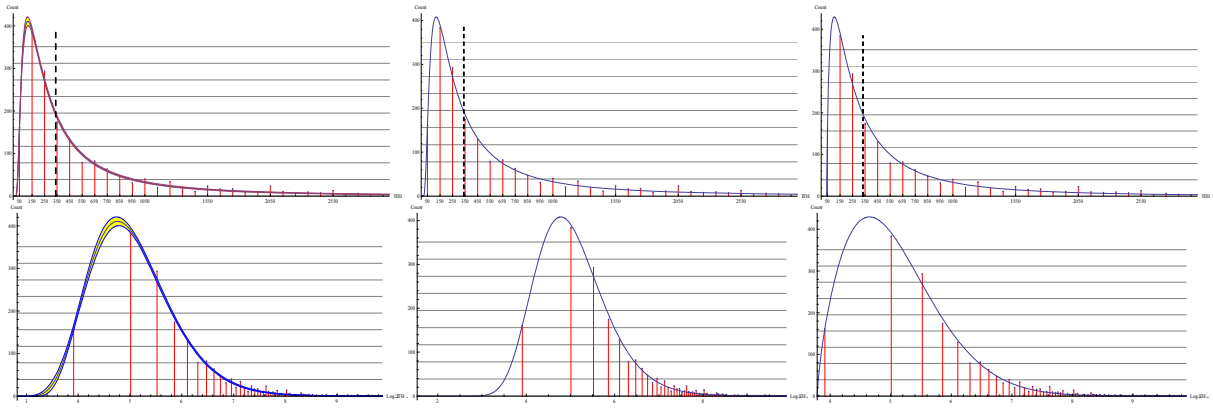
APPENDIX A

Comparing data fits for the eight datasets of Table 1. The x-axis represents TFH, the y-axis is accident frequency count. Dashed vertical lines show $\tilde{x}_{TFH}$, the median value of TFH dividing the area under the modeling curve area in half. A .95CI brackets each gamma pdf. The first row of models has a linear x-axis, the second row has a ln(x)-axis, which allows easier inspection of data fits between model classes.

Model Class

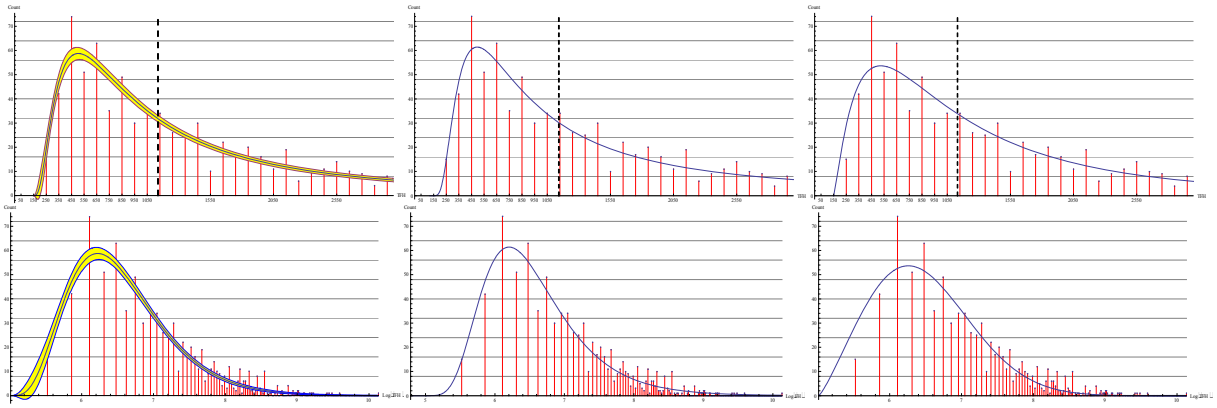
Gamma Log-Normal Weibull

1983-2000 non-instrument-rated fatal



n_{pilots}=1823	R^2=0.994	$\tilde{x}_{TFH}$=340	n_{pilots}=1823	R^2=0.994	$\tilde{x}_{TFH}$=340	n_{pilots}=1823	R^2=0.993	$\tilde{x}_{TFH}$=334
Parameter	**.95CI$_{lower}$**	**.95CI$_{upper}$**	**Parameter**	**.95CI$_{lower}$**	**.95CI$_{upper}$**	**Parameter**	**.95CI$_{lower}$**	**.95CI$_{upper}$**
A 776.3	765.8	786.8	A 777.1	766.9	787.4	A 782.0	767.2	796.9
α 7.789	5.996	9.572	μ 1.230	1.118	1.343	α 1.721	1.646	1.795
β 0.286	0.253	0.319	σ 0.228	0.204	0.250	β 1.431	1.376	1.486
δ 2.821	2.559	3.085	δ 1.533	1.151	1.915	δ 3.786	3.746	3.826
r 0.9968	0.9965	0.9971	r 0.9969	0.9966	0.9972	r 0.9964	0.9960	0.9967

