# Chapter 2

# Principles of Maximum Likelihood Estimation and The Analysis of Censored Data

**January 22, 2003**



Developed as part of NSF/ILI grant DUE9751644.

## Objectives

This chapter explains

- Motivation for the use of parametric likelihood as a tool for data analysis and inference.
- Principles of likelihood and how likelihood is related to the probability of the observed data.
- Likelihood for parametric distributions.
- The use of likelihood confidence intervals for model parameters and other quantities of interest.
- Censoring mechanisms that restrict one's ability to observe actual response values.
- Likelihood for samples containing right and left censored observations.

## Overview

This chapter introduces some ideas and methods for using maximum likelihood (ML) to estimate model parameters. The basic ideas presented here are used throughout the rest of the book. Section 2.2 describes a number of more the commonly-used parametric probability distributions and their characteristics. Section 2.3 shows how to construct the likelihood (probability of the data) function and describes the basic ideas behind using this function to estimate a parameter. Sections 2.4 and 2.5 describe, respectively, methods for computing confidence intervals for parameters and functions of parameters. Section 2.6 describes and illustrates the use of likelihood methods for right censored data. It also describes the commonly used probability density-approximation for the likelihood function for "exact observations." Section 2.7 describes and illustrates the use of likelihood methods for left censored data.

Table 2.1: Binned $\alpha$-particle interarrival time data in units of 1/5,000 seconds.

| Time | | Interarrival Times Frequency of Occurrence | | | |
|---|---|---|---|---|---|
| Interval Endpoint | | All Times | Random Samples of Times | | |
| lower | upper | $n = 10220$ | $n = 2000$ | $n = 200$ | $n$=20 |
| 0 | 100 | 1609 | 292 | 41 | 3 |
| 100 | 300 | 2424 | 494 | 44 | 7 |
| 300 | 500 | 1770 | 332 | 24 | 4 |
| 500 | 700 | 1306 | 236 | 32 | 1 |
| 700 | 1000 | 1213 | 261 | 29 | 3 |
| 1000 | 2000 | 1528 | 308 | 21 | 2 |
| 2000 | 4000 | 354 | 73 | 9 | 0 |
| 4000 | $\infty$ | 16 | 4 | 0 | 0 |

## 2.1 Introduction

The method of Maximum Likelihood (ML) is perhaps the most versatile method for fitting statistical models to data. The method dates back to early work by Fisher (1925) and has been used as a method for constructing estimators ever since. In typical applications, the goal is to use a parametric statistical model to describe a set of data or a process that generated a set of data. The appeal of ML stems from the fact that it can be applied to a wide range of statistical models and kinds of data (e.g., continuous, discrete, categorical, censored, truncated, etc.), where other popular methods, like least squares, do not, in general, provide a satisfactory method of estimation. Indeed, when assuming an underlying normal (also known as Gaussian) distribution, the least squares estimates of regression coefficients are equivalent to ML estimates. The ML method is, however, much more general because it allows one to use other distributions as well as more general assumptions about the model and the form of the data, as will be illustrated later in this chapter as well as in subsequent chapters. Statistical theory (e.g., the references in the bibliographic notes) shows that, under standard regularity conditions, ML estimators are optimal in large samples.

Modern computing hardware and software have tremendously expanded the feasible areas of application for ML methods. There have been a few reported problems with some applications of the ML method. These are typically due to the use of inappropriate approximations or inadequate numerical methods used in calculations. This chapter illustrates the method of ML on a range of practical problems involving censored data.

**Example 2.1 Time between $\alpha$-particle emissions of americium-241.**

Berkson (1966) investigates the randomness of $\alpha$-particle emissions of americium-241 (which has a half-life of about 458 years). Physical theory suggests that, over a short period of time, particles are generated randomly at a constant rate per unit time. The expected number of particles detected in an interval of time would be proportional to the americium-241 decay rate, size of the sample, and the the efficiency of the counter.

The original data consisted of 10,220 observed interarrival times of $\alpha$ particles (time unit equal to 1/5,000 second). The observed interarrival times were put into intervals (or bins) running from 0 to 4,000 time units with interval lengths ranging from 25 to 100 time units, with one additional interval for observed times exceeding 4,000 time units. To simplify our analysis, examples in this chapter use a smaller number of larger bins; reducing the number of bins in this way will not seriously affect the precision of ML estimates. These data are shown in Table 2.1.

To illustrate the effects of sample size on ML estimation, simple random samples (i.e., each interval-censored interarrival time had an equal probability of being selected) of sizes $n = 2000$, 200, and 20 were drawn with replacement from these interarrival times. The following examples compare the results that one obtains with these different sample sizes. When focusing on just one sample, the sample of size $n = 200$ interarrival times is used. Figure 2.1 is a histogram of this sample. ∎

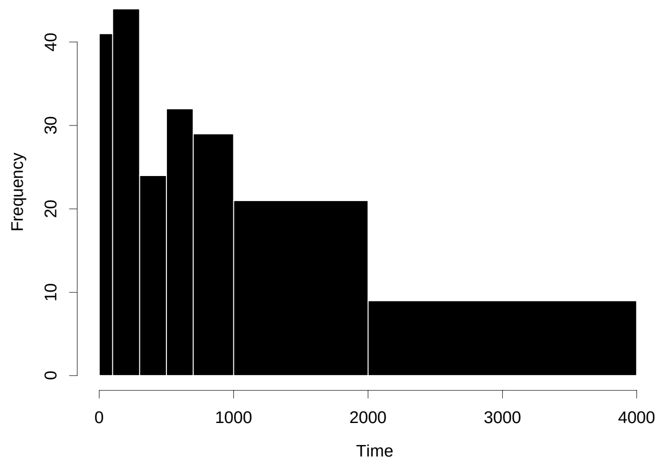


Figure 2.1: Histogram of the $n = 200$ sample of $\alpha$-particle interarrival time data.

