Statistics Preprints

Statistics

9-2008

The importance of identifying different components of a mixture distribution in the prediction of field returns

Yili Hong
Iowa State University

William Q. Meeker
Iowa State University, wqmeeker@iastate.edu

The importance of identifying different components of a mixture distribution in the prediction of field returns

Abstract

Data from a mixture of distributions with two different increasing hazard functions can behave, over some period of time, like a distribution with decreasing hazard functions. As a result, reliability predictions based on data from a mixture of units with two or more different physical designs could be seriously wrong if the pooled data are used to extrapolate in time. Thus, it is important to identify components of the mixture and do statistical inference based on the stratified data. In this paper, the importance of this principle is investigated analytically and illustrated with lifetime data on high-voltage transformers. From engineering knowledge, the lifetime distribution of a transformer has an increasing hazard due, largely, to insulation aging. However, data from a population of units could indicate a decreasing hazard due to a mixture of units with different designs or environmental conditions. Comparisons are made between the predictions based on the pooled-data and stratified-data models an the importance of correct stratification in practice is shown. Some suggestions for practitioners are also given.

Disciplines

Statistics and Probability

Comments

The Importance of Identifying Different Components of a Mixture Distribution in the Prediction of Field Returns

Yili Hong and William Q. Meeker

Department of Statistics

Iowa State University

Ames, IA, 50011

September 25, 2008

Abstract

Data from a mixture of distributions with two different increasing hazard functions can behave, over some period of time, like a distribution with decreasing hazard functions. As a result, reliability predictions based on data from a mixture of units with two or more different physical designs could be seriously wrong if the pooled data are used to extrapolate in time. Thus, it is important to identify components of the mixture and do statistical inference based on the stratified data. In this paper, the importance of this principle is investigated analytically and illustrated with lifetime data on high-voltage transformers. From engineering knowledge, the lifetime distribution of a transformer has an increasing hazard due, largely, to insulation aging. However, data from a population of units could indicate a decreasing hazard due to a mixture of units with different designs or environmental conditions. Comparisons are made between the predictions based on the pooled-data and stratified-data models and the importance of correct stratification in practice is shown. Some suggestions for practitioners are also given.

Key Words: Hazard function; maximum likelihood; stratification; transformer reliability; Weibull

1

1   Introduction

1.1   The Problem

It is well known that data from a mixture of two different distributions with increasing hazard functions can behave, over some period of time, like from a distribution with a decreasing hazard function (see Meeker and Escobar 1998, page 119 for a simple example). Thus, it is possible for predictions based on data from this kind of heterogeneous population to lead to seriously incorrect conclusions. The lifetime distribution of a product is expected to have an increasing hazard function if the unit fails due to aging (wearout). For example, if a preliminary analysis of the pooled data indicates a decreasing or constant hazard function for what is known to be a wearout failure mode, it may be that one should stratify the data into relatively homogeneous subgroups and do prediction based on the stratified data.

This paper is motivated by reliability prediction problems that arise in engineering applications. In our particular application, an energy company wanted to predict future failures for a population of high-voltage transformers. The predictions are to be based on lifetime data collected up to a given date. The power transformer population consists of a mixture of two different designs, an old design and a new design. Both engineering knowledge and the data suggest that there is a difference between the old-design and the new-design transformers because the old transformers were often over-engineered. The life distribution estimate from the pooled data suggests a nearly constant hazard function, a result engineers who work with these transformers know to be wrong. However, the estimates from stratified data suggest different increasing hazard functions for the different designs. Extrapolation to predict future failures from a constant hazard function model would, give incorrect answers.

1.2   Related Literature and Contributions of This Work

Proschan (1963) showed that pooled data on the times between failures of an air-conditioning system from a fleet of airplanes would indicate that the distribution of times between failures has a decreasing hazard function and gave some theoretical explanation of this phenomena.

Gurland and Sethuraman (1994) gave two examples of a mixture of two distributions with rapidly increasing hazard functions that also behave as a distribution with a decreasing hazard function if the data are pooled. Block and Joe (1997) studied the tail behavior of the hazard function for mixtures. Block, Savits, and Wondmagegnehu (2003) studied the shape and the overall behavior of the hazard function of a mixture of two distributions with linearly increasing hazard functions. In this paper, we use results from White (1982) who gave general asymptotic theory for the properties of a maximum likelihood estimator under a misspecified model.

We study the asymptotic properties of the maximum likelihood (ML) estimator under an incorrectly specified model to be used for prediction. We compare the asymptotic mean square error (AMSE) of predictions for the cumulative number of failing at a future time based both on the pooled-data model (inappropriate) and stratified-data model (appropriate). Results show that the prediction based on the pooled-data model can be seriously biased. We present an analysis of the power transformer data as an illustration.

1.3 Overview

The remainder of this paper is organized as follows. Section 2 introduces the lifetime model, the data, and the ML estimation for the pooled and stratified data, and the asymptotic properties of the ML estimator. Section 3 gives details on predicting the cumulative number of failing in a future time interval, based on a pooled-data and a stratified-data model. Section 3 also compares the asymptotic mean square error for the predictions from these two models. Section 4 gives an application to the power transformer data. Section 5 contains some conclusions and provides some suggestions for practical applications.