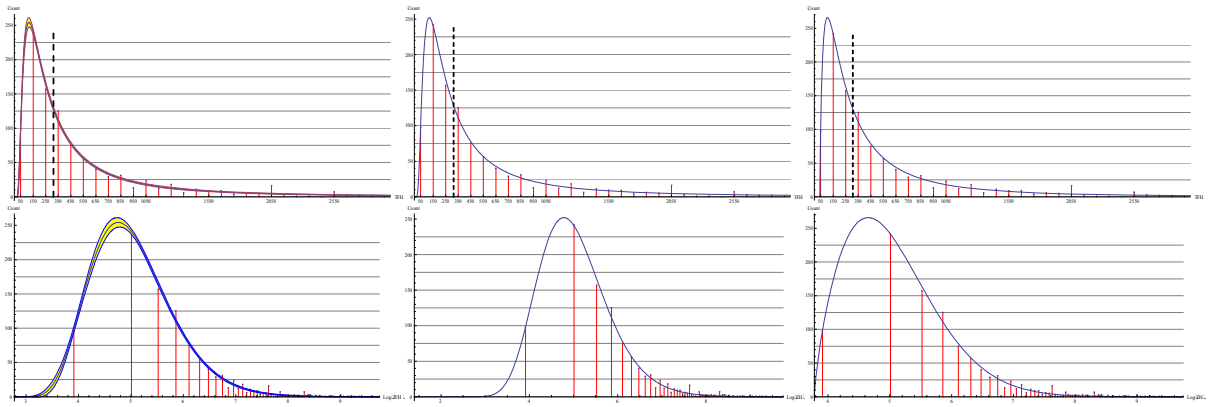
1983-2000 instrument-rated fatal



n_{pilots}=831	R^2=0.940	$\tilde{x}_{TFH}$=1135	n_{pilots}=831	R^2=0.942	$\tilde{x}_{TFH}$=1142	n_{pilots}=831	R^2=0.923	$\tilde{x}_{TFH}$=1131
Parameter	**.95CI$_{lower}$**	**.95CI$_{upper}$**	**Parameter**	**.95CI$_{lower}$**	**.95CI$_{upper}$**	**Parameter**	**.95CI$_{lower}$**	**.95CI$_{upper}$**
A 92.30	88.87	95.73	A 91.57	88.35	94.79	A 95.33	89.58	101.1
α 4.902	3.189	6.616	μ 0.500	0.360	0.640	α 2.217	1.819	2.614
β 0.311	0.252	0.369	σ 0.389	0.336	0.442	β 1.636	1.311	1.962
δ 5.011	4.730	5.291	δ 4.787	4.552	5.022	δ 5.011	4.656	5.365
r 0.9694	0.9650	0.9732	r 0.9706	0.9664	0.9743	r 0.9607	0.9551	0.9656

Gamma Log-Normal Weibull

1983-2000 non-instrument-rated serious



n_{pilots}=1051 R^2=0.994 $\tilde{x}_{TFH}$=312

Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 472.4	465.5	479.2
α 8.322	6.163	10.48
β 0.271	0.236	0.306
δ 2.776	2.476	3.077
r^A 0.9968	0.9964	0.9972

n_{pilots}=1051 R^2=0.994 $\tilde{x}_{TFH}$=313

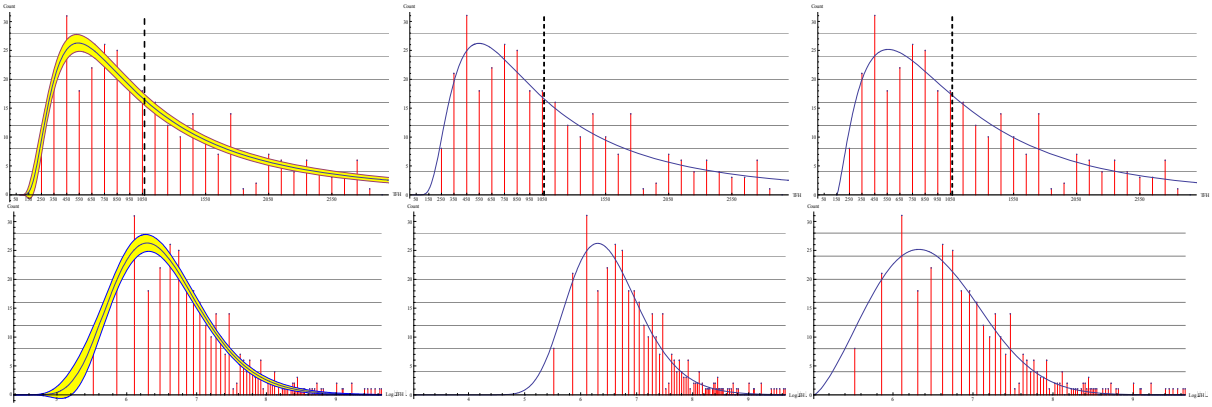
Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 472.4	465.6	479.2
μ 1.266	1.133	1.398
σ 0.216	0.189	0.243
δ 1.396	0.930	1.861
r^A 0.9968	0.9964	0.9972

