

Use of Truncated Regression Methods to Estimate the Shelf Life of a Product from Incomplete Historical Data

William Q. Meeker
Dept. of Statistics
Iowa State University
Ames, IA 50011

Luis A. Escobar
Dept. of Experimental Statistics
Louisiana State University
Baton Rouge, LA 70803

August 10, 2001
17h 27min

Abstract

Over a period of time, experiments were conducted to estimate the shelf life of a product. Each trial used a combination of a temperature level and an additive concentration that was used to inhibit spoilage. The policy was to terminate each trial after 270 days, even if the product sample had not yet failed. Particularly at the lower temperatures, some trials ended before the product sample reached the failed state. No records were kept on the number of unfailed samples. Thus the resulting data were truncated. This paper describes the analysis of the resulting data and the methods that were used to estimate the shelf life distribution of the product.

Key words: Censored data, Maximum likelihood, Reliability, Response Surface, Time-Varying Environment.

1 Introduction

1.1 Background and motivation

An additive is used to extend the shelf life of a product (to protect the identity of this product, we will refer to it as Product A). Higher levels of the additive will extend shelf life, but will also add cost and detract from product performance. The producer would like to use the least amount of additive possible, consistent with adequate control of the risk of spoilage before use (an event that we will refer to as “failure”).

Over a long period of time, data were collected to investigate the statistical relationship between Product A shelf life and concentration of additive for different levels of temperature and to monitor other product characteristics. If all specimens put on test had been observed until failure, classical response surface methods could have been used to estimate the effect that concentration and temperature have on Product A shelf life. If the survival times for all unfailed specimens had been recorded, then standard censored data methods could have been used to estimate the relationship. The available data for Product A, however, had been entered into a computer database from old paper records. Only the failure times had been recorded. Notations concerning what would have been censored observations were non-numerical and were never entered into the data base. Although it was known that the upper time limit of observation had been 270 days for the product units that were monitored, there was no available information on the number of trials for which the specimens were unspoiled.

The goal of the analysis of the historical data was to estimate the relationship between life and the two explanatory variables: additive concentration and temperature. The relationship was to be used to make better decisions on how much additive should be added to the product. Although all experimentation was done at 45°C or above, there was interest in the distribution of shelf life down to 35°C. The original Product A data are proprietary. Therefore we have used simulated data with some modification to the scaling of the parameters in order to illustrate the steps that were followed in the original analysis.

1.2 The Product A shelf life experimental data

The available Product A shelf life data are available in the SPLIDA/S-PLUS life data object ShelfLifeA.ld and also in the SPLIDA distribution as a text file. The concentration of the additive is in units of parts per million (ppm) and the storage temperature is in degrees Celsius. The time to failure (indicated by evidence of a particular observable event related to spoilage) was recorded in terms of number of days. The last two columns in the data set give the truncation information (indicating the time at which the trial would have been terminated had the product sample not failed). In this application, the right truncation time was the same for all units, but the methodology used here allows for different truncation times for different trials, as long as the truncation times are known for all observations.

Figure 1 shows the number of recorded shelf lives at the different levels of temperature and

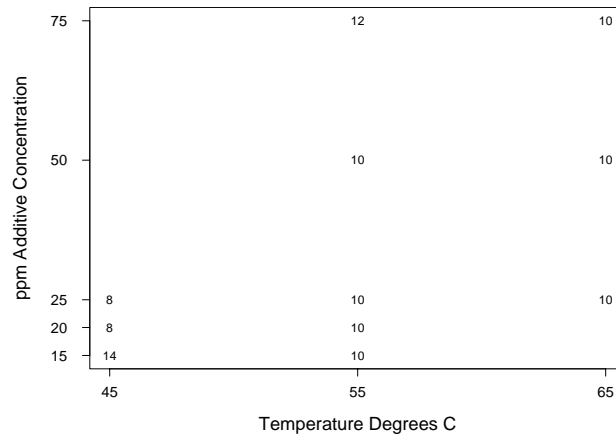


Figure 1: Temperature and concentration combinations used in the shelf life study.

concentration levels. Figure 1 suggests that the allocation of samples to the particular levels of concentration and temperature over time were chosen in a somewhat haphazard manner. Recall that the actual number of samples at each condition is unknown due to the truncation. The logic for avoiding the SE (high temperature with low concentration) and NW (low temperature with high concentration) regions is that these were conditions that would not be used in practice. Experimentation below 45°C was avoided because of the large amount of time needed to observe failures.

Figure 2 is a plot of Product A shelf life versus temperature, conditioned on the unique values of concentration. This plot suggests the propensity for shorter shelf life at higher temperatures. Figure 3 is a similar plot of Product A shelf life versus concentration, conditioned on the unique values of temperature. This plot suggests the propensity for longer shelf life at higher levels of concentration. In both of these figures, at combinations of the factors implying long life, the truncation point at 270 days can be seen.

1.3 Related literature

Truncated distributions and truncated data have been discussed in a number of places in the statistical literature (truncated data are data generated from a truncation distribution). The books by Cohen (1991) and Schneider (1986) describe parametric methods of estimation and inference for truncated and censored data from a single distribution. Turnbull (1976) and Nelson (1990a) provide nonparametric methods of estimation for truncated data. Kalbfleisch and Lawless (1992) describe several different reliability applications involving truncated data. Meeker (1987) shows that there is a relationship between an LFP model and right truncation. Kalbfleisch and Lawless (1988) illustrate the loss of efficiency of truncation relative to censoring. Escobar and Meeker (1998) present methods for computing asymptotic variances that would allow such comparisons in other situations. Chap-

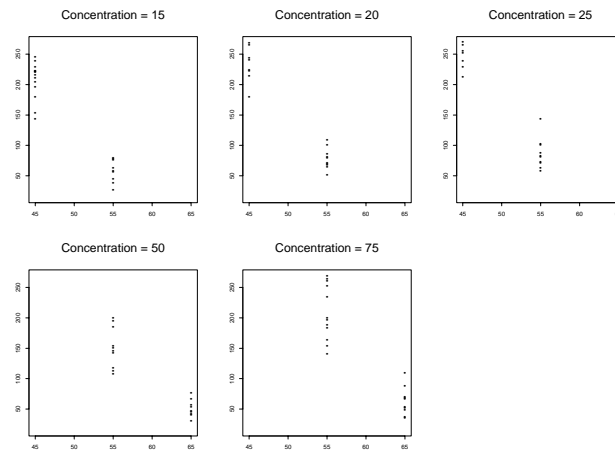


Figure 2: Plot of Product A shelf life versus temperature, conditioned on the unique values of concentration.