2 Lifetime Model, Data and ML Estimation

2.1 The Lifetime Model

The Weibull and the lognormal distributions are the most commonly used distributions to describe lifetime in reliability applications. These distributions are members of the log-location-scale family. Let T be a random variable with a distribution from the log-location-scale family. The cumulative distribution function (cdf) of T can be written as

$$F(t; \boldsymbol{\theta}) = \Phi \left[\frac{\log(t) - \mu}{\sigma} \right]$$

where $\Phi(\cdot)$ is the standard cdf for the location-scale family of distributions (location 0 and scale 1) and $\boldsymbol{\theta} = (\mu, \sigma)'$. The corresponding probability density function (pdf) is given by

$$f(t; \boldsymbol{\theta}) = \frac{1}{\sigma t} \phi \left[\frac{\log(t) - \mu}{\sigma} \right]$$

where $\phi(\cdot)$ is the standard pdf for the location-scale family of distributions. For example, the cdf and pdf of the Weibull random variable T are

$$F(t; \boldsymbol{\theta}) = \Phi_{\text{sev}} \left[\frac{\log(t) - \mu}{\sigma} \right] \text{ and } f(t; \boldsymbol{\theta}) = \frac{1}{\sigma t} \phi_{\text{sev}} \left[\frac{\log(t) - \mu}{\sigma} \right]$$

where $\Phi_{\text{sev}}(z) = 1 - \exp[-\exp(z)]$ and $\phi_{\text{sev}}(z) = \exp[z - \exp(z)]$ are the standard (i.e., $\mu = 0, \sigma = 1$) smallest extreme value cdf and pdf, respectively. The cdf and pdf of the Weibull random variable T can also be expressed as

$$F(t; \eta, \beta) = 1 - \exp \left[- \left(\frac{t}{\eta} \right)^{\beta} \right] \text{ and } f(t; \eta, \beta) = \left(\frac{\beta}{\eta} \right) \left(\frac{t}{\eta} \right)^{\beta - 1} \exp \left[- \left(\frac{t}{\eta} \right)^{\beta} \right]$$

where $\eta = \exp(\mu)$ is the scale parameter and $\beta = 1/\sigma$ is the shape parameter. The Weibull shape parameter β determines the monotonicity of the hazard function of the distribution. For more information about the log-location-scale family of distributions and applications in reliability, see Meeker and Escobar (1998, Chapter 4).

2.2 Data

Right censored data often arise in reliability applications because there are unfailed units at the time the data are analyzed. Type I censoring (time censoring) is one very common form

of censoring in lifetime data. Such data arise when interest is on a group of units that are put into service all at one time. Type I censoring is most common in life testing but also arises in field data when interest centers on a cohort of units that were manufactured over a short period of time and that suffered some manufacturing problem such as a bad batch of raw material. For notational simplicity, in our analytical development, we assume there is only a single censoring time, denoted by t_c. Usually with field data, however, staggered entry is involved. This leads to multiple time censoring. For example, the transformer data considered in Section 4 were multiply censored. The development of analytical results for multiply censored data would be similar to the single censoring situation treated here, but would require detailed specification of a model for the entry pattern and additional notation.

Denote the censored data by $(t_i, \delta_i), i = 1, 2, \cdots, n$. Here $\delta_i = \mathbf{1}_{(t_i \leq t_c)}$ is a censoring indicator. That is $\delta_i = 1$ for a failure and $\delta_i = 0$ for a censored observation. In the stratified-data model (true model), we assume the failure times T_i are independently distributed with pdf $f_1(x; \boldsymbol{\theta}_1)$, for $i = 1, 2, \cdots, n_1$ and $f_2(x; \boldsymbol{\theta}_2)$, for $i = n_1 + 1, n_1 + 2, \cdots, n = n_1 + n_2$. Further, assume that the proportion of units from sub-population 1 is $\lambda_{n_1,n} = n_1/n$ and that $\lambda_{n_1,n} \to \lambda \in (0,1)$ as $n \to \infty$. Here, $\boldsymbol{\theta}_1 = (\mu_1, \sigma_1)'$ and $\boldsymbol{\theta}_2 = (\mu_2, \sigma_2)'$ are parameters for the two sub-populations. Units can be identified as belonging to sub-population i. Suppose that sub-population 1 corresponds to the old-design and sub-population 2 corresponds to the new-design in the transformer data setting. Thus, the overall population is a mixture. If the pooled-data model is used, the (incorrect) assumption is that the failure times T_i are independent and identically-distributed with pdf $f(x; \boldsymbol{\theta}), i = 1, 2, \cdots, n$. That is, when analyzing the pooled data, we incorrectly assume that the lifetime of products using both the old design and the new design are from the same distribution with parameter $\boldsymbol{\theta} = (\mu, \sigma)$.

2.3   ML Estimation for the Stratified Data

The log-likelihood function for the data under the stratified-data model is $l_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} l_{ni}$ where $l_{ni} = \delta_i \log[f_1(t_i; \boldsymbol{\theta}_1)] + (1 - \delta_i) \log[1 - F_1(t_c; \boldsymbol{\theta}_1)]$ for $i = 1, 2, \cdots, n_1$ and $l_{ni} = \delta_i \log[f_2(t_i; \boldsymbol{\theta}_2)] + (1 - \delta_i) \log[1 - F_2(t_c; \boldsymbol{\theta}_2)]$ for $i = n_1 + 1, n_1 + 2, \cdots, n$. We call this the stratified-data model.