## 2.2 Basic Ideas of Parametric Modeling

### 2.2.1 Modeling variability with parametric distributions

The natural model for a continuous random variable, say $T$, is the cumulative distribution function (cdf) having the form

$$\Pr(T \leq t) = F(t; \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} > 0$ is, in general, a vector of parameters and the range of $T$ is specified. This function can also be interpreted as giving the fraction of $T$ values that will be below a given value of $t$. This section describes some commonly used distributions.

**The exponential distribution.** One simple example that we will use later in this chapter is the exponential distribution for which

$$\Pr(T \leq t) = F(t; \theta) = 1 - \exp\left(-\frac{t}{\theta}\right), \quad t > 0, \tag{2.1}$$

where $\theta$ is the single parameter of the distribution (equal to the first moment or mean, in this example). The exponential probability density function is

$$f(t; \theta) = \frac{dF(t; \theta)}{dt} = \frac{1}{\theta} \exp\left(-\frac{t}{\theta}\right). \tag{2.2}$$

Figure 2.2 shows the exponential distribution pdf and cdf.

**The lognormal distribution.** The lognormal distribution is widely used in many areas of application, including engineering, medicine, and finance. The lognormal cdf is

$$F(t; \mu, \sigma) = \Phi_{\text{nor}}\left[\frac{\log(t) - \mu}{\sigma}\right], \quad t > 0, \tag{2.3}$$

and the lognormal pdf is

$$f(t; \mu, \sigma) = \frac{dF(t; \theta)}{dt} = \frac{1}{\sigma t}\phi_{\text{nor}}\left[\frac{\log(t) - \mu}{\sigma}\right], \quad t > 0. \tag{2.4}$$

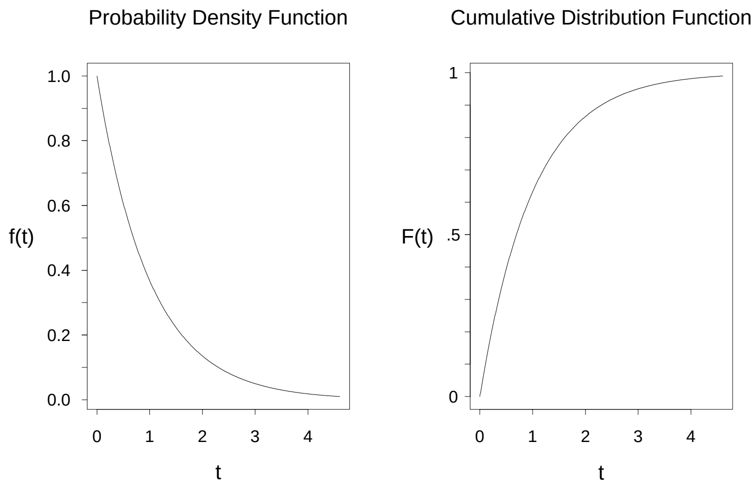


Figure 2.2: Exponential distribution pdf and cdf.

where $\phi_{\text{nor}}(z) = (1/\sqrt{2\pi})\exp(-z^2/2)$ and $\Phi_{\text{nor}}(z) = \int_{-\infty}^{z} \phi_{\text{nor}}(w)dw$ are, respectively, the pdf and cdf for a standardized normal ($\mu = 0, \sigma = 1$).

The natural logarithms of a lognormal random variable follows the well-known normal distribution with mean and standard deviation $\mu$ and $\sigma$, respectively. This can be seen from equation (2.3). Thus $\mu$ and $\sigma$ can be interpreted, respectively, as the mean and standard deviation of the logarithms, respectively. Following from the central limit theorem, the lognormal distribution can be motivated as the distribution of the product of a large number of similarly distributed positive quantities. Figure 2.3 shows the lognormal distribution pdf and cdf.

**The Weibull distribution.** The Weibull distribution is, in many ways (including the general shape of the cdf and pdf), similar the lognormal distribution. It is also commonly used in different areas of science and engineering. Its cdf is commonly written as

$$F(t; \alpha, \beta) = 1 - \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right], \quad t > 0 \tag{2.5}$$

where $\beta$ is a shape parameter and $\alpha$ is a scale parameter. The Weibull cdf also can be written as

$$F(t; \mu, \sigma) = \Phi_{\text{sev}}\left[\frac{\log(t) - \mu}{\sigma}\right], \tag{2.6}$$

and the Weibull pdf is

$$f(t; \mu, \sigma) = \frac{1}{\sigma t}\phi_{\text{sev}}\left[\frac{\log(t) - \mu}{\sigma}\right], \quad t > 0. \tag{2.7}$$

where $\phi_{\text{sev}}(z) = \exp[z - \exp(z)]$ and $\Phi_{\text{sev}}(z) = \int_{-\infty}^{z} \phi_{\text{sev}}(w)dw = 1 - \exp[-\exp(z)]$ are, respectively, the pdf and cdf for a standardized smallest extreme value distribution ($\mu = 0, \sigma = 1$). This is so because natural logarithms of Weibull random variables follow the simpler smallest extreme value distribution with location and scale parameters $\mu = \log(\alpha)$ and $\sigma = 1/\beta$. The expressions in (2.6) and (2.7) is convenient because one can immediately switch to other distributions (e.g., the lognormal) by changing the definition of $\Phi$.

For more information on the technical characteristics and motivations for these and other similar distributions, see Chapters 4 and 5 of Meeker and Escobar (1998).