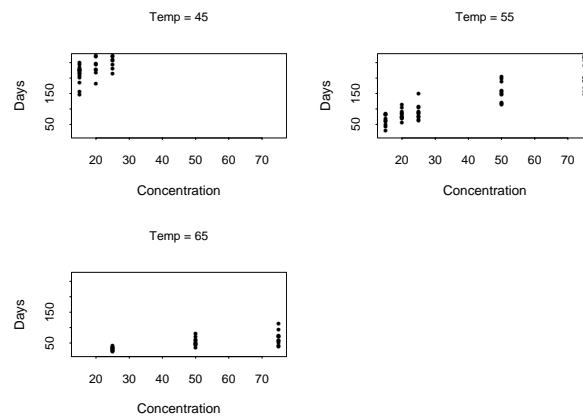


Figure 3: Plot of Product A shelf life versus concentration, conditioned on the unique values of temperature.

ter 11 of Meeker and Escobar (1998) presents a number of examples of left and right truncation, describes ML estimation, and suggests a probability plotting method to help assess distributional goodness of fit. Here we will review some of the important ideas, as they pertain to right truncation.

There are fewer references involving truncated data in regression analysis, and we were unable to find any applications in reliability data analysis. Amemiya (1973) describes economic applications and provides a detailed description of the theory behind ML estimators for regression with normally distributed residuals and *censored* data (which he refers to as “truncated”). In the last section of the paper, however, the author presents an argument that the results in the paper apply to an alternative model that is the same as the truncated data model presented here and elsewhere. Kalbfleisch and Lawless (1991) use a right-truncated regression model to analyze AIDS incubation times. Pagano, Tu, De Gruttola, and MaWhinney (1994) use truncated regression for the same application. Gross and Huber-Carol (1992) also describe regression models for truncated data.

1.4 Overview

Section 2 provides some background for estimation with truncated data, using data at a single experimental condition to make the presentation simpler. Section 3 describes and illustrates the use of models and methods for truncated data with explanatory variables. Section 4 extends the example given in Section 2 to give a comparison between the analysis of truncated data and censored data. Section 5 contains some concluding remarks and suggestions for further research.

2 Truncated Data Background

2.1 Analysis of truncated data at a single condition

This section introduces the ideas behind the analysis of truncated data by focusing on the results of the trials with a concentration of 25 ppm and a temperature of 45°C. There were 8 observations (trials with failures) at this condition. Figure 4 is an event plot of these data. The line at row 9 extending to 270 days indicates the unknown number of truncated observations. If we knew the number of samples that had been terminated at 270 days, the data would be right-censored instead of right-truncated.

Figure 5 is a lognormal probability plot. The lognormal cdf and pdf are

$$F(t; \mu, \sigma) = \Phi_{\text{nor}} \left[\frac{\log(t) - \mu}{\sigma} \right] \quad (1)$$

$$f(t; \mu, \sigma) = \frac{1}{\sigma t} \phi_{\text{nor}} \left[\frac{\log(t) - \mu}{\sigma} \right], \quad t > 0 \quad (2)$$

where ϕ_{nor} and Φ_{nor} are, respectively, the standardized normal pdf and cdf. The lognormal median $t_{.5} = \exp(\mu)$ is a scale parameter and $\sigma > 0$ is the lognormal shape parameter.

The straight line in Figure 5 is the ML estimate of the untruncated lognormal distribution of the Product A shelf life distribution at 25 ppm and 45°C, based on the truncated data. The agreement

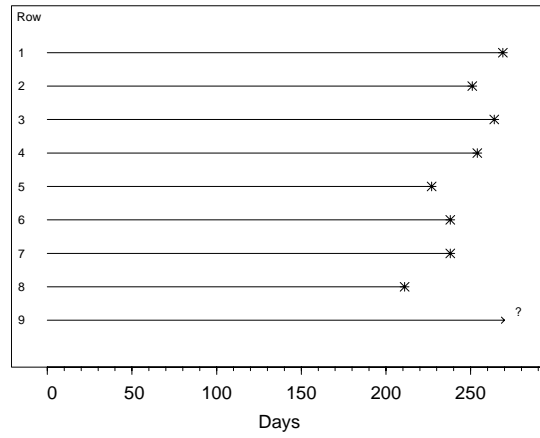


Figure 4: Event plot of the Product A shelf life data at 25 ppm and 45°C.

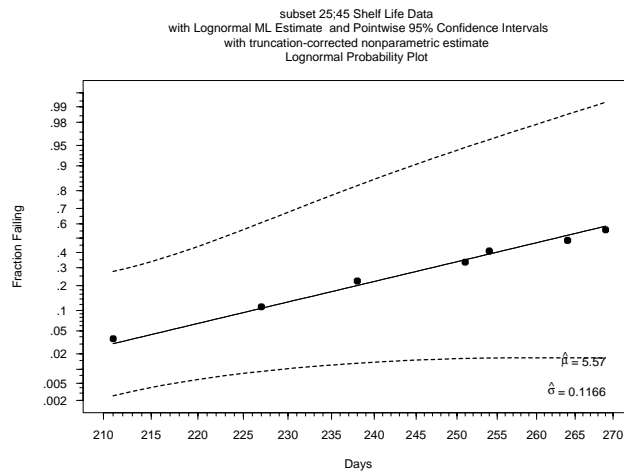


Figure 5: Lognormal probability plots of the Product A shelf life data at 25 ppm and 45°C.