n_{pilots}=1051 R^2=0.994 $\tilde{x}_{TFH}$=307

Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 475.8	466.6	485.0
α 1.746	1.667	1.825
β 1.418	1.361	1.475
δ 3.780	3.738	3.822
r^A 0.9968	0.9964	0.9972

[A]The equality of values for *r=0.9968* across distributions here is simply a coincidence.

1983-2000 instrument-rated serious



n_{pilots}=362 R^2=0.932 $\tilde{x}_{TFH}$=1067

Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 42.12	39.85	44.38
α 10.33	2.150	18.51
β 0.208	0.126	0.289
δ 4.357	3.449	5.266
r 0.9653	0.9575	0.9717

n_{pilots}=362 R^2=0.931 $\tilde{x}_{TFH}$=1063

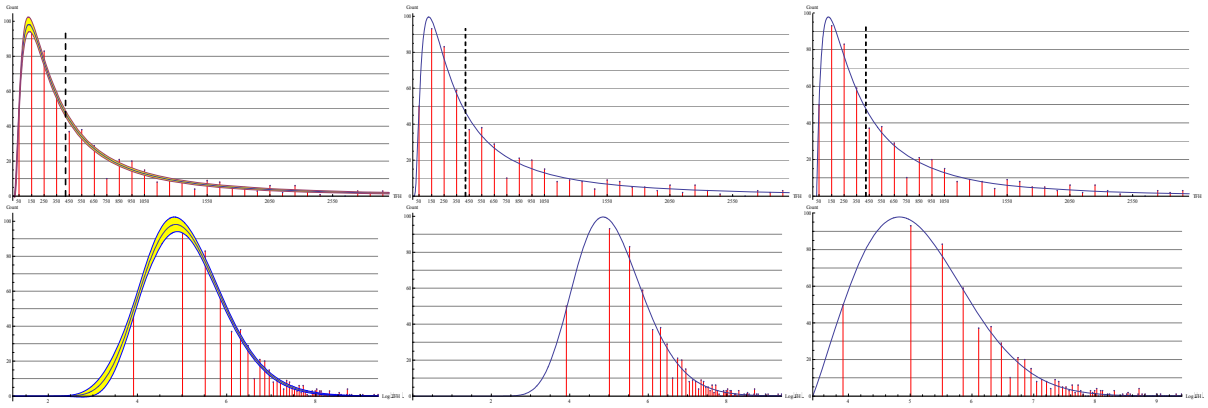
Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 42.36	40.07	44.65
μ 1.226	0.765	1.687
σ 0.193	0.109	0.277
δ 3.023	1.434	4.611
r 0.9651	0.9572	0.9715

n_{pilots}=362 R^2=0.931 $\tilde{x}_{TFH}$=1064

Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 42.35	39.61	45.09
α 2.397	1.878	2.916
β 1.644	1.260	2.028
δ 5.011	4.600	5.421
r 0.9650	0.9571	0.9715