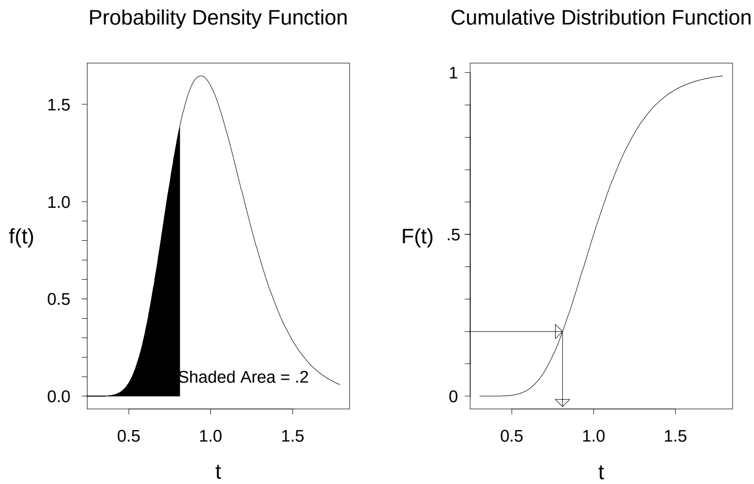


Figure 2.3: Lognormal distribution pdf and cdf.

The most commonly used parametric probability distributions have between 1 and 4 parameters, although there are also distributions with more than 4 parameters. More complicated models involving mixtures or other combinations of distributions and models that include explanatory variables could contain many more parameters.

### 2.2.2 Parameters and functions of parameters

This chapter shows how to estimate the model parameters $\boldsymbol{\theta}$ and important functions of $\boldsymbol{\theta}$.

In practical applications, interest often centers on quantities that are functions of $\boldsymbol{\theta}$ like

- The probability $p = \Pr(T \leq t) = F(t; \boldsymbol{\theta})$ for a specified $t$. For example, if $T$ is the tensile strength of a particular unit, then $p$ is the probability that the unit's tensile strength $T$ is less than $t$.

- The $p$ quantile of the distribution of $T$ is the value of $t_p$ such that $F(t_p; \boldsymbol{\theta}) = p$. We will express this quantity as $t_p = F^{-1}(p; \boldsymbol{\theta})$. For the tensile strength example, $t_p$ is the strength below which we will find $100p\%$ of the units in the product population. The arrows in Figure 2.3 illustrate the .2 quantile.

- The mean of $T$

$$\mathrm{E}(T) = \int_{-\infty}^{\infty} t f(t) dt$$

  which is sometimes, but not always, one of the model parameters.

Explicit formulas for these quantities are given in Chapters 4 and 5 of Meeker and Escobar (1998).

The choice of $\boldsymbol{\theta}$, a set of parameters (the values of which are often assumed to be unknown) to describe a particular distribution or more general model, is somewhat arbitrary and may depend on tradition, on physical interpretation, or on having parameters with desirable numerical properties. For example, the traditional definition for the parameters of a normal distribution are $\theta_1 = \mu$ and $\theta_2 = \sigma^2 > 0$, the mean and variance respectively. An alternative unrestricted parameterization would be $\theta_1 = \mu$ and $\theta_2 = \log(\sigma)$. Another parameterization, which has better numerical properties for some data sets is $\theta_1 = \mu - 2\sigma$ and $\theta_2 = \log(\sigma)$.

### 2.2.3 Data from and Underlying Continuous Model

Many scientific observations are modeled on a continuous scale. Because of inherent limitations in measurement precision, however, actual data are *always* discrete. Moreover, other limitations in our ability or in our measuring instrument's ability to observe can cause our data to be censored or truncated. The examples in this chapter will illustrate several different types of observations, give examples, show how to compute the "probability of the data" as a function of $\boldsymbol{\theta}$, and show how to use this function for statistical inference.

### 2.2.4 Interval-censored observations

"Interval-censored" data arise due to roundoff or binning, or when the response is the time to some event and the occurrence of an event is detected at inspection that are done at discrete time points. Subsequent sections describe and illustrate other kinds of censoring.

**Example 2.2 Physical model for $\alpha$-particle emissions of americium-241.** As described in Example 2.1, physical theory suggests that, over a short period of time, particles are generated randomly at a constant rate per unit time. This, and some other plausible physically based mathematical assumptions, implies that interarrival times will be independent and come from an exponential distribution with cdf given in (2.1). A mathematically equivalent model can be based on the homogeneous Poisson process that counts the number of emissions on the real-time line. This counting process has arrival rate or the intensity $\lambda = 1/\theta$ where, as in (2.1), $\theta$ is the mean time between arrivals. ∎

### 2.2.5 Using data to help choose a distributional model

Statistical modeling, in practice, is an iterative procedure of fitting proposed models in search of a model that provides an adequate description of the population or process of interest, without being unnecessarily complicated. Application of ML methods generally starts with a set of data and a tentative statistical model for the data. The tentative model is often suggested by the initial graphical analysis, physical theory, previous experience with similar data, or other expert knowledge. In order to describe the variability in a response without any explanatory variables it is often suggested that one start with a histogram and/or a probability plot of the data.

**Example 2.3 Graphical analyses for the $\alpha$-particle data.** Figure 2.1 gives a simple histogram and Figure 2.4 shows an exponential probability plot of the $\alpha$-particle data sample with $n = 200$. The points on the probability plot show the observed fraction of observations below $t$ as a function of $t$. The nonlinear probability axis on an exponential probability plot is chosen such that the exponential distribution cdf given in (2.1) plots as a straight line as a function of $t$. Thus the approximate linearity of the points on the plot indicates that the exponential distribution provides a good fit to these data. The simultaneous nonparametric confidence bands (described in Chapter 3 of Meeker and Escobar 1998) express statistical uncertainty—an indication of variability that one would see in such an estimate if the experiment were to be repeated over and over. ∎

## 2.3 Parametric Likelihood

### 2.3.1 Probability of the data

The likelihood function can be viewed as the *probability of the observed data*, written as a function of the model's parameter(s). The exponential distribution in (2.1) has only one parameter.