The ML estimator $\left(\widehat{\boldsymbol{\theta}}'_1, \widehat{\boldsymbol{\theta}}'_2\right)'$ maximizes $l_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. The Fisher information matrices for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are $I_{1n}(\boldsymbol{\theta}_1) = \mathrm{E}\left[-\partial^2 l_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}'_1\right]$, and $I_{2n}(\boldsymbol{\theta}_2) = \mathrm{E}\left[-\partial^2 l_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}'_2\right]$, respectively. In this paper, the expectation is always taken with respect to the true model (stratified-data model). Let

$$I_1(\boldsymbol{\theta}_1) = \lambda \int_0^\infty \left[-\frac{\partial^2 l_{n1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}'_1}\right] f_1(t; \boldsymbol{\theta}_1)dt,$$

$$I_2(\boldsymbol{\theta}_2) = (1-\lambda) \int_0^\infty \left[-\frac{\partial^2 l_{n,n_1+1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}'_2}\right] f_2(t; \boldsymbol{\theta}_2)dt.$$

Note that $I_1(\boldsymbol{\theta}_1), I_2(\boldsymbol{\theta}_2)$ are the limiting values of their finite sample average quantities (i.e., $I_1(\boldsymbol{\theta}_1) = \lim_{n\to\infty} I_{1n}(\boldsymbol{\theta}_1)/n$ and $I_2(\boldsymbol{\theta}_2) = \lim_{n\to\infty} I_{2n}(\boldsymbol{\theta}_2)/n$). As $n \to \infty$,

$$\sqrt{n}\left[\begin{pmatrix}\widehat{\boldsymbol{\theta}}_1 \\ \widehat{\boldsymbol{\theta}}_2\end{pmatrix} - \begin{pmatrix}\boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2\end{pmatrix}\right] \xrightarrow{d} N\left[\mathbf{0}, \begin{pmatrix}I_1^{-1}(\boldsymbol{\theta}_1) & \mathbf{0} \\ \mathbf{0} & I_2^{-1}(\boldsymbol{\theta}_2)\end{pmatrix}\right]. \tag{1}$$

2.4 ML Estimation for the Pooled Data

The log-likelihood function for *pooled* data, fitting the (incorrect) single distribution, is $l_n(\boldsymbol{\theta}) = \sum_{i=1}^n l_{ni}(\boldsymbol{\theta})$ where $l_{ni}(\boldsymbol{\theta}) = \delta_i \log[f(t_i; \boldsymbol{\theta})] + (1 - \delta_i) \log[1 - F(t_c; \boldsymbol{\theta})]$. The ML estimator $\widehat{\boldsymbol{\theta}}$ maximizes $l_n(\boldsymbol{\theta})$. White (1982) calls $\widehat{\boldsymbol{\theta}}$ a quasi-maximum likelihood (QML) estimator. Define matrices $A_n(\boldsymbol{\theta}) = \mathrm{E}[\partial^2 l_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']$ and $B_n(\boldsymbol{\theta}) = \mathrm{E}\left[\sum_{i=1}^n \partial l_{ni}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} \cdot \partial l_{ni}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'\right]$. Note that $A_n(\boldsymbol{\theta})$ is the expectation of the loglikelihood curvature (Hessian) matrix. Let

$$A(\boldsymbol{\theta}) = \int_0^\infty \left[\frac{\partial^2 l_{n1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] g(t; \lambda, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)dt, \text{ and } B(\boldsymbol{\theta}) = \int_0^\infty \left[\frac{\partial l_{n1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \frac{\partial l_{n1}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right] g(t; \lambda, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)dt$$

where $g(t; \lambda, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \lambda f_1(t; \boldsymbol{\theta}_1) + (1 - \lambda)f_2(t; \boldsymbol{\theta}_2)$. Note that $A(\boldsymbol{\theta}) = \lim_{n\to\infty} A_n(\boldsymbol{\theta})/n$ and $B(\boldsymbol{\theta}) = \lim_{n\to\infty} B_n(\boldsymbol{\theta})/n$. In order to study the asymptotic behavior of the QML estimator $\widehat{\boldsymbol{\theta}}$, we need the expected score function, $U_n(\boldsymbol{\theta}) = \mathrm{E}\left[\partial l_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right]$. Let $U(\boldsymbol{\theta}) = \int_0^\infty [\partial l_{n1}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}]$ $g(t; \lambda, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)dt$. Note that $U(\boldsymbol{\theta}) = \lim_{n\to\infty} U_n(\boldsymbol{\theta})/n$. Let $\boldsymbol{\theta}_* = (\mu_*, \sigma_*)'$ be the root of the equation $U(\boldsymbol{\theta}) = 0$. We call $\boldsymbol{\theta}_*$ the *wrong-model parameter*. Note that $\boldsymbol{\theta}_*$ depends on the censoring time t_c. In particular, $\lambda = 1$ and $\lambda = 0$ lead into $\boldsymbol{\theta}_* = \boldsymbol{\theta}_i, i = 1, 2$, respectively. The $\boldsymbol{\theta}_*$ obtained here, for the log-location-scale distribution, is the same as the parameter defined

6

λ	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	t_c	$\boldsymbol{\theta}_*$
0.55	$(5, 0.65)'$	$(3, 0.15)'$	70	$(4.2, 0.89)'$
0.55	$(5, 0.65)'$	$(3, 0.15)'$	50	$(4.0, 0.73)'$
0.55	$(5, 0.65)'$	$(3, 0.15)'$	30	$(3.5, 0.41)'$

Table 1: Values of the wrong-model parameters $\boldsymbol{\theta}_*$ for different censoring times.

in White (1982) which is obtained by minimizing the Kullback-Leibler information criterion (Kullback and Leibler 1952).