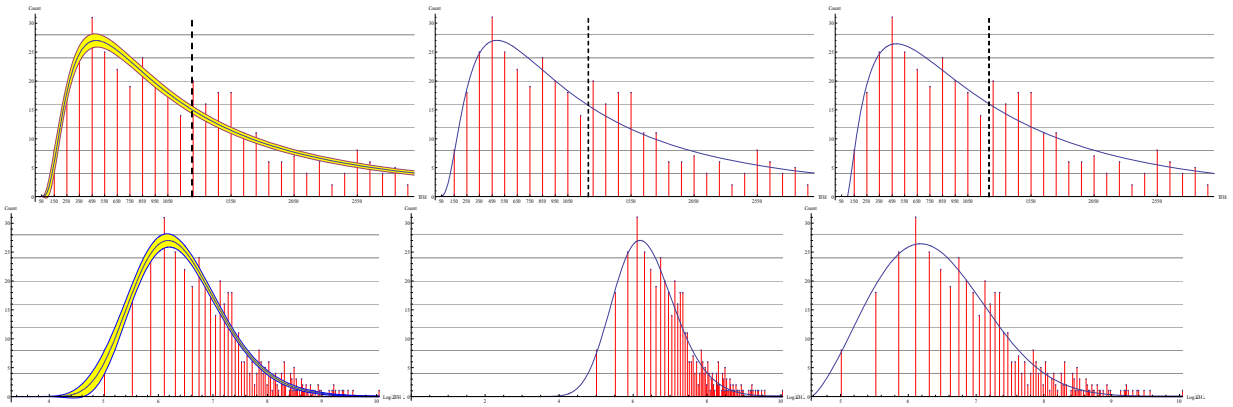
Gamma Log-Normal Weibull

2000-2011 non-instrument-rated fatal



n_{pilots}=571	R^2=0.983	$\tilde{x}_{TFH}$=421	n_{pilots}=571	R^2=0.982	$\tilde{x}_{TFH}$=421	n_{pilots}=571	R^2=0.982	$\tilde{x}_{TFH}$=424
Parameter	**.95CI$_{lower}$**	**.95CI$_{upper}$**	**Parameter**	**.95CI$_{lower}$**	**.95CI$_{upper}$**	**Parameter**	**.95CI$_{lower}$**	**.95CI$_{upper}$**
A 219.7	213.5	226.0	A 219.7	213.5	225.9	A 231.2	207.5	218.9
α 17.94	5.006	30.87	μ 1.611	1.305	1.916	α 2.083	1.855	2.309
β 0.216	0.141	0.290	σ 0.179	0.128	0.230	β 1.921	1.704	2.138
δ 1.203	0	2.695	δ 0	-1.534	1.534	δ 3.422	3.207	3.636
r 0.9914	0.9899	0.9927	r 0.9912	0.9896	0.9925	r 0.9912	0.9896	0.9925

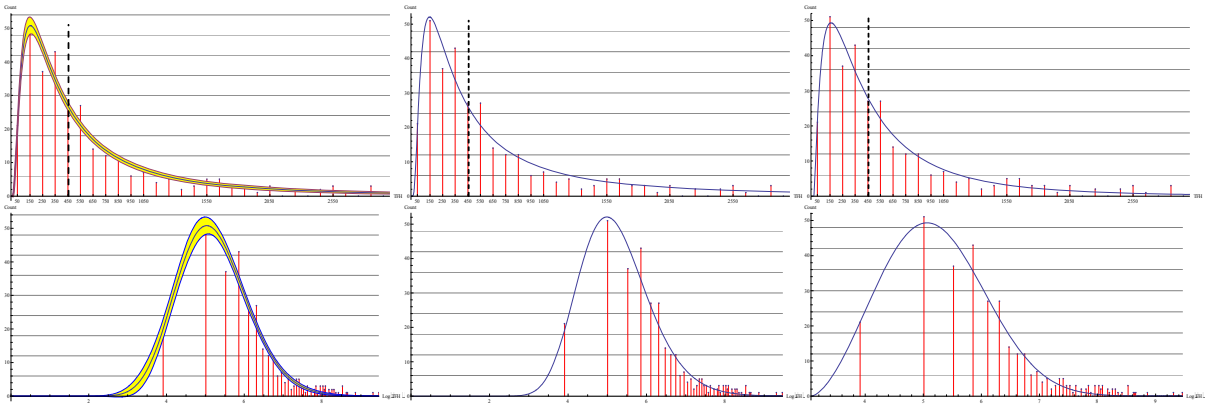
2000-2011 instrument-rated fatal



n_{pilots}=465	R^2=0.946	$\tilde{x}_{TFH}$=1241	n_{pilots}=465	R^2=0.947	$\tilde{x}_{TFH}$=1212	n_{pilots}=465	R^2=0.948	$\tilde{x}_{TFH}$=1218
Parameter	**.95CI$_{lower}$**	**.95CI$_{upper}$**	**Parameter**	**.95CI$_{lower}$**	**.95CI$_{upper}$**	**Parameter**	**.95CI$_{lower}$**	**.95CI$_{upper}$**
A 54.03	51.70	56.37	A 55.87	53.41	58.34	A 55.15	52.87	57.43
α 14.27	4.592	23.96	μ 1.840	1.351	2.330	α 2.380	2.099	2.660
β 0.218	0.146	0.289	σ 0.132	0.071	0.193	β 2.027	1.775	2.279
δ 3.288	2.148	4.429	δ 0	-3.104	3.105	δ 4.568	4.303	4.833
r 0.9726	0.9672	0.9771	r 0.9732	0.9679	0.9776	r 0.9736	0.9684	0.9779

Gamma Log-Normal Weibull

2000-2011 non-instrument-rated serious



n_{pilots}=328 R^2=0.967 $\tilde{x}_{TFH}$=455

Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 113.9	109.2	118.7
α 32.82	-4.546	70.19
β 0.158	0.071	0.246
δ 0	-3.082	3.082
r 0.9831	0.9791	0.9864