For a set of $n$ independent observations, the likelihood function can be written as the following joint probability

$$L(\theta) = \mathcal{C} \prod_{i=1}^{n} L_i(\theta; \text{data}_i) \tag{2.8}$$

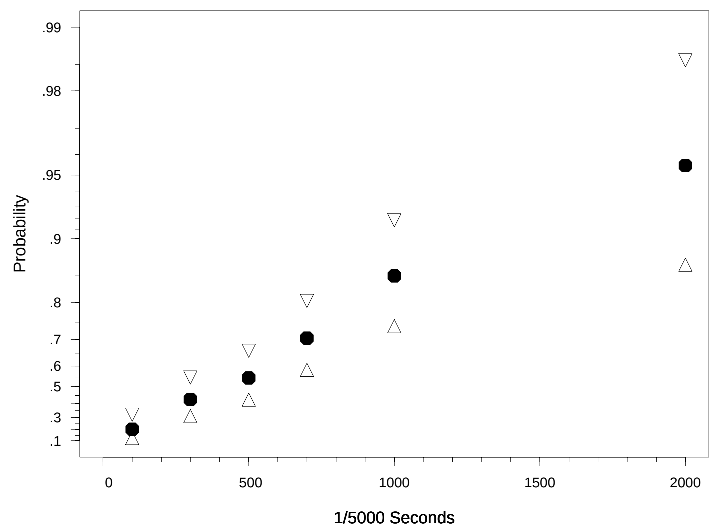


Figure 2.4: Exponential probability plot of the $n = 200$ sample of $\alpha$-particle interarrival time data. The plot also shows simultaneous nonparametric approximate 95% confidence bands.

where $\text{data}_i$ is used to denote the information that is available from observation $i$ and $L_i(\theta; \text{data}_i)$ is the probability of observation $i$. The quantity $\mathcal{C}$ in (2.8) is a constant term that depends on the underlying sampling scheme, but that does not depend on the data or on $\theta$. For computational purposes, let $\mathcal{C} = 1$. For example, if an event is known to have occurred between times $t_{i-1}$ and $t_i$, the probability of this event is

$$\begin{aligned} L_i(\theta) = L_i(\theta; \text{data}_i) &= \Pr(t_{i-1} < T \leq t_i) \\ &= \int_{t_{i-1}}^{t_i} f(t;\theta)\, dt = F(t_i;\theta) - F(t_{i-1};\theta). \end{aligned} \tag{2.9}$$

This probability is illustrated by the interval-censored observation in Figure 2.5. The probability of left and right censored observations will be described and illustrated in subsequent sections.

For a given set of data, $L(\theta)$ in (2.9) can be viewed as a function of the parameter $\theta$. The dependence of $L(\theta)$ on the data will be understood and, for the sake of simplicity, is usually suppressed in notation. The values of $\theta$ for which $L(\theta)$ is relatively large are more plausible than values of $\theta$ for which the probability of the data is relatively small. There may or may not be a unique value of $\theta$ that maximizes $L(\theta)$. Regions in the space of $\theta$ with relatively large $L(\theta)$ can be used to define confidence regions for $\theta$. One can also use ML to estimate *functions* of $\theta$. The rest of this chapter shows how to make these concepts operational for several different examples and statistical models. Subsequent chapters describe and use similar ideas, for other kinds of censoring.

### 2.3.2 Likelihood function and its maximum

Given a sample of $n$ independent observations, denoted generically by $\text{data}_i$, $i = 1, \ldots, n$, and a specified model, the total likelihood $L(\theta)$ for the sample is given by equation (2.8). For some purposes, it is convenient to use the log likelihood $\mathcal{L}_i(\theta) = \log[L_i(\theta)]$. For all practical problems $\mathcal{L}(\theta)$ will be representable in computer memory without special scaling (which is not so for $L(\theta)$ because of possible extreme exponent values), and some theory for ML is developed more naturally in terms of sums like

$$\mathcal{L}(\theta) = \log[L(\theta)] = \sum_{i=1}^{n} \mathcal{L}_i(\theta)$$

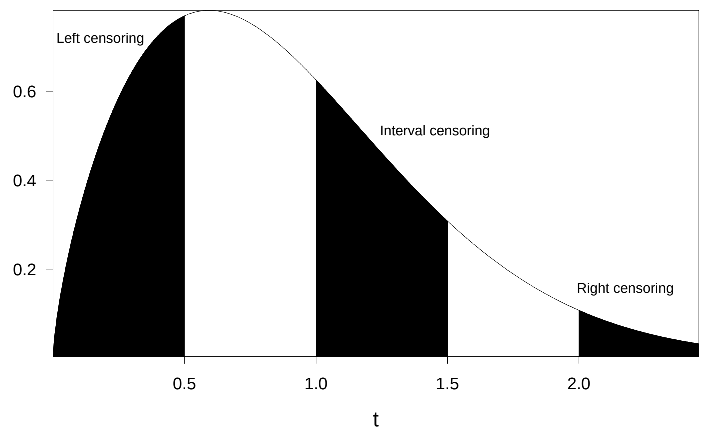


Figure 2.5: Illustration of probability (area under the curve) for censored observations.

rather than in terms of the products in equation (2.8). Note that the maximum of $\mathcal{L}(\theta)$, if one exists, occurs at the same value of $\theta$ as the maximum of $L(\theta)$.