Table 1: Lognormal ML estimates for the Product A shelf life data at 25 ppm and 45°C.

```
subset 25;45 Shelf Life Data
Maximum likelihood estimation results:
Response units: Days
Lognormal Distribution
Log likelihood at maximum point: -33.34
```

Parameter	Approx Conf. Interval			
	MLE	Std.Err.	95% Lower	95% Upper
mu	5.5703	0.15216	5.27204	5.8685
sigma	0.1166	0.07326	0.03406	0.3994

between the plotted points and the parametric ML estimate is very strong in part because of the parametric adjustment to the conditional nonparametric estimate described in Section 2.2.

If the random variable T_i is truncated when it lies above τ_i^U then the likelihood of an observation at $t_i \leq \tau_i^U$ is

$$L_i(\boldsymbol{\theta}) = \frac{f(t_i; \boldsymbol{\theta})}{F(\tau_i^U; \boldsymbol{\theta})} \quad (3)$$

where $\boldsymbol{\theta} = (\mu, \sigma)$. The total log likelihood for a sample of n independent truncated observations is

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_i^n \log[L_i(\boldsymbol{\theta})]. \quad (4)$$

The values of $\boldsymbol{\theta} = (\mu, \sigma)$ that maximize (4) are the maximum likelihood estimates and these are denoted by $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma})$. The ML estimation results for the Product A shelf life data at 25 ppm and 45°C are summarized in Table 1. The maximum of the log likelihood is at $\hat{\mu} = 5.5703$ and $\hat{\sigma} = 0.1166$. The standard errors in Table 1 were obtained from the observed information matrix.

Figure 6 is a plot of the relative likelihood function $L(\mu, \sigma)/L(\hat{\mu}, \hat{\sigma})$ for the Product A shelf life data at 25 ppm and 45°C. Although the likelihood appears to have a clear maximum, the likelihood contours indicate a ridge-like behavior running from the SW to the NE corner of this plot. This is an indication of poor identifiability of the lognormal distribution parameters. In particular, the data at this condition are not entirely capable of distinguishing between a distribution with a small value of μ with a small amount of truncation and a large value of μ with a large amount of truncation.

2.2 Adjustment for probability plotting of truncated data

Probability plots are useful for displaying data and parametric estimates of the cdf, and for making an assessment of distributional goodness of fit. A truncated (log) location-scale distribution is not, in general, a (log) location-scale distribution. Thus special methods are needed to do a probability plot. Section 11.6 of Meeker and Escobar (1998) shows how to parametrically adjust a truncated-data nonparametric estimator so that it can be used for making a probability plot. The basic idea

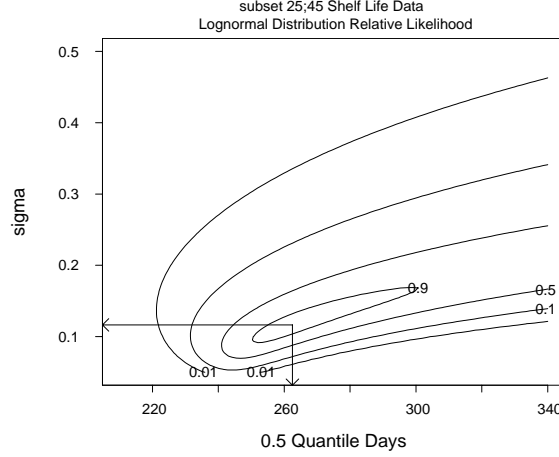


Figure 6: Contour plot of the lognormal likelihood function for the Product A shelf life data at 25 ppm and 45°C.

is to use the ML estimate of an underlying parametric distribution to provide an estimate of the fraction truncated and to use this probability to adjust the nonparametric estimate, providing an estimate of the unconditional distribution that generated the truncated response. This approach can be used in situations with right truncation, left truncation or both left and right truncation. We will describe the method as it applies to the right-truncated shelf life data and extend the application of the method to situations in which there are explanatory variables. With right truncation, the nonparametric estimate provides an estimate for the conditional probability given by the truncated distribution

$$F_C(t) = \Pr(T \leq t | 0 \leq T < \tau_{\max}^U) = \frac{F(t)}{F(\tau_{\max}^U)}, \quad 0 < t \leq \tau_{\max}^U \quad (5)$$

where τ_{\max} is the largest right-censoring time in the sample. Then the parametrically-adjusted unconditional nonparametric (NPU) estimate of $F(t)$ is

$$\hat{F}_{NPU}(t) = \hat{F}_{NPC}(t)F(\tau_{\max}^U; \hat{\theta}), \quad 0 < t \leq \tau_{\max}^U \quad (6)$$

where \hat{F}_{NPC} is the nonparametric estimate of F_C . This adjusted nonparametric estimate can be plotted on probability paper in the usual way (i.e., plotting the middle probability point in the jump at each observation, as described in Section 6.4 of Meeker and Escobar 1998).

Figure 5 is an adjusted probability plot for the Product A shelf life data at 25 ppm and 45°C. The points, corresponding to the adjusted nonparametric estimate, fall very much along a straight line. This indicates that there is no evidence in the data to suggest a departure from a lognormal distribution. Note, however, that in this case, there is little information in the data about the shape of the underlying distribution. A similar plot, done for the Weibull distribution, also provided an adjusted estimate of $F(t)$ that plotted almost as a straight line on the Weibull probability plot.

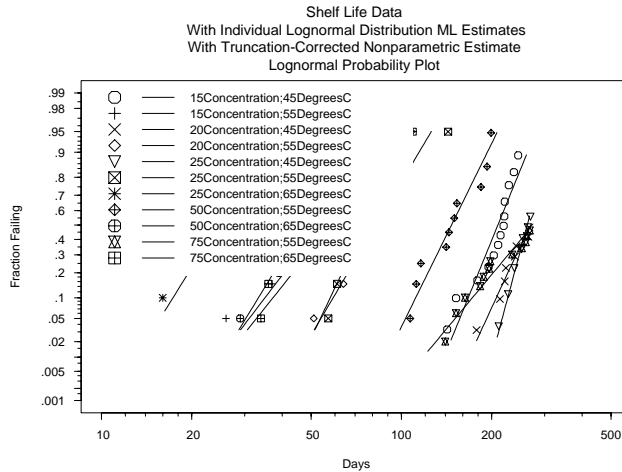


Figure 7: Multiple individual lognormal ML estimates and probability plots (with truncation correction) for the Product A shelf life data.