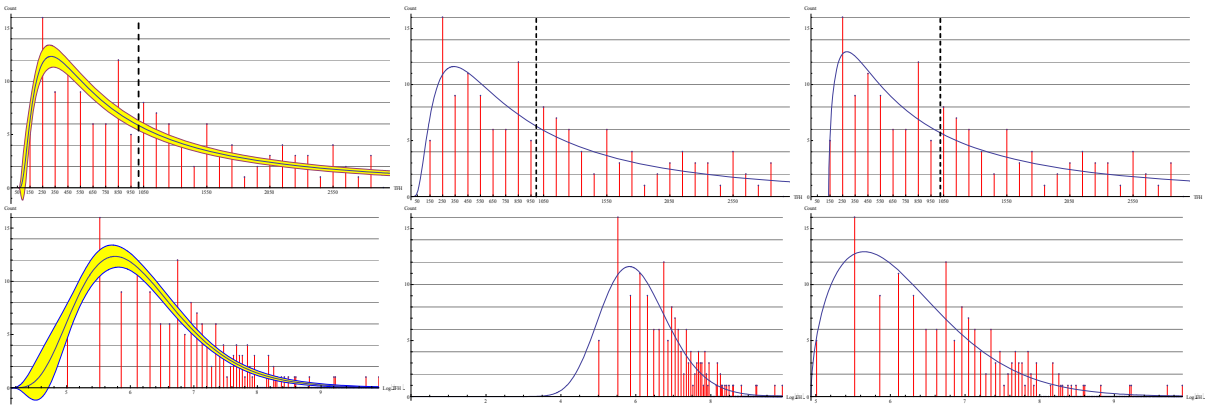
n_{pilots}=328 R^2=0.964 $\tilde{x}_{TFH}$=456

Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 113.9	109.1	118.7
μ 1.637	1.258	2.016
σ 0.172	0.110	0.234
δ 0	-1.947	1.947
r 0.9817	0.9773	0.9853

n_{pilots}=328 R^2=0.968 $\tilde{x}_{TFH}$=455

Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 112.2	107.6	116.8
α 2.635	2.188	3.081
β 2.396	1.972	2.821
δ 3.066	2.635	3.497
r 0.9841	0.9803	0.9872

2000-2011 instrument-rated serious



n_{pilots}=164 R^2=0.864 $\tilde{x}_{TFH}$=1010

Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 24.78	22.97	26.59
α 5.383	1.285	9.482
β 0.376	0.233	0.518
δ 4.115	3.248	4.982
r 0.9295	0.9051	0.9477

n_{pilots}=164 R^2=0.862 $\tilde{x}_{TFH}$=992

Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 26.11	23.58	28.64
μ 1.786	0.841	2.731
σ 0.152	0.021	0.283
δ 0	-5.711	5.711
r 0.9283	0.9035	0.9468

n_{pilots}=164 R^2=0.872 $\tilde{x}_{TFH}$=1024

Parameter	.95CI$_{lower}$	.95CI$_{upper}$
A 24.00	22.67	25.33
α 1.547	1.390	1.703
β 1.396	1.282	1.509
δ 4.936	4.851	5.022
r 0.9336	0.9106	0.9508

APPENDIX B

Estimates for the gamma pdf's parameters α and β can be calculated using the method of moments. The expected value (mean, or first population moment) of the gamma pdf is

$$E(x) = \alpha\beta \tag{16}$$

while the second population moment is

$$E(x^2) = \alpha(\alpha + 1)\beta^2 \tag{17}$$

From our data, we can estimate the first moment as

$$m_1 = \frac{\sum_{i=1}^{n} x_i}{n} \tag{18}$$

and the second moment as

$$m_2 = \frac{\sum_{i=1}^{n} x_i^2}{n} \tag{19}$$

where the binned data have been shifted to start at zero by subtracting the *x*-value of the first bin from all others.

Letting

$$\alpha\beta = m_1 \tag{20}$$

and

$$\alpha(\alpha + 1)\beta^2 = m_2 \tag{21}$$

we can solve for α and θ as

$$\alpha = \frac{m_1^2}{m_2 - m_1^2} \tag{22}$$

$$\beta = \frac{m_2 - m_1^2}{m_1} \tag{23}$$

Now, using these estimates for α and θ, plus the expected *x*-value for the mode (x_{Mo}) of the pdf,

$$Mo = (\alpha - 1)\beta \tag{24}$$

plus y_{max}, the maximum observed *y*-value in the data, the amplitude term (A) can be estimated as

$$A_{est} = \frac{y_{\max}}{y_{\mathrm{mod}\,e}} = \frac{y_{\max}}{\frac{e^{-x_{Mo}/\beta} x_{Mo}^{\alpha-1} \beta^{-\alpha}}{\Gamma(\alpha)}} = \frac{y_{\max}}{\frac{e^{-(\alpha-1)}((\alpha-1)\beta)^{\alpha-1} \beta^{-\alpha}}{\Gamma(\alpha)}} \tag{25}$$