**Example 2.4 Likelihood for the $\alpha$-particle data.** In this example, the unknown parameter $\theta$ is a scalar and this makes the analysis particularly simple and provides a useful first example to illustrate basic concepts. Substituting equation (2.1) into (2.9), and (2.9) into (2.8) gives the exponential distribution likelihood function (joint probability) for interval data (e.g., Table 2.1) as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} L_i(\theta) = \prod_{i=1}^{n} [F(t_i;\theta) - F(t_{i-1};\theta)] \qquad (2.10) \\ &= \prod_{j=1}^{8} [F(t_j;\theta) - F(t_{j-1};\theta)]^{d_j} = \prod_{j=1}^{8} \left[ \exp\left(-\frac{t_{j-1}}{\theta}\right) - \exp\left(-\frac{t_j}{\theta}\right) \right]^{d_j} \end{aligned}$$

where $d_j$ is the number of interarrival times in interval $j$. Notice that in the first line of (2.10), the product is over the $n$ observed times and in the second line, it is over the 8 bins into which the data have been grouped. ▮

The ML estimate of $\theta$ is found by maximizing $L(\theta)$. When there is a unique global maximum, $\widehat{\theta}$ denotes the value of $\theta$ that maximizes $L(\theta)$. In general, however, the maximum may not be unique. The function $L(\theta)$ may have multiple local maxima or can have flat spots along which $L(\theta)$ changes slowly, if at all. Such flat spots may or may not be at the maximum value of $L(\theta)$. The shape and magnitude of $L(\theta)$ relative to $L(\widehat{\theta})$ over all possible values of $\theta$ describe the information on $\theta$ that is contained in $\text{data}_i, i = 1, \ldots, n$. This suggests the importance of the relative likelihood function

$$R(\theta) = \frac{L(\theta)}{L(\widehat{\theta})}.$$

In particular, the relative likelihood function allows one to judge the probability of the data for values of $\theta$, *relative* to the probability at the ML estimate. For example, $R(\theta_0) = .1$ implies that the probability of the data is 10 times larger at $\widehat{\theta}$ than at $\theta_0$. The next section explains how to use $R(\theta)$ to compute confidence intervals for $\theta$.

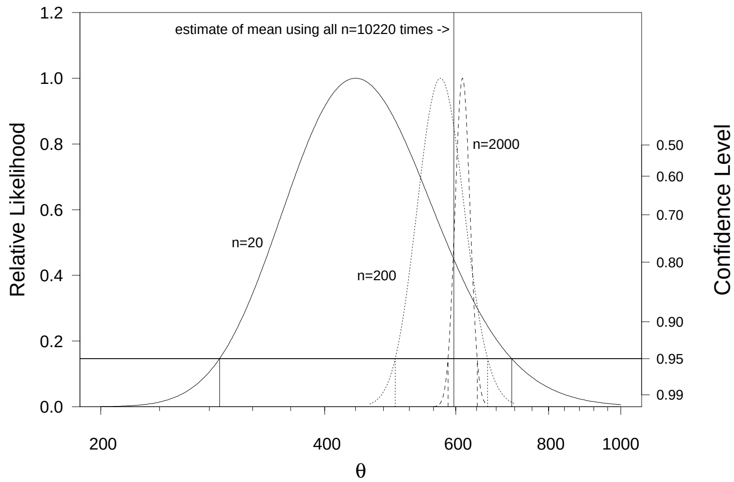


Figure 2.6: Relative likelihood functions $R(\theta) = L(\theta)/L(\widehat{\theta})$ for the $n = 20$, 200, and 2000 samples and ML estimate for the $n =$10,220 sample of the $\alpha$-particle data. Vertical lines give corresponding approximate 95% likelihood confidence intervals.

**Example 2.5 Relative likelihood for the $\alpha$-particle data.** Figure 2.6 shows the *relative likelihood functions* for the samples of size $n = 2000$, 200 and 20 and a vertical line at the mean of all $n =$10,220 times.

Figure 2.6 indicates that the spread of the likelihood function tends to decrease as the sample size increases. The relative likelihood functions for the larger samples are much tighter than those for the smaller samples, indicating that the larger samples contain more information about $\theta$. The $\theta$ value at which the different likelihood functions are maximized is random and depends, in this comparison, on the results of the sampling described in Example 2.1. The four $\widehat{\theta}$'s differ, but they are consistent with the variability that one would expect from random sampling using the corresponding sample sizes. The fact that the ML estimates $\widehat{\theta}$ increase with $n$ is purely coincidental. ▮

### 2.3.3 Comparison of $\alpha$-particle data analyses

Figure 2.7 shows another exponential probability plot for the $n = 200$ sample. The solid line on this graph is the ML estimate of the exponential cdf $F(t;\theta)$. The dotted lines are drawn through a set of pointwise parametric normal-approximation 95% confidence intervals for $F(t;\theta)$; these parametric intervals quantify statistical uncertainty and will be explained in Section 2.4. Table 2.2 summarizes the results of fitting exponential distributions to the four different samples in Table 2.1; it includes ML estimates, standard errors, and confidence intervals. Section 2.4 provides results specifically for $\theta$, the mean (which is also the .632 quantile, $t_{.632}$) of the exponential distribution. Section 2.5.1 shows how to obtain similar estimates and confidence intervals for $\lambda = 1/\theta$, the arrival intensity rate (mean number of particles arriving per unit of time).

## 2.4 Confidence Intervals for $\theta$

The likelihood function provides a versatile method for assessing the information that the data contains on parameters, or functions of parameters. In particular, the likelihood function provides a generally useful method for testing hypotheses about or finding approximate confidence intervals for parameters and functions of parameters.

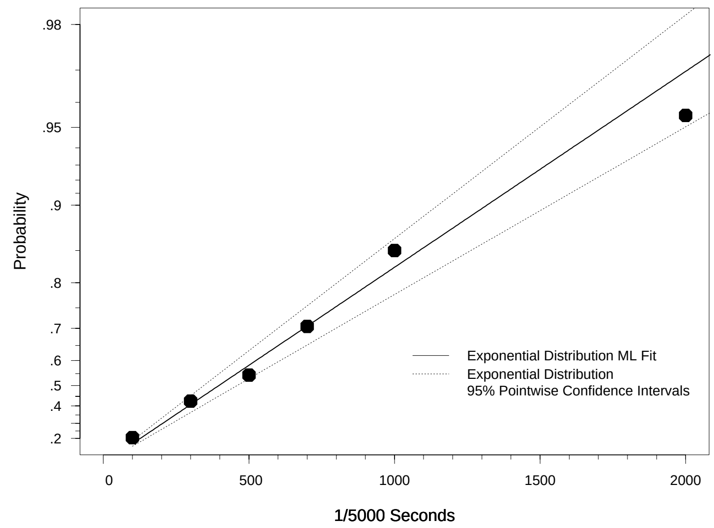


Figure 2.7: Exponential probability plot of the $n = 200$ sample of $\alpha$-particle interarrival time data. The plot also shows the parametric exponential ML estimate and approximate 95% confidence intervals for $F(t;\theta)$.