One conclusion from this analysis is that there is rather little information in the available data at this condition, relative to an untruncated sample or even a censored sample with the same number of reported failures. In Section 4, we will use these same data to do a direct comparison between truncated data and censored data to show the differences in information content.

We have presented detailed results for one combination of the explanatory variables. Similar results (with similar conclusions) were obtained for the other combinations of temperature and concentration and these will be summarized in the next section.

3 Truncated Regression Model for the Product A Shelf Life Data

This section describes models, graphical methods, and model estimation methods for the shelf life data, using all of the combinations of the explanatory variables. The models and methods presented here are similar to those used in Nelson (1990b) and Meeker and Escobar (1998), extended to truncated data.

3.1 Individual analyses at different experimental conditions

Figure 7 is a multiple probability plot, similar to Figure 5, but showing the results simultaneously for all of the combinations of levels of temperature and concentration. Figure 8 is the same as Figure 7, but without the legend, making it possible to see all of the results clearly. Each of the nonparametric estimates was adjusted individually with (6), based on ML estimates of the lognormal $F(\tau_{\max}^U; \hat{\theta})$, using the data from the corresponding individual levels of concentration and temperature. The

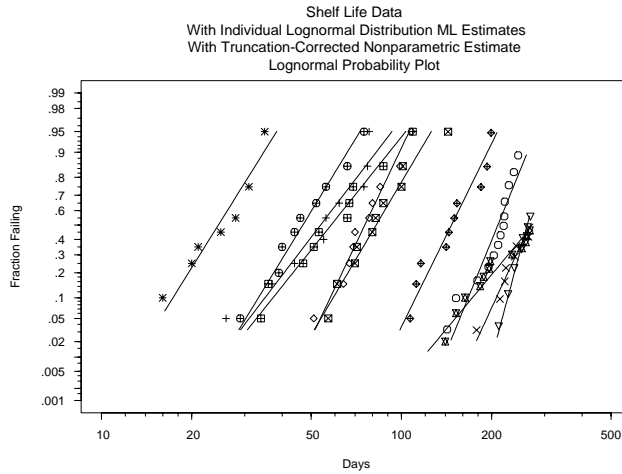


Figure 8: Multiple individual lognormal ML estimates and probability plots (with truncation correction) for the Product A shelf life data.

different slopes of the ML estimate lines in Figure 8 indicate some variation in the estimates of the lognormal shape parameters. Table 2 provides a summary of the estimation results for the individual combinations of levels.

3.2 Floating-scale model

Figure 9 is a multiple probability plot that is similar to Figure 8, but the fitted model is analogous to the standard one-way ANOVA model. In this model, there is a separate lognormal scale parameter for each combination of temperature and concentration. In contrast with the individual analysis, however, there is a common lognormal shape parameter. We call this the “floating-scale model.” The ML estimate lines are parallel because of the constraint that the lognormal shape parameters be the same.

In Figure 9, each of the nonparametric estimates (represented by the plotted points) was adjusted individually using (6), based on ML estimates of the lognormal $F(\tau_{\max}^U; \hat{\theta})$ from the constant shape parameter floating-scale model. The ML estimation results for the floating-scale model are summarized in Table 3. The intercept corresponds to the lognormal $\mu_{15\text{Concentration};45\text{DegreesC}}$ parameter. The other coefficients are estimates of the deviation $\mu_i - \mu_{15\text{Concentration};45\text{DegreesC}}$ for each of the other combinations of conditions. A comparison of the results between Figures 8 and 9 (corresponding to the individual analyses and floating-scale model, respectively) allows an assessment of whether the differences in the shape parameter estimates at different combinations of temperature and concentration can be explained by chance. The log likelihood ratio statistic for the comparison is $Q = 2 \times [-492.3 - (-495.8)] = 7.0$. The difference in the number of parameters estimated in the two models is $22 - 12 = 10$. The approximate p -value for the test comparing these

Table 2: ML estimates of the individual lognormal parameters (μ, σ) at distinct factor-level combinations for the Product A shelf life data.

Shelf Life Data

Maximum likelihood estimation results:

Response units: Days

Lognormal Distribution

	Concentration, DegreesC	Log likelihood	mu	se_mu	sigma	se_sigma
1	15Concentration;45DegreesC	-67.43	5.345	0.06398	0.1775	0.04919
2	15Concentration;55DegreesC	-43.03	3.984	0.10531	0.3330	0.07447
3	20Concentration;45DegreesC	-36.58	5.591	0.32337	0.2014	0.14034
4	20Concentration;55DegreesC	-41.69	4.324	0.06553	0.2072	0.04633
5	25Concentration;45DegreesC	-33.34	5.570	0.15216	0.1166	0.07326
6	25Concentration;55DegreesC	-44.71	4.410	0.08134	0.2572	0.05753
7	25Concentration;65DegreesC	-33.16	3.199	0.08600	0.2720	0.06081
8	50Concentration;55DegreesC	-48.55	4.989	0.06864	0.2146	0.05041
9	50Concentration;65DegreesC	-39.43	3.848	0.08415	0.2661	0.05950
10	75Concentration;55DegreesC	-60.11	5.614	0.51552	0.3627	0.21500
11	75Concentration;65DegreesC	-44.28	4.065	0.11008	0.3481	0.07788

Total log likelihood= -492.3

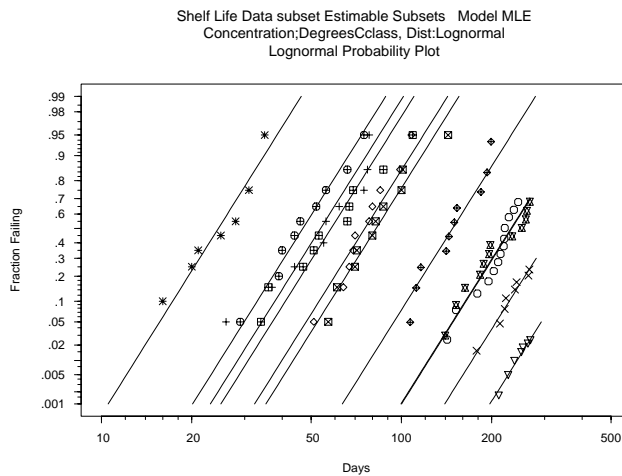


Figure 9: Multiple individual lognormal probability plots (with floating-scale model truncation correction) and ML estimate with common σ for the Product A shelf life data.