APPENDIX C

```
(*This is the Mathematica code used to generate the  data for the paper *)
SetDirectory["M:\\New Projects\\NTSB Rec Popup\\Modeling accidents x FH"];  FileNames[];
filename = "Data 2000-2011 IR SERS.txt";  (* "Data 1983-2000 IR FATL.txt" is the only file that wouldn't
    converge with pre-estimated parameters *)
data = actualYs = predictedYs = {};    maxLogFH = 0;
myFile = OpenRead[filename];
onePoint = Read[myFile, Number];
While[ToString[onePoint] ≠ "EndOfFile",  AppendTo[data, onePoint];   onePoint = Read[myFile, Number] ];
Close[myFile];
n = Length[data];
F̃H = Median[data];   (*median flight hours *)

binSize = 100;
base = Floor[data[[1]], binSize];
binnedData = binnedLnData = BinCounts[data, binSize];  (*Tally data into bins of the specified size*)
nBins = Length[binnedLnData];  modeLnFH = 0;
For[i = 1, i ≤ nBins, i++,
    temp1 = binnedData[[i]];
    temp2 = binnedLnData[[i]];
          (* Add x-values for each bin, whose midpoint is halfway into the bin *)
    binnedData[[i]] = {N[base + (i - 0.5) *binSize], temp1 } ;
    binnedLnData[[i]] = {N[Log[base + (i - 0.5) *binSize]], temp2 } ;  (* Add Log-transformed x-bin values, similarly *)
    If[binnedLnData[[i, 2]] > modeLnFH, modeLnFH = binnedLnData[[i, 2]]]
  ];
minFH = binnedData[[1, 1]];  maxFH = binnedData[[nBins, 1]];
minLogFH = binnedLnData[[1, 1]];  maxLogFH = binnedLnData[[nBins, 1]];

Clear[A, α, β, μ, σ, δ, mu, μPDF, constraints, initialValues];
gammaModel = A *PDF[GammaDistribution[α, β], x - δ];
LogNormalModel = Re[A *PDF[LogNormalDistribution[μ, σ], x - δ]];
WeibullModel = Re[A *PDF[WeibullDistribution[α, β], x - δ]];  (* α is shape parm and β is scale parm *)
selectedModel =  gammaModel;

If[selectedModel == gammaModel,  modelName = "GammaModel"; constraints = {0 < A < 2000, α > 0, β > 0, minLogFH > δ > 0};
    logDataZeroed = N[Log[data]];  (*Here, in the ln[x] domain, we begin estimation of starting values for parameters *)
    logDataZeroed = logDataZeroed - logDataZeroed[[1]];
    m1 = Mean[logDataZeroed];     (* 1st moment *)
    m2 = Mean[logDataZeroed^2];  (* 2nd moment *)
    aa = m1^2 / (m2 - m1^2);  (*Estimate of α, based on data*)
    bb = (m2 - m1^2) / m1;  (*Estimate of β, based on data*)
    modeEstX = (aa - 1) bb;  (* Est. of the x-value of the mode *)
    modeEstY = PDF[GammaDistribution[aa, bb], modeEstX];  (* Est. of the y-value of gamma pdf at the mode *)
    AA = modeLnFH / modeEstY;  (* Est. of amplitude factor  needed, based on max val of data / y-value at the esti x-val of  mode*)
    dataMax = {1, -99};
    For[i = 1, i ≤ Length[binnedLnData], i++, If[binnedLnData[[i, 2]] > dataMax[[2]], dataMax = binnedLnData[[i]]]];
    dataXMax = dataMax[[1]];  dataYMax = dataMax[[2]];
    temp = FindMaximum[AA *PDF[GammaDistribution[aa, bb], x], x];
    pdfYMax = temp[[1]];   pdfXMax = temp[[2, 1, 2]];
    dd = dataXMax - pdfXMax;  (* pdfXMax will also = (aa-1) bb *)
    Print["m1: ", m1, "   m2: ", m2, "   aa: ", aa, "   bb:  ", bb, "   modeExtX: ", modeEstX, "   modeEstY: ",
      modeEstY, "   AA: ", AA, "   dd: ", dd ];
    initialValues = {{A, AA}, {α, aa}, {β, bb}, {δ, dd}};
    If[filename == "Data 1983-2000 IR FATL.txt",  initialValues =  {{A, 162}, {α, 6.5}, {β, .37}, {δ, 4}}];
  ];
```

```