Table 2.2: Comparison of $\alpha$-particle ML results.

| | All Times | Sample of Times | | |
|---|---|---|---|---|
| | $n = 10{,}220$ | $n = 2{,}000$ | $n = 200$ | $n = 20$ |
| ML Estimate $\widehat{\theta}$ | 596.3 | 612.8 | 572.3 | 440.2 |
| Standard Error $\widehat{\text{se}}_{\widehat{\theta}}$ | 6.084 | 14.13 | 41.72 | 101.0 |
| Approximate 95% Confidence Intervals for $\theta$ Based on | | | | |
| Likelihood | [584, 608] | [586, 641] | [498, 662] | [289, 713] |
| $Z_{\log(\widehat{\theta})} \dot\sim \text{NOR}(0,1)$ | [584, 608] | [586, 641] | [496, 660] | [281, 690] |
| $Z_{\widehat{\theta}} \dot\sim \text{NOR}(0,1)$ | [584, 608] | [585, 640] | [490, 653] | [242, 638] |
| ML Estimate $\widehat{\lambda} \times 10^5$ | 168 | 163 | 175 | 227 |
| Standard Error $\widehat{\text{se}}_{\widehat{\lambda} \times 10^5}$ | 1.7 | 3.8 | 13 | 52 |
| Approximate 95% Confidence Intervals for $\lambda \times 10^5$ Based on | | | | |
| Likelihood | [164, 171] | [156, 171] | [151, 201] | [140, 346] |
| $Z_{\log(\widehat{\lambda})} \dot\sim \text{NOR}(0,1)$ | [164, 171] | [156, 171] | [152, 202] | [145, 356] |
| $Z_{\widehat{\lambda}} \dot\sim \text{NOR}(0,1)$ | [164, 171] | [156, 171] | [149, 200] | [125, 329] |

### 2.4.1 Relationship between confidence intervals and significance tests

Significance testing (sometimes called hypothesis testing) is a statistical technique that is widely used in many areas of science. The basic idea is to assess the reasonableness of a claim or hypothesis about a model or parameter value, relative to observed data.

A likelihood ratio test for a single-parameter model can be done by comparing the value of the likelihood under the "null hypothesis" (say $\theta_0$) to the maximum of the likelihood over all possible values for $\theta$. A likelihood (probability of the data) that is much smaller under the null hypothesis provides evidence to refute the hypothesis. Specifically, the null hypothesis that $\theta = \theta_0$ should be rejected at the $\alpha$ level of significance if

$$X^2 = -2 \log \left[ L(\theta_0)/L(\widehat{\theta}) \right] > \chi^2_{(1-\alpha;1)} \tag{2.11}$$

where $\widehat{\theta}$ is the ML estimate of $\theta$. This test is justified because when $\theta = \theta_0$, $X^2$ has a chisquare distribution with one degree of freedom. Rejection of the null hypothesis implies that the data are not consistent with the value of $\theta_0$, again at the $\alpha$ level of significance.

**Example 2.6 Likelihood-ratio test for the mean time between arrivals of $\alpha$ particles.** Suppose that investigators conducted the $\alpha$-particle experiment to test the hypothesis that the mean time between arrivals of $\alpha$ particles is $\theta = 650$. Based on the confidence intervals for $n$ =2000 in Table 2.2, we would conclude that there is not enough evidence to reject this hypothesis. Correspondingly,

$$X^2 = -2 \log [L(650)/L(572.3)] = 2.94 < \chi^2_{(.95;1)} = 3.84$$

again showing that there is not sufficient evidence in the $n = 200$ sample to reject the hypothesis. Using the $n$ =2000 sample, however, does provide sufficient evidence to reject the hypothesis that $\theta = 650$ at the 5% level of significance, as 650 is not in the 95% confidence interval. ∎

In many areas of application (especially engineering, medicine, and business) it has been recognized that confidence intervals provide a much more informative method of presentation of the results of a study than the simple yes-no conclusion of an hypothesis test. A confidence interval allows one to immediately visualize the *size* of an effect or deviation from a standard. Thus, a confidence interval also allows one to assess practical significance (also known as engineering significance or clinical significance in engineering and medical applications, respectively) in addition to statistical significance. See pages 39-40 of Hahn and Meeker (1991) and other references given there for further discussion of this subject.

A $100(1-\alpha)$% confidence interval for $\theta$ can be obtained by determining the set of all possible values of $\theta_0$ for which the null hypothesis would not be rejected at the $100\alpha$% level of significance. Correspondingly, one can test a hypothesis by first constructing a $100(1-\alpha\%)$ confidence interval for the quantity of interest and then checking to see if the interval encloses the hypothesized value or not. If not, then the hypothesis is rejected "at the $\alpha$ level of significance." If the interval encloses the hypothesized value, then the appropriate conclusion is that the data are consistent with the hypothesis (it is important, however, to note that failing to reject a hypothesis is not the same as saying that the hypothesis is true—see the following examples).