Because an explicit form of $\boldsymbol{\theta}_*$ usually is not available, $\boldsymbol{\theta}_*$ must be computed numerically. Table 1 gives values of $\boldsymbol{\theta}_*$ under the Weibull distribution for example values of the parameter $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \lambda$, and several values of t_c. The values of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \lambda$ used here are close to those in the numerical example in Section 4. Figure 1 gives plots for the hazard functions of two different sub-populations, and the mixture of these two sub-populations. The hazard function for the mixture is $\{\lambda[f_1(t; \boldsymbol{\theta}_1)] + (1 - \lambda)[f_2(t; \boldsymbol{\theta}_2)]\}/\{1 - \int_0^t \lambda[f_1(s; \boldsymbol{\theta}_1)] + (1 - \lambda)[f_2(s; \boldsymbol{\theta}_2)]ds\}$. The true hazard function for the mixture is increasing rapidly when t is less 20, decreases for some time, and then returns to the hazard function of the "strong" component ($\mu_1 = 5$, $\sigma_1 = 0.65$). This is not surprising because units in the "weak" mixture component ($\mu_2 = 3$, $\sigma_2 = 0.15$) fail rapidly and dominate in the hazard function when t is less 20. The "strong" component dominates the hazard function after most of the weak units have failed. The hazard functions corresponding to the wrong-model parameters are also plotted in Figure 1. The plot shows that the pooled data hazard function is seriously incorrect. For example, with $t_c = 70$ we have $\boldsymbol{\theta}_* = (4.2, 0.89)'$ and $\sigma_* = 0.89$, or equivalently $\beta_* = 1/\sigma_* = 1.12$, which suggests a nearly constant hazard function for the Weibull distribution.

Result 1 *Using the general results in White (1982), the following results hold for the prediction model in this paper.*

1. $\widehat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_*$.

2. $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \xrightarrow{d} N[\mathbf{0}, V(\boldsymbol{\theta}_*)]$ where $V(\boldsymbol{\theta}_*) = A^{-1}(\boldsymbol{\theta}_*)B(\boldsymbol{\theta}_*)A^{-1}(\boldsymbol{\theta}_*)$.
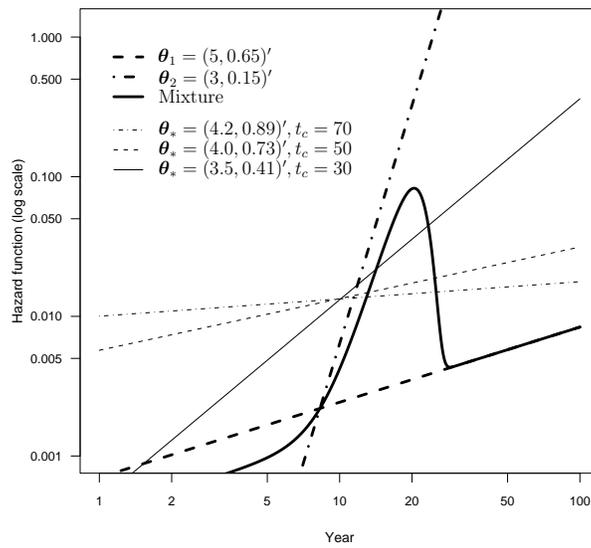
7

Figure 1: Comparison of hazard functions for the two sub-populations, the mixture of these two sub-populations, and the wrong model for three different values of the wrong-model parameters $\boldsymbol{\theta}_*$.

Result 1 shows that the QML estimator $\widehat{\boldsymbol{\theta}}$ under the wrong-model is asymptotically normally distributed and consistent in the sense that it converges to $\boldsymbol{\theta}_*$ almost surely (a.s.), the wrong-model parameter. An estimator of $V(\boldsymbol{\theta}_*)$ is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = \widehat{A}^{-1}(\widehat{\boldsymbol{\theta}})\widehat{B}(\widehat{\boldsymbol{\theta}})\widehat{A}^{-1}(\widehat{\boldsymbol{\theta}}) \tag{2}$$

where $\widehat{A}(\widehat{\boldsymbol{\theta}}) = n^{-1}[\partial^2 l_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$ and $\widehat{B}(\widehat{\boldsymbol{\theta}}) = n^{-1}\left[\sum_{i=1}^{n}\partial l_{ni}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} \cdot \partial l_{ni}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'\right]|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$. Note that $\widehat{V}(\widehat{\boldsymbol{\theta}})$ is the so-called "sandwich" estimator and is a robust estimator of $n\mathrm{Var}(\widehat{\boldsymbol{\theta}})$ under model misspecification (e.g., Kalbfleisch and Prentice 2002, page 210). If one is concerned that an incorrect model might have been fit, then the robust variance estimator should be used because the ordinary variance estimator, $n[\partial^2 l_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}']^{-1}$, is no longer consistent for estimating $n\mathrm{Var}(\widehat{\boldsymbol{\theta}})$ due to **Result 1**.

3 Prediction for the Mixture Population

In this section, we consider the problem described in Section 1.1 of predicting the cumulative number of future failures, denoted by K_t, at a future time t $(> t_c)$ for those units were survived until time t_c. Our approach is similar to that used in Escobar and Meeker (1999, Section 4). K_t can be interpreted as the number of field returns in the warranty data context. The expected number of failures is

$$\mathrm{E}\,(K_t) = N_n \times \rho(t).$$

Here, $N_n = \sum_{i=1}^{n}(1-\delta_i)$ is the number of units at-risk at time t_c and $\rho(t)$ is the probability of failing between time t_c and time t conditional on surviving to time t_c. An estimator of $\mathrm{E}\,(K_t)$ is

$$\widehat{\mathrm{E}\,(K_t)} = N_n \times \widehat{\rho}(t)$$

which is also a prediction for K_t. Because the size of the risk set N_n can be treated as a constant when the dataset is given at time t_c, we focus on estimating $\rho(t)$ for the N_n at-risk units in the subsequent development in this paper. Also, we assume that $N_n > 0$ in the observed data. Otherwise, there is no need to do prediction because all units have failed at time t_c.