Table 3: Lognormal ML estimate of the floating-scale model for the Product A shelf life data.

Response units: Days

Lognormal Distribution

Log likelihood at maximum point: -495.8

Parameter	Approx Conf. Interval			
	MLE	Std.Err.	95% Lower	95% Upper
Intercept	5.451724	0.10645	5.2431	5.6604
15Concentration;55DegreesC	-1.467888	0.13719	-1.7368	-1.1990
20Concentration;45DegreesC	0.331074	0.22107	-0.1022	0.7644
20Concentration;55DegreesC	-1.128057	0.13719	-1.3969	-0.8592
25Concentration;45DegreesC	0.677770	0.00000	0.6778	0.6778
25Concentration;55DegreesC	-1.041426	0.13718	-1.3103	-0.7725
25Concentration;65DegreesC	-2.252341	0.13719	-2.5212	-1.9835
50Concentration;55DegreesC	-0.454238	0.13810	-0.7249	-0.1836
50Concentration;65DegreesC	-1.603737	0.13719	-1.8726	-1.3349
75Concentration;55DegreesC	-0.005434	0.14846	-0.2964	0.2855
75Concentration;65DegreesC	-1.386938	0.13719	-1.6558	-1.1181
sigma	0.273665	0.02176	0.2342	0.3198

two different models is $\Pr(\chi_{10}^2 > 7) = 0.2745$, indicating that the differences in the slopes of the fitted lines relative to the the common slope of the floating-scale model can be explained by chance alone.

3.3 Response surface model for life as a function of concentration and temperature

The response surface model suggested by the scientists responsible for quantifying the shelf life of Product A was

$$\begin{aligned}\mu &= \beta_0 + \beta_1 \log(\text{Concentration}) + \beta_2 \left(\frac{120.279}{\text{temp}^\circ\text{C} + 273.15} \right) \\ \sigma &= \text{Constant.}\end{aligned}\tag{7}$$

The first and second terms in (7) correspond to a power function relationship relating life and concentration [i.e., $\text{Days} = \alpha (\text{Concentration})^{\beta_1}$]. The third term is based on the well-known Arrhenius rate reaction model and 120.279 is the reciprocal of the gas constant in units of kJ/mole. Table 4 provides a numerical summary of the lognormal ML estimation results and Figure 10 provides a graphical summary. The ML estimate lines are again parallel because of the constant-shape parameter restriction. In Figure 10, each of the nonparametric estimates (represented by the plotted points) was adjusted individually using (6), based on ML estimates of the lognormal $F(\tau_{\max}^U; \hat{\theta})$ from

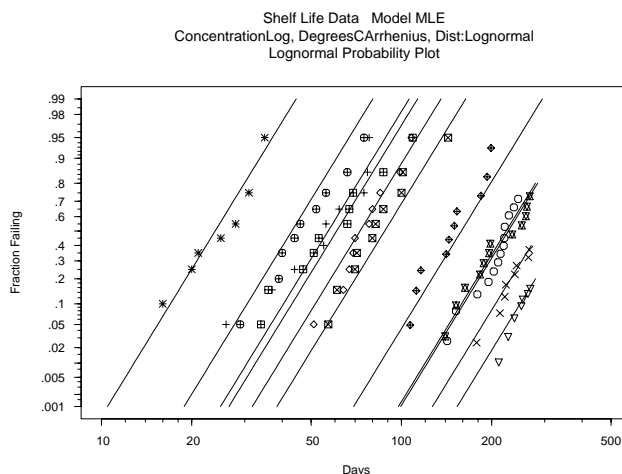


Figure 10: Multiple individual lognormal probability plots (with regression-model based truncation correction) and ML estimate of the concentration-temperature response surface model for the Product A shelf life data.

the response surface model in (7).

The coefficient $\hat{\beta}_1 = 0.8489$ for Concentration is the estimated power in a power law relationship between life and concentration. The coefficient $\hat{\beta}_2 = 120.05$ for DegreesC can be interpreted as the activation energy (in units of kJ/mole per °C) for the chemical reaction affecting the shelf life of Product A at a constant concentration. The value of $\hat{\sigma} = 0.2674$ implies that the variability in shelf life corresponding to a one- σ deviation at a given temperature and concentration is approximately 27% of the nominal response level.

Figure 10 can be used to assess the adequacy of the response surface model. The assessment is done by comparing the nonparametric estimates (the plotted points) with the fitted $F(t)$ lines, at each of the combinations of the experimental factors. The regression relationship in (7) is a further constraint in fitting a model to the data. Thus, there will be more deviations between the nonparametric estimates and the fitted $F(t)$ lines in Figure 10 than in Figure 9.

To test whether such deviations are statistically important, as opposed to being explainable by the natural variability in the data under the model in (7), we can again do a likelihood ratio test, this time comparing the results in Tables 3 and 4 (corresponding to the floating-scale model and the response surface model, respectively). The total log likelihood values from these models are -495.8 and -498.1 , respectively. The log likelihood ratio statistic for the comparison is $Q = 2 \times [-495.8 - (-498.1)] = 4.6$. The difference in the number of parameters estimated in the two models is $12 - 4 = 8$. The approximate p -value for the test comparing these two different models is $\Pr(\chi_8^2 > 4.6) = 0.201$, indicating that the differences in positions of the fitted lines from the response surface model, relative to the floating-scale model estimates, can be explained by chance alone.

We also fit a response surface model with a term for interaction between concentration and

Table 4: ML estimates of the linear response surface model for the Product A shelf life data.