### 2.4.2 Likelihood confidence intervals for $\theta$

A likelihood-based confidence interval is the set of all values of $\theta$ that would not be rejected under the likelihood ratio test defined in (2.11). Thus, an approximate $100(1-\alpha)$% likelihood-based confidence interval for $\theta$ is the set of all values of $\theta$ such that

$$-2 \log[R(\theta)] \leq \chi^2_{(1-\alpha;1)}$$

or, equivalently, the set defined by

$$R(\theta) \geq \exp\left[-\chi^2_{(1-\alpha;1)}/2\right].$$

**Example 2.7 Likelihood confidence intervals for the mean time between arrivals of $\alpha$ particles.** Figure 2.6 illustrates likelihood confidence intervals. The horizontal line at $\exp[-\chi^2_{(.95;1)}/2] = .147$, corresponds to approximate 95% confidence intervals. The vertical lines dropping from the respective curves give the endpoints of the confidence intervals for the different samples. Table 2.2 gives numerical values of likelihood-based approximate 95% confidence intervals (as well as intervals based on other methods to be explained subsequently). Figure 2.8 shows that increasing sample size tends to reduce confidence interval length. Approximate (large-sample) theory shows that confidence interval length under standard regularity conditions is approximately proportional to $1/\sqrt{n}$. ∎

A one-sided approximate $100(1-\alpha)\%$ confidence bound can be obtained by drawing a horizontal line at $\exp[-\chi^2_{(1-2\alpha;1)}/2]$ and using the appropriate end point of the resulting two-sided confidence interval.

**Example 2.8 One-sided likelihood-based confidence bounds for the mean time between arrivals of $\alpha$ particles.** Referring to Figure 2.6, the horizontal line at $\exp[-\chi^2_{(.95;1)}/2] = .147$ would provide one-sided approximate 97.5% confidence bounds for $\theta$. For one-sided approximate 95% confidence bounds the line would be drawn at $\exp[-\chi^2_{(.90;1)}/2] = .259$ (corresponding to .90 on the right-hand scale on Figure 2.6). ∎

### 2.4.3 Normal-approximation confidence intervals for $\theta$

A $100(1-\alpha)\%$ normal-approximation confidence interval for $\theta$ is

$$[\underset{\sim}{\theta}, \quad \tilde{\theta}] = \widehat{\theta} \pm z_{(1-\alpha/2)} \widehat{\text{se}}_{\widehat{\theta}}. \tag{2.12}$$

A one-sided approximate $100(1-\alpha)\%$ confidence bound can be obtained by replacing $z_{(1-\alpha/2)}$ with $z_{(1-\alpha)}$ in (2.12) and using the appropriate endpoint of the resulting two-sided confidence interval.

An estimate of the standard error of $\widehat{\theta}$ is typically computed from the "observed information" as

$$\widehat{\text{se}}_{\widehat{\theta}} = \sqrt{\left[-\frac{d^2\mathcal{L}(\theta)}{d\theta^2}\right]^{-1}} \tag{2.13}$$

where the second derivative is evaluated at $\widehat{\theta}$. The second derivative measures curvature of $\mathcal{L}(\theta)$ at $\widehat{\theta}$. If $\mathcal{L}(\theta)$ is approximately quadratic, large curvature implies a narrow likelihood and thus a small estimate of the standard error of $\widehat{\theta}$.

The approximate confidence interval in (2.12) is based on the assumption that the distribution of

$$Z_{\widehat{\theta}} = \frac{\widehat{\theta} - \theta}{\widehat{\text{se}}_{\widehat{\theta}}} \tag{2.14}$$

can be approximated by a NOR$(0, 1)$ distribution. Then

$$\Pr\left[z_{(\alpha/2)} < Z_{\widehat{\theta}} \le z_{(1-\alpha/2)}\right] \approx 1 - \alpha \tag{2.15}$$

where $z_{(\alpha/)} = \Phi^{-1}_{\text{nor}}(\alpha/)$ is the $\alpha$ quantile of the standard normal distribution. Equation (2.15) implies that

$$\Pr\left[\widehat{\theta} - z_{(1-\alpha/2)}\widehat{\text{se}}_{\widehat{\theta}} < \theta \le \widehat{\theta} + z_{(1-\alpha/2)}\widehat{\text{se}}_{\widehat{\theta}}\right] \approx 1 - \alpha \tag{2.16}$$

because $z_{(1-\alpha/2)} = -z_{(\alpha/2)}$. The approximation is usually better for large samples, but may be poor for small samples. Meeker and Escobar (1995) show that a normal-approximation confidence interval can be viewed as an approximation to a likelihood-ratio based interval in which the loglikelihood is approximated by a quadratic function. See Appendix Section B.6 of Meeker and Escobar (1998) for more information on such large-sample approximations.

**Example 2.9 Normal-approximation confidence intervals for the mean time between arrivals of $\alpha$ particles.** For the $n = 200$ $\alpha$-particle data $\widehat{\text{se}}_{\widehat{\theta}} = 41.72$ and an approximate 95% confidence interval for $\theta$ based on the assumption that $Z_{\widehat{\theta}} \dot\sim \text{NOR}(0, 1)$ is

$$[\underset{\sim}{\theta}, \quad \tilde{\theta}] = 572.3 \pm 1.960(41.72) = [490, \quad 653].$$

Thus we are (approximately) 95% confident that $\theta$ is in this interval. ▮

An alternative approximate confidence interval for positive quantities like $\theta$ is

$$[\underset{\sim}{\theta}, \quad \tilde{\theta}] = [\widehat{\theta}/w, \quad \widehat{\theta} \times w] \tag{2.17}$$

where $w = \exp(z_{(1-\alpha/2)}\widehat{\text{se}}_{\widehat{\theta}}/\widehat{\theta})$. This interval is based on the assumption that the distribution of

$$Z_{\log(\widehat{\theta})} = \frac{\log(\widehat{\theta}) - \log(\theta)}{\widehat{\text{se}}_{\log(\widehat{\theta})}} \tag{2.18}$$

can be approximated by a NOR(0, 1) distribution where $\widehat{\text{se}}_{\log(\widehat{\theta})} = \widehat{\text{se}}_{\widehat{\theta}}/\widehat{\theta}$ is obtained by using the delta method (described in Appendix Section B.2 of Meeker and Escobar 1998). The confidence interval in (2.17) follows because an approximate $100(1 - \alpha)\%$ confidence interval for $\log(\theta)$ is

$$[\underset{\sim}{\log}(\theta), \quad \tilde{\log}(\theta)] = \log(\widehat{\theta}) \pm z_{(1-\alpha/2)}\widehat{\text{se}}_{\log(\widehat{\theta})}.$$