9

3.1 The Stratified-Data Model

For the stratified-data model, $\rho_1(t; \boldsymbol{\theta}_1) = [F_1(t; \boldsymbol{\theta}_1) - F_1(t_c; \boldsymbol{\theta}_1)]/[1 - F_1(t_c; \boldsymbol{\theta}_1)]$, and $\rho_2(t; \boldsymbol{\theta}_2) = [F_2(t; \boldsymbol{\theta}_2) - F_2(t_c; \boldsymbol{\theta}_2)]/[1 - F_2(t_c; \boldsymbol{\theta}_2)$, $t > t_c$ are the distributions of remaining life (the probability of failing between t_c and t given that survived at t_c) for the two sub-populations. Thus, the cumulative fraction of the N_n remaining at-risk units failing between t_c and a future time t, based on the mixture population, is $\rho(t; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \nu_{n_1,n}\rho_1(t; \boldsymbol{\theta}_1) + (1 - \nu_{n_1,n})\rho_2(t; \boldsymbol{\theta}_2)$ where $\nu_{n_1,n} = \sum_{i=1}^{n_1}(1 - \delta_i)/\sum_{i=1}^{n}(1 - \delta_i)$. The ML estimator of $\rho(t; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is $\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)$. The large-sample approximate variance of $\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)$ is

$$\text{AVar}[\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)] = \nu_{n_1,n}^2 \text{AVar}[\rho_1(t; \widehat{\boldsymbol{\theta}}_1)] + (1 - \nu_{n_1,n})^2 \text{AVar}[\rho_2(t; \widehat{\boldsymbol{\theta}}_2)] \tag{3}$$

where $\text{AVar}[\rho_1(t; \widehat{\boldsymbol{\theta}}_1)] = [\partial\rho_1(t; \boldsymbol{\theta}_1)/\partial\boldsymbol{\theta}_1]' [nI_1(\boldsymbol{\theta}_1)]^{-1} [\partial\rho_1(t; \boldsymbol{\theta}_1)/\partial\boldsymbol{\theta}_1]$ and $\text{AVar}[\rho_2(t; \widehat{\boldsymbol{\theta}}_2)] = [\partial\rho_2(t; \boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_2]' [nI_2(\boldsymbol{\theta}_2)]^{-1} [\partial\rho_2(t; \boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_2]$. We obtain an estimate of $\text{AVar}[\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)]$ by evaluating (3) at $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2$. By the asymptotic properties of ML estimator (see, for example, Cox and Hinkley 1974, page 309-310), the estimator $\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)$ is asymptotically unbiased for $\rho(t; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Thus, $\text{ABias}^2[\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)] = 0$.

3.2 The Pooled-Data Model

Under the pooled-data model, the cumulative fraction of the N_n remaining at-risk units failing between t_c and a future time t based on the pooled population is $\rho(t; \boldsymbol{\theta}) = [F(t; \boldsymbol{\theta}) - F(t_c; \boldsymbol{\theta})]/[1 - F(t_c; \boldsymbol{\theta})]$, $t > t_c$. The ML estimator of this quantity is $\rho(t, \widehat{\boldsymbol{\theta}})$. The large-sample approximate variance of $\rho(t, \widehat{\boldsymbol{\theta}})$ is

$$\text{AVar}[\rho(t; \widehat{\boldsymbol{\theta}})] = \left[\frac{\partial\rho(t; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} \right]' \left[\frac{V(\boldsymbol{\theta}_*)}{n} \right] \left[\frac{\partial\rho(t; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} \right]. \tag{4}$$

We estimate $\text{AVar}[\rho(t; \widehat{\boldsymbol{\theta}})]$ by substituting $\widehat{\boldsymbol{\theta}}$ into (4). The square of the asymptotic bias for $\rho(t; \widehat{\boldsymbol{\theta}})$ is

$$\text{ABias}^2[\rho(t; \widehat{\boldsymbol{\theta}})] = \left| \rho(t; \boldsymbol{\theta}_*) - \rho(t; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \right|^2. \tag{5}$$

We estimate the square of the asymptotic bias for $\rho(t; \widehat{\boldsymbol{\theta}})$ by substituting $\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2$ into (5). That is $\widehat{\text{ABias}}^2[\rho(t; \widehat{\boldsymbol{\theta}})] = \left| \rho(t; \widehat{\boldsymbol{\theta}}) - \rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2) \right|^2$.

3.3 The Asymptotic Mean Square Error

Following the approach of Pascual (2006), we use the AMSE as a criterion for prediction under model misspecification. The AMSE of $\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)$ is

$$\text{AMSE}[\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)] = \text{AVar}[\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)]. \tag{6}$$

The AMSE of $\rho(t, \widehat{\boldsymbol{\theta}})$ is

$$\text{AMSE}[\rho(t, \widehat{\boldsymbol{\theta}})] = \text{AVar}[\rho(t, \widehat{\boldsymbol{\theta}})] + \text{ABias}^2[\rho(t, \widehat{\boldsymbol{\theta}})]. \tag{7}$$

Result 2 *As $n \to \infty$,*

1. $\text{AMSE}[\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)] \to 0$,

2. $\text{AMSE}[\rho(t, \widehat{\boldsymbol{\theta}})] \to \text{ABias}^2[\rho(t, \widehat{\boldsymbol{\theta}})]$.

The proof of **Result 2** is straightforward because $\text{AVar}[\rho(t; \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)]$ in (3) and $\text{AVar}[\rho(t; \widehat{\boldsymbol{\theta}})]$ in (4) go to 0 as $n \to \infty$. **Result 2** shows that the asymptotic bias dominates the AMSE of the pooled-data model when n is large enough.