Lognormal Distribution

Relationship(s)

1 Concentration: Log

2 DegreesC: Arrhenius kJ/mole

Log likelihood at maximum point: -498.1

Parameter	Approx Conf. Interval			
	MLE	Std.Err.	95% Lower	95% Upper
Intercept	-42.2604	1.97280	-46.1270	-38.3938
Concentration	0.8489	0.05939	0.7324	0.9653
DegreesC	120.0534	5.08663	110.0838	130.0230
sigma	0.2674	0.02061	0.2299	0.3110

temperature (details are not given here), but there was no indication of interaction in the data and all inferences were similar to those provided by using the response surface model in (7).

3.4 Estimates of shelf life at particular storage conditions from the response surface model

Management wanted an estimate of the shelf life distribution as a function of concentration for a nominal storage temperature of 35°C. Figure 11 is similar to Figure 10, but shows the extrapolation to the storage conditions of 25 ppm and 35°C. Figure 12 is a conditional model plot showing the ML estimates of the lognormal shelf life distributions for various fixed levels of concentration ranging between 10 and 80 ppm for fixed temperature at 35°C, based on the results for the response surface model summarized in Table 4. Table 5 gives the ML estimates and approximate confidence intervals for the quantiles of the shelf life distribution at the storage conditions of 25 ppm and 35°C. Consider $t_{.01}$, the time at which a fraction .01 of the product population will fail at these conditions. The estimation results indicate that we are 95% confident that $t_{.01}$ is between 650 and 1036 days.

4 Comparison of Truncated and Censored Data Analysis

Section 2 illustrated estimation from truncated data of Product A shelf at 25 ppm and 45°C. Here we compare the truncated data analysis with a censored data analysis, assuming that there had been a particular number of right-censored observations at 270 days. To get good agreement between the point estimates in the comparison, we used the truncated data model estimates in Section 2 to estimate that there had been 6 right-censored observations. Then, the actual number of trials at 25 ppm and 45°C would have been $8 + 6 = 14$. We refer to these data as the “pseudo-censored” data. The main purpose of this comparison is to illustrate the important loss in precision that results from

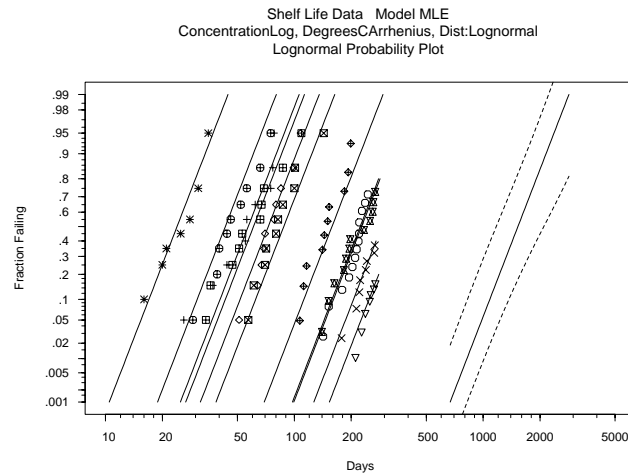


Figure 11: Multiple individual lognormal probability plots (with regression-model based truncation correction) and ML estimate of the concentration-temperature response surface model for the Product A shelf life data showing extrapolation to the storage conditions of 25 ppm and 35°C.

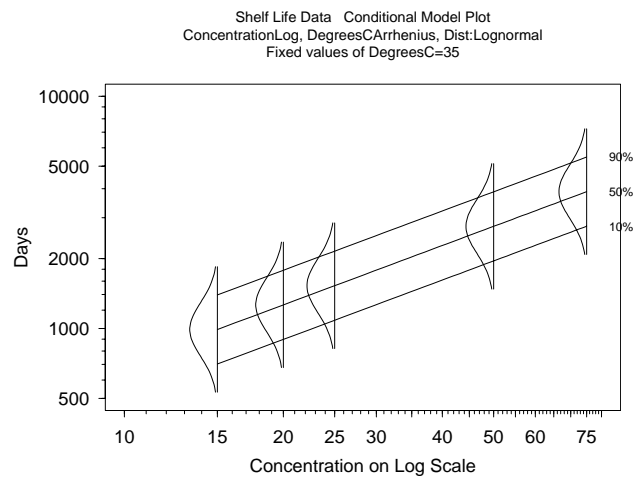


Figure 12: Conditional model plot of Product A shelf life versus concentration at 35°C.

Table 5: ML estimates and approximate confidence intervals for the quantiles of the shelf life distribution at the storage conditions of 25 ppm and 35°C.

Using Shelf Life Data at 25Concentration;35DegreesC

Parametric ML Quantile Estimates

Pointwise Approximate 95% Confidence Intervals

Lognormal Distribution

p	Quantile	Std.Err.	95% Lower	95% Upper
0.001	668.84	79.462	529.90	844.20
0.005	767.47	91.116	608.14	968.54
0.010	820.42	97.649	649.72	1035.98
0.050	984.43	119.075	776.65	1247.80
0.100	1084.88	133.045	853.09	1379.65
0.200	1220.32	152.832	954.71	1559.84
0.300	1328.37	169.345	1034.67	1705.42
0.400	1428.23	185.145	1107.78	1841.37
0.500	1528.35	201.472	1180.36	1978.93
0.600	1635.48	219.452	1257.27	2127.46
0.700	1758.44	240.696	1344.66	2299.53
0.800	1914.12	268.467	1454.06	2519.74
0.900	2153.09	312.834	1619.52	2862.45
0.990	2847.12	451.851	2086.01	3885.94

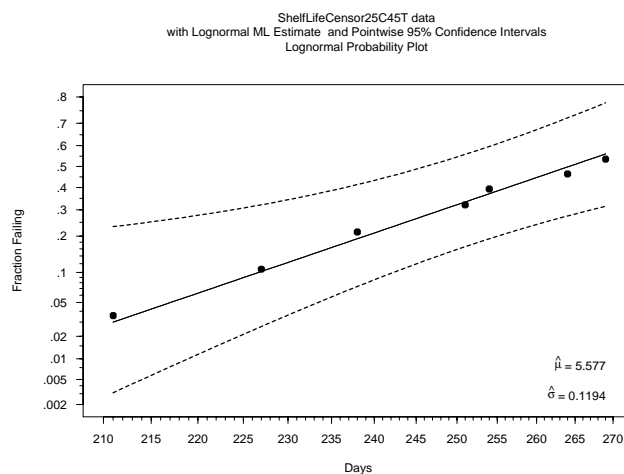


Figure 13: Lognormal probability plots of the pseudo-censored Product A shelf life data at 25 ppm and 45°C.

the truncated data caused by not recording the number of censored observations.