If[selectedModel == LogNormalModel, modelName = "LogNormal model"; constraints = {0 < A < 5000, σ > 0, minLogFH > δ > 0};
    If[filename == "Data 1983-2000 NIR FATL.txt", initialValues = {{A, 777}, {μ, 1.23}, {σ, 0.228}, {δ, 1.53}}];
    If[filename == "Data 1983-2000 IR FATL.txt", initialValues = {{A, 92}, {μ, 0.5}, {σ, .39}, {δ, 4.8}}];
    If[filename == "Data 1983-2000 NIR SERS.txt", initialValues = {{A, 472}, {μ, 1.27}, {σ, .22}, {δ, 1.4}}];
    If[filename == "Data 1983-2000 IR SERS.txt", initialValues = {{A, 42}, {μ, 1.23}, {σ, 0.19}, {δ, 3}}];
    If[filename == "Data 2000-2011 NIR FATL.txt", initialValues = {{A, 220}, {μ, 1.6}, {σ, .18}, {δ, 0.001}}];
    If[filename == "Data 2000-2011 IR FATL.txt", initialValues = {{A, 56}, {μ, 1.86}, {σ, .13}, {δ, 0.001}}];
    If[filename == "Data 2000-2011 NIR SERS.txt", initialValues = {{A, 114}, {μ, 1.6}, {σ, .17}, {δ, 0.01}}];
    If[filename == "Data 2000-2011 IR SERS.txt", initialValues = {{A, 26}, {μ, 1.8}, {σ, .15}, {δ, 0.01}}];
];
If[selectedModel == WeibullModel, modelName = "Weibull Model"; constraints = {0 < A < 2000, α > 0, β > 0, minLogFH > δ > 0};
    If[filename == "Data 1983-2000 NIR FATL.txt", initialValues = {{A, 782}, {α, 1.72}, {β, 1.43}, {δ, 3.8}}];
    If[filename == "Data 1983-2000 IR FATL.txt", initialValues = {{A, 95}, {α, 2.22}, {β, 1.64}, {δ, 5}}];
    If[filename == "Data 1983-2000 NIR SERS.txt", initialValues = {{A, 475}, {α, 1.75}, {β, 1.4}, {δ, 3.8}}];
    If[filename == "Data 1983-2000 IR SERS.txt", initialValues = {{A, 42}, {α, 2.4}, {β, 1.64}, {δ, 5.0}}];
    If[filename == "Data 2000-2011 NIR FATL.txt", initialValues = {{A, 213}, {α, 2}, {β, 1.9}, {δ, 3.4}}];
    If[filename == "Data 2000-2011 IR FATL.txt", initialValues = {{A, 55}, {α, 2.38}, {β, 2.0}, {δ, 4.6}}];
    If[filename == "Data 2000-2011 NIR SERS.txt", initialValues = {{A, 112}, {α, 2.6}, {β, 2.4}, {δ, 3.1}}];
    If[filename == "Data 2000-2011 IR SERS.txt", initialValues = {{A, 24}, {α, 1.55}, {β, 1.4}, {δ, 4.9}}]
];
nlm = NonlinearModelFit[binnedLnData, {selectedModel, constraints}, initialValues, x, AccuracyGoal → 4, PrecisionGoal → 4,
    MaxIterations → 200, Method → "Automatic"] (*, Gradient→"FiniteDifference"] ]*)

If[selectedModel == gammaModel, A = nlm[[1, 2, 1, 2]]; α = nlm[[1, 2, 2, 2]]; β = nlm[[1, 2, 3, 2]]; δ = nlm[[1, 2, 4, 2]];
    linearModel[x_] := A β^-α (Log[x] - δ)^(α-1) E^(-(Log[x]-δ)/β) / Gamma[α]; (*This is the linear version, using parameters est'd in the log domain *)
    minf = Solve[linearModel[x] == 0, x][[1, 1, 2]];
    μPDF = FindArgMin[Abs[(NIntegrate[linearModel[x], {x, minf, mu}] / (n * binSize)) - 0.5], {mu, Median[data]}][[1]];
    Print["A→", A, " α=", α, " β=", β, " δ=", δ, " μPDF=", μPDF, " n=", n]; (* n*binSize = the area under the curve *)
];
If[selectedModel == LogNormalModel, A = nlm[[1, 2, 1, 2]]; μ = nlm[[1, 2, 2, 2]]; σ = nlm[[1, 2, 3, 2]]; δ = nlm[[1, 2, 4, 2]];
    linearModel[x_] := Re[A E^(-(-μ+Log[Log[x]-δ])^2 / (2 σ^2)) / ((Log[x]-δ) σ)] / Sqrt[2 π];
    (*The numerical estimate of the same definite integral. Throws an error msg, but works *)
    μPDF = FindArgMin[Abs[(NIntegrate[linearModel[x], {x, E^δ, mu}] / (n * binSize)) - 0.5], {mu, Median[data]}][[1]];
    Print["A→", A, " μ=", μ, " σ=", σ, " δ=", δ, " μPDF=", μPDF, " n=", n]
];
If[selectedModel == WeibullModel, A = nlm[[1, 2, 1, 2]]; α = nlm[[1, 2, 2, 2]]; β = nlm[[1, 2, 3, 2]]; δ = nlm[[1, 2, 4, 2]];