Figure 2 gives comparisons of the AMSE in (6) and (7) for the parameters $\lambda = 0.55$, $\boldsymbol{\theta}_1 = (5, 0.65)'$, $\boldsymbol{\theta}_2 = (3, 0.15)'$, $t_c = (70, 17)'$, $\boldsymbol{\theta}_* = (4.86, 0.85)'$, and $\nu_{n_1, n} = 0.55$. Here, we chose two values for the censoring time t_c so that the two sub-populations are censored approximately at the same proportion. Hence, the proportion of units from the two sub-populations in the risk-set are approximately the same. If these proportions were seriously unbalanced, it would make it difficult to see the effect of asymptotic variance from two mixture components in (3). The sample size n in Figures 2a and 2b is 10 and 100, respectively. When $n = 10$, the AMSE of the predictions for the stratified-data model are larger over some period of time than that for the pooled-data model. This is because the stratified-data model has more parameters that need to be estimated than the pooled-data model, increasing the variability in estimation. Note in Figure 2a, there are two modes in the plot of the AMSE for the stratified-data model. This is because the variance from the "weak" component is dominating the AMSE for small times and the variance from the "strong" component is dominating the AMSE for large times.
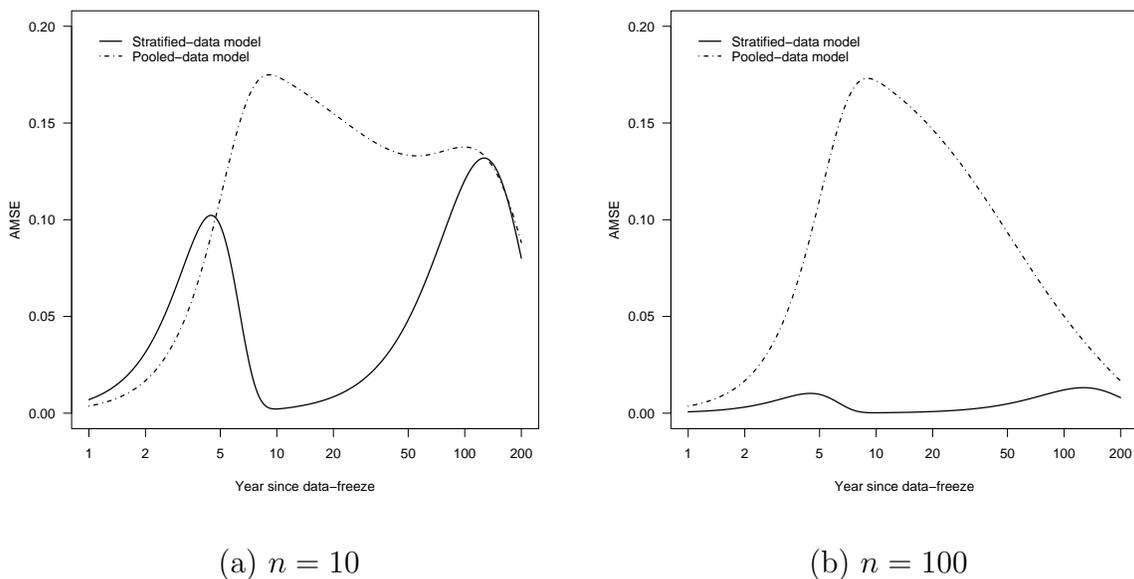
(a) $n = 10$ (b) $n = 100$

Figure 2: Comparison of the AMSE for the prediction based on the pooled-data and stratified-data models.

When $n = 100$, however, the AMSE for the stratified-data model is much smaller than that of the pooled-data model. The AMSE for the pooled-data model does not decrease much when the sample size is increased from 10 to 100. The AMSE of the stratified-data model is relatively small, which is consistent with **Result 2**.

4   The High-Voltage Transformer Data

4.1   Background

In this section, we use a particular dataset to illustrate the importance of stratification. As mentioned in Section 1.1, an energy company wanted to predict the cumulative number of failures between the data-freeze time and a future point in time denoted by t among those at-risk units, based on the lifetime data collected up to the data-freeze time. Our data set is based on 95 units of which 12 were failures. In particular, there are 45 units with the new design (6 failures) and 50 units with the old design (6 failures). Figure 3 gives an event plot

12
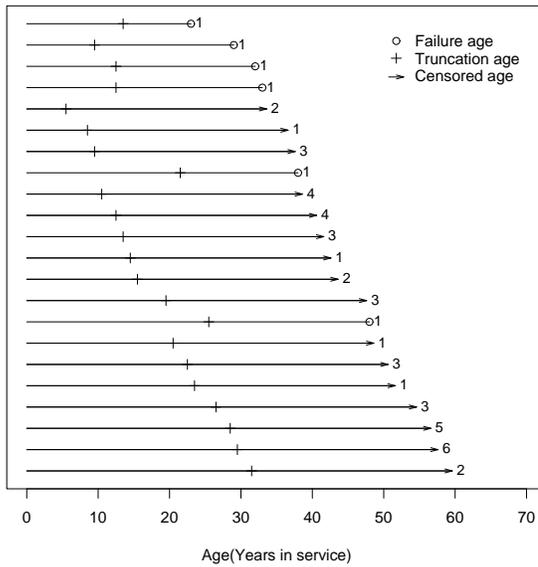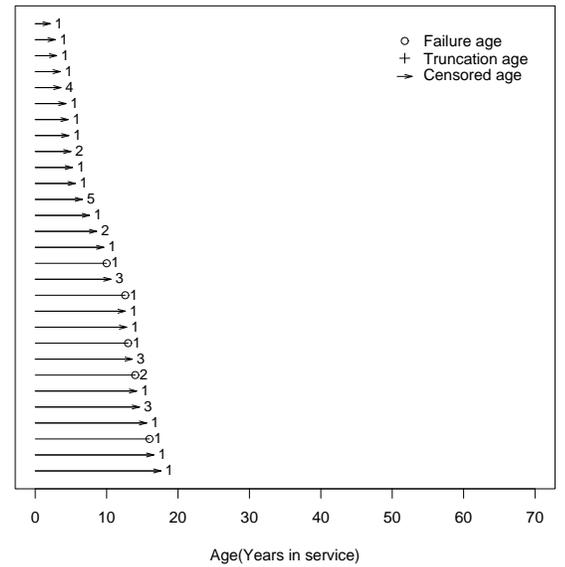
(a) Old design   (b) New design

Figure 3: Event plot of the power transformer data. The numbers to the right are the multiplicity of the corresponding events.

of the data.