Figure 13 shows the ML estimation results for the pseudo-censored Product A shelf life data at 25 ppm and 45°C. The corresponding numerical results are summarized in Table 6. Comparing Figures 5 and 13 shows the large improvement in precision that would result if the number of censored observations could be recovered. Relatedly, note the improvement in the standard errors and confidence interval widths given in Table 6. Some insight into the reasons for the improvements in precision can be obtained by comparing the contour plots in Figures 6 and 14. In particular, the closed likelihood contours in Figure 13 indicate that the shelf life median can be bounded statistically over the region of the plot, which was not possible in Figure 5. The ambiguity between the amount of truncation and the location of the upper tail of the shelf life distribution has been, to a large

Table 6: Lognormal ML estimates for the Product A shelf life data at 25 ppm and 45°C.

```
subset 25;45 Shelf Life Data
Maximum likelihood estimation results:
Response units: Days
Lognormal Distribution
Log likelihood at maximum point: -42.9
```

Parameter		Approx Conf. Interval	
	MLE Std.Err.	95% Lower	95% Upper
mu	5.5766	0.03677	5.50449 5.6486
sigma	0.1194	0.03282	0.06969 0.2046

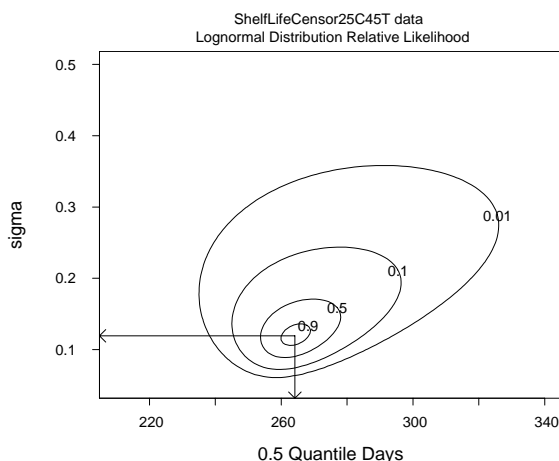


Figure 14: Contour plot of the lognormal likelihood function for the pseudo-censored Product A shelf life data at 25 ppm and 45°C.

degree, resolved by “knowing” the number of censored observations. In effect, there is no good information about the fraction failing after 270 days. These remarks extend to the fitting of the regression model considered in Section 3.3.

5 Concluding Remarks and Extensions

This paper has shown how to extract useful information out of an historical set of data with a certain type of missing data. In this case, especially because there was probably a number of conditions with complete or nearly complete data (as indicated by an absence of observations near to 270 degrees at some combinations of concentration and temperature), the final answers were rather precise. This application provides a good example of how statistical methods can be adapted to deal with “messy data.”

In other situations, it will be impossible to address completely important questions of interest without having to make strong assumptions or to rely on information coming from outside of the available data. For example, it may be necessary to assume that certain parameters are known. A more flexible alternative would be to obtain prior distributions for such parameters (e.g., from engineering knowledge or physical/chemical theory) and to use Bayesian methods instead. Because of recent advances in computing technology and the development of methods for Monte Carlo Markov Chain methods of inference, such methods are now practicable.

For the Product A shelf life example, needed inferences were for storage/shipping at a given constant temperature. In other applications, however, temperature (or other environmental factors) will vary over time. In such cases, it is possible to use a model similar to the one used in this paper to make predictions about life under a specified environmental profile. Methods for doing this are

presented in Nelson (2001).

Acknowledgments

We would like to thank Victor Chan and Katherine Meeker for helpful comments on a previous version of this paper.

A SPLIDA Commands for the Analyses

This appendix gives explicit direction on how to use the SPLIDA/S-PLUS software to do the the analyses described in this paper.

1. The data [Concentration in ppm (parts per million), failure “times” in Days, and truncation information (all observations were right-truncated at 270 days)] are in the SPLIDA data frame `ShelfLifeA`. Use `Splida -> Make/edit/summary/view data object -> Make life data object` to create the SPLIDA life data object `ShelfLifeA.ld`. Choose `Days` as the response, `Temp` and `Concentration` as the explanatory variables, `Truntime` as the truncation time, and `TrunType` as the Truncation ID.
2. Make a conditional scatter plot by using the SPLIDA/S-PLUS commands `ConditionalPlot("Days","Temp","Concentration", data = ShelfLifeA)` and `ConditionalPlot("Days","Concentration","Temp", data = ShelfLifeA)`.
3. To make a life data object for the data at a concentration of 25 ppm and a temperature of 45°C, use `Splida -> Multiple regression (ALT) data analysis -> Make a subset life data object`. Choose the `ShelfLifeA.ld` data object and highlight both explanatory variables. Then highlight 25;45 and click on “Apply.”
4. To make the event plot of the data at 25 ppm and 45°C, use `SPLIDA -> Single distribution data analysis -> Life Data event plot`. Choose the data object `ShelfLifeA.25Concentration.45DegreesC.ld` and click on “Apply.”
5. To make the probability plot and fit the truncated lognormal distribution to the data at 25 ppm and 45°C, use `SPLIDA -> Single distribution data analysis -> Probability plot with parametric ML fit`, choose the lognormal distribution, and click on “Apply.”
6. To make the likelihood contour plot for the data at 25 ppm and 45°C, use `SPLIDA -> Single distribution data analysis -> Likelihood contour/perspective/profile plots`, choose the lognormal distribution, check the contour box, change the quantile to .5, and click on “Apply.”
7. To obtain a probability plot showing lognormal distributions fitted to the individual conditions, use `Splida -> Multiple regression (ALT) data analysis -> Probability plot`

and ML fit for individual conditions. On the Basic page, choose the ShelfLife.lf life data object, the lognormal distribution, and both Concentration and DegreesC as the explanatory variables. Click on “Apply.” Visit the Plot options page and choose the suppress option in the legend box. Click on “Apply” again.