    linearModel[x_] := Re[A E^-((Log[x]-δ)/β)^α α β^-α (Log[x] - δ)^(-1+α)];
    (*The x-value of the log prediction fcn, above & below which = 50% area under the curve*)
    μPDF = FindArgMin[Abs[(NIntegrate[linearModel[x], {x, E^δ, mu}] / (n * binSize)) - 0.5], {mu, Median[data]}][[1]];
    Print["A→", A, " α=", α, " β=", β, " δ=", δ, " μPDF=", μPDF, " n=", n]
];
Block[{x}, For[i = 1, i ≤ Length[binnedLnData], i++, AppendTo[actualYs, binnedLnData[[i, 2]]]; x = binnedLnData[[i, 1]];
    AppendTo[predictedYs, selectedModel]]]; (*End Block*)
maxActualYs = Max[actualYs];
gridlines = {}; For[i = 0, i < Ceiling[maxActualYs, 10], i += (Ceiling[maxActualYs, 10] / 10), AppendTo[gridlines, i]];
r = Correlation[actualYs, predictedYs];
```

```
Z = 0.5 Log[(1 + Abs[r])/(1 - Abs[r])];        (* Fisher's Z-transform of the correlation r.  See Glass & Hopkins, p. 304-307 *)
σz = N[1/Sqrt[n - 3]];        (* Standard error of Fisher's Z *)
r_upper CI = Tanh[Z + 1.96 σz];  (*Hyperbolic tangent.  See Glass & Hopkins, p. 305.10 *)
r_lower CI = Tanh[Z - 1.96 σz];
Print["r = ", r, "   Upper .95CI = ", r_upper CI, "   Lower .95CI = ", r_lower CI, "   r^2 = ", r^2];
Print["MMCA R squared:  ", nlm["RSquared"], "    MMCA Adj R squared:  ", nlm["AdjustedRSquared"]];
Print["MMCA ParameterConfidenceIntervalTable: ", nlm["ParameterConfidenceIntervalTable", ConfidenceLevel → .95]];
(*Print["MMCA CovarianceMatrix: ",TableForm[nlm["CovarianceMatrix"]]];*)
(*Print["MMCA ANOVATable: ",hd=nlm["ANOVATable"],"   MSE: ",nlm["ANOVATableMeanSquares"]];*)
linearXRange = 3000;
linearAxesTicks = {50, 150, 250, 350, 450, 550, 650, 750, 850, 950, 1050, 1550, 2050, 2550};
g1 = ListPlot[binnedLnData, PlotRange → All, Filling → Axis, FillingStyle → Red];  (*These already have a log-scaled x *)
CI95[x_] = Table[nlm["MeanPredictionBands", ConfidenceLevel → .95], {x}];
g2 = Plot[{nlm[x], CI95[x]}, {x, δ, maxLogFH}, Filling → {1 → {{2}, Yellow}}, PlotRange → All, PlotStyle → {Thickness[.002],
        RGBColor[0, 0, 1]}];
g3 = Graphics[{Thickness[.002], Dashing[.01], Line[{{μPDF, 0}, {μPDF, maxActualYs}}]}];

g5 = ListPlot[binnedData, PlotRange → {{0, linearXRange}, All}, Filling → Axis, FillingStyle → Red];  (*These  have a log-scaled x *)
g6 = Plot[{linearModel[x], CI95[Log[x]]}, {x, E^δ, linearXRange}, Filling → {1 → {{2}, Yellow}}, PlotRange → All,
   PlotStyle → {Thickness[.002]}];
theLabel = StringJoin[modelName, ",  ", filename, ",   n_accidents = ", ToString[n], ",   n_bins = ", ToString[nBins],
   "  .50 Quantile (μPDF) = ", ToString[μPDF]]
Show[g1, g2, ImageSize → {800, 400}, PlotRange → {{δ, maxLogFH}, All}, AspectRatio → Full, AxesLabel → {"Log[TFH]", "Count"},
  GridLines → {None, gridlines}, AxesOrigin → {δ, 0}]
Show[g5, g6, g3, ImageSize → {800, 400}, PlotRange → {{0, linearXRange}, All}, AspectRatio → Full, AxesLabel → {"TFH", "Count"},
  GridLines → {None, gridlines}, AxesOrigin → {0, 0}, Ticks → {linearAxesTicks, Automatic}]
NIntegrate[linearModel[x], {x, E^δ, μPDF}]/(n * binSize)  (* Double-check calculation of the median area under the curve *)

(* This is the method of finding quantiles *)
qn = 0.9;  (* 0<qn<1 *)
FindArgMin[Abs[(NIntegrate[linearModel[x], {x, minf, mu}]/(n * binSize)) - qn], {mu, Quantile[data, qn]}][[1]]
```