One complication in the data is that records for transformers that were in service before January 1, 1980, but that did not survive until January 1, 1980, are not available. Thus, no information is available on transformers that failed before January 1, 1980. For this reason, transformers that were installed before January 1, 1980 and survived until January 1, 1980 should be considered as transformers sampled from truncated distributions. For transformers that were installed before January 1, 1980 and that survived at least until January 1, 1980, we know either the failure time or that the transformer is still in service and the corresponding service time. Thus, the power transformer data are left truncated and right censored. Note that the truncation points (for those units installed before January 1, 1980) and censoring time (for those units still in service) vary from transformer to transformer because of staggered entry of transformers into service.

13

4.2 The ML Estimates

From engineering knowledge, the lifetime distribution of transformers is expected to have an increasing hazard function due to insulation aging. There is a difference between old transformers and new transformers because old transformers were often over-engineered and for this reason, old transformers tend to have longer lifetimes. Thus, there is a mixture of two sub-populations: the old-design sub-population and the new-design sub-population. Figure 4 gives the probability plot of the ML estimate for the pooled data obtained by fitting a single Weibull distribution. Because all of the old design units are truncated, the plot of the points from the nonparametric plot of the data points do not align well with the parametric ML estimate. However, this does not indicate for lack of fit because the Kaplan-Meier estimator is inconsistent for such truncated data. We also tried the lognormal and other distributions and the results were similar. Figure 5 gives probability plots of the ML estimate for the stratified data. As with the pooled data, the probability plot for the old design data exhibits curvature due to the truncation. For the new-design data, there was no truncation and the Weibull distribution fits the data well.

Table 2 gives the ML estimates for the Weibull scale parameters ($\eta = \exp(\mu)$) and shape parameters ($\beta = 1/\sigma$) for both models. The standard errors using both the ordinary estimator and the "sandwich" estimator (the robust estimator) in (2) are also reported. The estimate ($\widehat{\beta} = 1.095$) from the pooled data incorrectly suggests a nearly constant hazard function for the overall population. The estimates ($\widehat{\beta}_1 = 1.499, \widehat{\beta}_2 = 7.063$) from the stratified data suggest different increasing hazard functions for the two different sub-populations.

4.3 Prediction

In this section, we give the prediction results for the power transformer data. Figure 6 gives the predictions for the fraction failing as a function of time for the $N = 83$ at-risk units using the methods described in Section 3. The predictions based on the incorrectly pooled-data model and stratified-data model are plotted together as a comparison. The prediction based
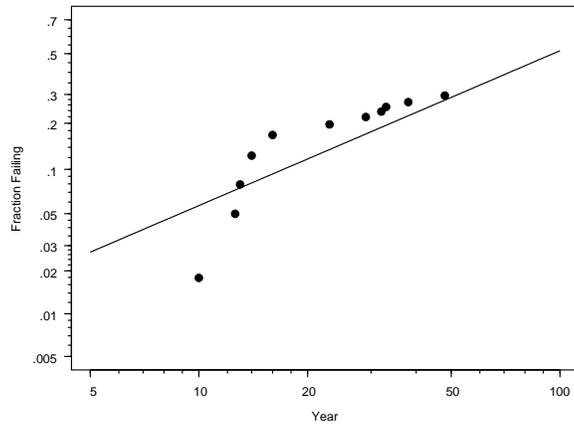
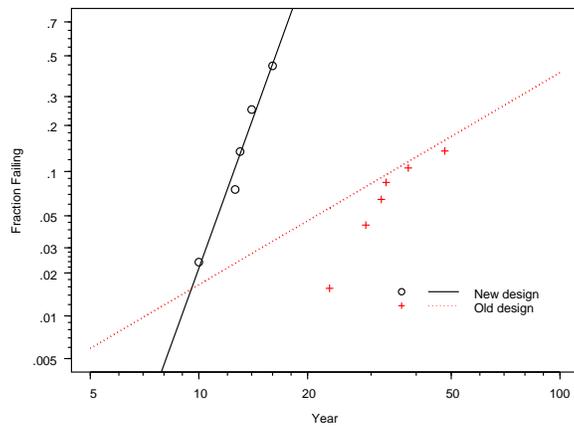Figure 4: Weibull probability plot and the ML estimate based on pooled-data model.

Figure 5: Weibull probability plot and the ML estimates based on the stratified-data model.