8. To obtain a multiple probability plot and fit of the floating-scale model, use Splida -> Multiple regression (ALT) data analysis -> Prob plot and ML fit for indiv conditions. On the Basic page, choose the lognormal distribution and both Concentration and DegreesC as the explanatory variables. Visit the Plot options page and choose the suppress option in the legend box. Click on “Apply.”
9. The Arrhenius model will be used to fit the response model. To match the output in this paper is is necessary to override the default choice for the units of the Arrhenius model regression coefficient. Use Splida -> Change SPLIDA default options (preferences), visit the Misc page, and choose kJ/mole.
10. To obtain a multiple probability plot and fit of the response surface model, use Splida -> Multiple regression (ALT) data analysis -> Probability plot and ML fit for a regression(acceleration) model. On the Basic page, choose the lognormal distribution and both Concentration and DegreesC as the explanatory variables. Visit the Model page and click on the relationships button. Choose log for Concentration and Arrhenius for DegreesC. Enter 25;35 in the Additional levels for evaluation box. Visit the Plot options page and choose the suppress option in the legend box. Visit the Tabular output page, check “print table” under Quantile estimates, and choose 25;35 from the “Level(s)” list. Click on “Apply.”
11. To obtain a conditional model plot of the shelf life distributions as a function of concentration at 35°C, use Splida -> Multiple regression (ALT) data analysis -> Conditional model plot. On the Basic page, choose the the results object ShelfLifeA.groupm.Lognormal.CncnLog.DgrCArrhlognormal and select Concentration as the variable to vary. Click on the choose button and enter 35 for the temperature. Click on “Apply.”

Questions:

1. Make scatter plots of the response versus temperature, conditioning on concentration level and vice-versa. Describe how you might modify these plot to reflect the truncation.
2. Re-do the scatter plots with different transformations on the variables. From looking at the the scatter plots, decide which scales make the relationship between life and the explanatory variables appear to be approximately linear.
3. Is it reasonable to assume that the shape parameter of the lognormal distribution σ does not depend on Concentration? How can you tell?

4. Use the results of the analysis in Section 3 to compute the ML estimate of the .01 quantile of life at 25°C if the additive concentration is 40 ppm.
5. What, approximately, would be a safe level of concentration for product that needs to last at least 360 days at 25°C with the probability of failing being no more than .05?
6. Try fitting some alternative relationships. In particular, try the linear and square root relationships for the explanatory variables. What effect does this have on the answer to part 4?
7. The analyses in this paper used the lognormal distribution. Re-do the analyses with a Weibull distribution. How do the results compare? Comment both on the model diagnostics and the final answers.
8. Conduct the following simulation to compare the effects of censoring and truncation on estimation precision.
 - (a) Generate a sample from a lognormal distribution with $\eta = \exp(\mu) = 200$ hours and $\sigma = 0.5$.
 - (b) Find the ML estimates of the parameters and $t_{.1}$, treating any observations beyond 250 hours as right censored.
 - (c) Repeat the simulation 1000 times. Make appropriate plots of the sample estimates (including scatter plots to see correlation). Compute and use histograms or other graphical displays to compare the estimates from the censored and the truncated samples. Also compute the sample variances for the parameter estimates and for the estimates of $t_{.1}$.
 - (d) What can you conclude from this simulation experiment?
9. Repeat Question 8 using $t_{.1}$ in place of $t_{.9}$. What do you conclude?
10. Consider the results from Question 8. Provide an intuitive explanation for the reason that precision from the censored distribution is much better than that from the truncated distribution.
11. Derive the truncated cdf used to construct the likelihood contribution in (3).

References

- Amemiya, T. (1973), Regression analysis when the dependent variable is truncated normal, *Econometrica*, **40**, 997-1016.
- Cohen, A. C. (1991), *Truncated and Censored Samples. Theory and Applications*, New York: Marcel Dekker.

- Escobar, L.A. and Meeker, W.Q. (1998), The asymptotic covariance matrix for maximum likelihood estimators with models based on location-scale distributions involving censoring, truncation, and explanatory variables. *Statistica Sinica* **8**, 221-237.
- Gross, S. T., and Huber-Carol, C. (1992), Regression models for truncated survival data, *Scandinavian Journal of Statistics*, **19**, 193-213.
- Kalbfleisch, J. D., and Lawless, J. F. (1988), Estimation of reliability in field-performance studies, *Technometrics*, **30**, 365-388 (with discussion).
- Kalbfleisch, J. D., and Lawless, J. F. (1991), Regression models for right truncated data with applications to AIDS incubation times and reporting lags, *Statistica Sinica*, **1**, 19-32.
- Kalbfleisch, J. D., and Lawless, J. F. (1992), Some useful statistical methods for truncated data, *Journal of Quality Technology*, **24**, 145-152.
- Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, New York: John Wiley & Sons.
- Meeker, W. Q. (1987) Limited failure population life tests: application to integrated circuit reliability. *Technometrics* **29**, 151-165.
- Meeker, W. Q., and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*. New York: John Wiley & Sons.
- Nelson, W. (1990a), Hazard plotting of left truncated life data, *Journal of Quality Technology*, **22**, 230-238.
- Nelson, W. (1990b), *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses*, New York: John Wiley & Sons.
- Nelson, W. (2001), Prediction of field reliability of units, each under differing dynamic stresses from accelerated test data, in *Reliability* Rao, C. R. and Balakrishnan, N., Elsevier Science, Amsterdam.
- Pagano, M., Tu, X. M., De Gruttola, V., and MaWhinney, S. (1994), Regression analysis of censored and truncated data: Estimating reporting-delay distributions and AIDS incidence from surveillance data, *Biometrics*, **50**, 1203-1214.
- Schneider, H. (1986), *Truncated and Censored Samples for Normal Populations*, New York: Marcel Dekker.
- Turnbull, B. W. (1976), The empirical distribution function with arbitrary grouped, censored, and truncated data, *Journal of the Royal Statistical Society*, **38**, 290-295.