15

Models		MLE	SE	Robust SE	95% CI		95% CI (Robust)	
					Lower	Upper	Lower	Upper
Stratified (old)	$\widehat{\eta}_1$	153.270	106.752	77.054	39.138	600.229	57.218	410.566
	$\widehat{\beta}_1$	1.499	1.105	0.695	0.353	6.359	0.604	3.720
Stratified (new)	$\widehat{\eta}_2$	17.176	1.195	1.262	14.986	19.686	14.871	19.837
	$\widehat{\beta}_2$	7.063	2.161	1.817	3.877	12.867	4.266	11.695
Pooled	$\widehat{\eta}$	133.807	48.966	40.998	65.311	274.139	73.397	243.939
	$\widehat{\beta}$	1.095	0.293	0.149	0.648	1.850	0.838	1.429

Table 2: ML estimates of the scale ($\eta = \exp(\mu)$) and shape ($\beta = 1/\sigma$) parameters of the Weibull distribution and 95% confidence intervals (CIs) based on the pooled-data and stratified-data models.

the pooled-data model is much less than that of the stratified-data model. This is because the estimate from the pooled-data model incorrectly indicates a nearly constant hazard function which means, not recognizing the large number of early failures that can be expected from the weak new-design part of the population. This makes the prediction based on pooled-data model seriously biased. For the stratified-data model, the predicted fraction failing increases less rapidly after 20 years. This is because almost all of the units with the new design are expected to have failed by that time. Figure 7 gives a comparison between the estimated AMSE for the prediction based on the pooled-data and the stratified-data models. The estimated AMSE of the stratified-data model is much less than that of the pooled-data model for $t < 50$. The estimated AMSE of the stratified-data model is larger than that of the pooled-data model for $t > 100$, which is different from in Figure 2b where the AMSE of the stratified-data model is smaller for all t. This is because the estimated large-sample approximate variance for $\rho_1(t, \widehat{\boldsymbol{\theta}}_1)$ in this particular dataset is larger than the large-sample approximate variance for $\rho_1(t, \widehat{\boldsymbol{\theta}}_1)$ in Figure 2b.
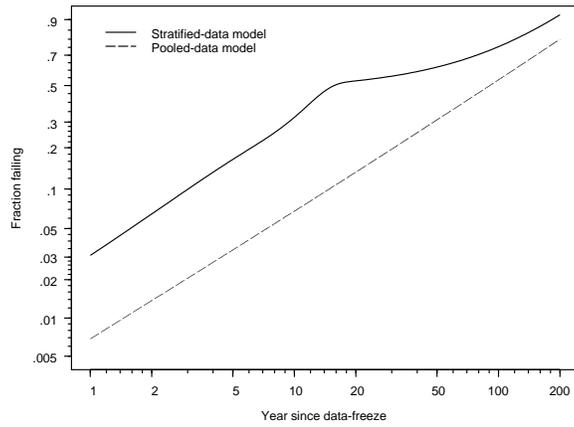
Figure 6: Comparison of estimates of fraction failing extrapolated to 200 years, based on pooled-data and stratified-data models.
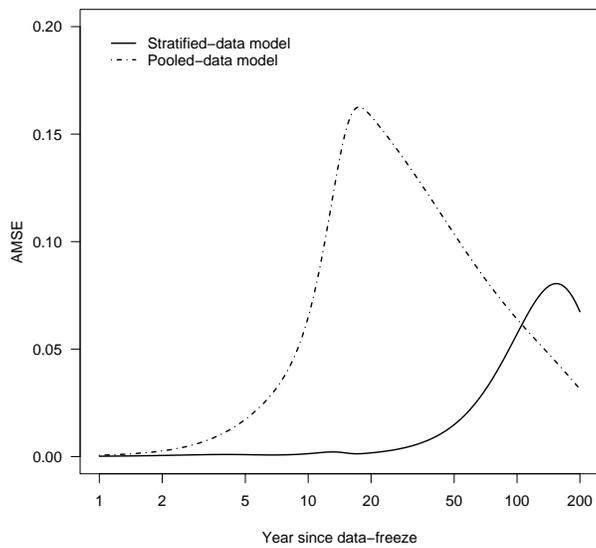

Figure 7: Comparison of the estimated AMSE for the prediction based on the pooled-data and stratified-data models.

5 Conclusions

In this paper, we have developed analytical results to show the importance of stratifying lifetime data into relatively homogenous subgroups, especially when extrapolation is involved. We used the prediction of future failures of the high-voltage transformers to illustrate the results. Stratification reduces prediction bias and provides sensible statistical results.

If preliminary analysis of pooled data suggests a decreasing or a constant hazard function when this is not consistent with the known increasing hazard failure mode; it is important to consider stratifying the data into relatively homogeneous subgroups. The stratification should take the knowledge of product failure mechanisms, explanatory variables, and data analysis into consideration. A statistical test of the significance of the difference between subgroups and the sensitivity analysis on the dividing rules can also be useful to guide in stratifying. However, stratifying of data into too many subgroups may increase the variance of the prediction when the number of units under observation is small. This is because stratification will increase the number of the parameters that need to be estimated from the data.

Acknowledgments

References

Block, H. and H. Joe (1997). Tail behavior of the failure rate functions of mixtures. *Lifetime Data Analysis 3*, 269–288.

Block, H. W., T. H. Savits, and E. T. Wondmagegnehu (2003). Mixtures of distributions with increasing linear failure rates. *Journal of Applied Probability 40*, 485–504.

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

Escobar, L. A. and W. Q. Meeker (1999). Statistical prediction based on censored life data. *Technometrics 41*(2), 113–124.

Gurland, J. and J. Sethuraman (1994). Reversal of increasing failure rates when pooling failure data. *Technometrics 36*(4), 416–418.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data* (2nd ed.). John Wiley and Sons Inc.

Kullback, S. and R. A. Leibler (1952). On information and sufficiency. *Annals of Mathematical Statistics 22*, 79–86.

Meeker, W. Q. and L. A. Escobar (1998). *Statistical Methods for Reliability Data.* John Wiley and Sons Inc.

Pascual, F. G. (2006). Accelerated life test plans robust to misspecification of the stress-life relationship. *Technometrics 48*(1), 14–25.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics 5*(3), 373–383.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.