For a parameter, like $\theta$, that must be positive, (2.17) is often suggested as providing a more accurate approximate interval than (2.12). Although there is no guarantee that (2.17) will be more accurate than (2.12) in a particular setting, the sampling distribution of $Z_{\log(\widehat{\theta})}$ is usually more symmetric than that of $Z_{\widehat{\theta}}$ and the log transformation ensures that the lower endpoint of the confidence interval will be positive [which is not always so for confidence intervals based on (2.12)].

**Example 2.10 Normal-approximation confidence intervals for the mean time between arrivals of $\alpha$ particles.** For the $\alpha$-particle data, an approximate 95% confidence interval for $\theta$ based on the assumption that $Z_{\log(\widehat{\theta})} \dot\sim \text{NOR}(0, 1)$ is

$$[\underset{\sim}{\theta}, \quad \tilde{\theta}] = [572.3/1.1536, \quad 572.3 \times 1.1536] = [496, \quad 660]$$

where $w = \exp\{1.960 \times 41.72/572.3\} = 1.1536$. ▮

## 2.5 Confidence Intervals for Functions of $\theta$

For one-parameter distributions like the exponential, confidence intervals for $\theta$ can be translated directly into confidence intervals for monotone decreasing or monotone increasing functions of $\theta$.

### 2.5.1 Confidence intervals for the arrival rate

The arrival rate $\lambda = 1/\theta$ is a *decreasing* function of $\theta$. Thus the upper limit $\tilde{\theta}$ is substituted for $\theta$ to get a lower limit for $\lambda$ and vice versa. A confidence interval for $\lambda$ obtained in this manner will contain $\lambda$ if and only if the corresponding interval for $\theta$ actually contains $\theta$. Thus the confidence interval for $\lambda$ has the same confidence level as the interval for $\theta$.

**Example 2.11 Likelihood-based confidence intervals for the arrival rate of $\alpha$ particles.** Using the likelihood-based confidence interval for the $n = 200$ sample in Table 2.2,

$$[\underset{\sim}{\lambda}, \quad \tilde{\lambda}] = [1/\tilde{\theta}, \quad 1/\underset{\sim}{\theta}] = [.00151, \quad .00201].$$

Also, the ML estimate of $\lambda$ is obtained as $\widehat{\lambda} = 1/\widehat{\theta} = .00175$. ▮

### 2.5.2 Confidence intervals for $F(t;\theta)$

Because $F(t;\theta)$ is a decreasing function of $\theta$, a confidence interval for the exponential distribution $F(t_e;\theta)$ for a particular $t_e$ is

$$[\underset{\sim}{F}(t_e), \quad \widetilde{F}(t_e)] = [F(t_e;\tilde{\theta}), \quad F(t_e;\underset{\sim}{\theta})].$$

One can compute a set of pointwise confidence intervals for a range of values of $t$. The set can, in this case, also be interpreted as simultaneous confidence bands for the entire exponential cdf $F(t;\theta)$. This is because $\theta$ is the only unknown parameter for this model and the bands will contain the unknown exponential cdf $F(t;\theta)$ if and only if the corresponding confidence interval for $\theta$ contains the unknown true $\theta$.

**Example 2.12 Confidence intervals for the $\alpha$-particle time between arrivals cdf.** The dotted lines in Figure 2.7 are drawn through a set of pointwise normal-approximation 95% confidence intervals for the exponential $F(t;\theta)$. For example, the fraction of $\alpha$-particle interarrival times less than 1/5 of a second (1000/5000 time units) is $F(1000; 572.3) = .826$. An approximate 95% confidence interval for $F(1000;\theta)$, based on the 95% likelihood confidence interval for $\theta$ given in Table 2.2, is

$$[\underset{\sim}{F}, \quad \widetilde{F}] = [F(1000; 662), \quad F(1000; 498)] = [.779, \quad .866]$$

Thus we are (approximately) 95% confident that $F(1000;\theta)$ lies in the interval [.779, .866]. ▮

Confidence intervals for other quantities, such as distribution quantiles are found in a similar manner.

In subsequent chapters where models have more than one parameter, interpretation of a collection of intervals must be handled differently because the confidence level applies only to the process of constructing an interval for a *single point* in time $t_e$. Generally, making a *simultaneous* statement would require either a wider set of bands or a lower level of confidence.

## 2.6 ML Estimation with Exact and Right-Censored Observations

Often some data from a study will be reported as "exact" values, even though data are always limited by the precision of the measuring instrument or are rounded in some manner. Additionally, when the response is the time to some event, there is often a form of censoring called right censoring.

**Example 2.13 Remission times for patients receiving a particular leukemia therapy.** Lawless (1982, page 136) gives the results of a study to investigate the effect of a certain kind of therapy for 20 leukemia patients. After the therapy, patients go into remission for some period of time, the length of which is random. The observed times were 1, 1, 2, 2, 2, 6, 6, 6, 7, 8, 9, 9, 10, 12, 13, 14, 18, 19 24, 26, 29, 31*, 42, 45*, 50*, 57, 60, 71*, 85*, 91 weeks. The times marked with a * indicate patients who were still in remission at the time that the data were analyzed. These are known as right-censored observations because all that is known about them is that they did not come out of remission up to the given time and, presumably, would have come out at some point in time (to the right) of the observed survival times. ▮

### 2.6.1 Correct likelihood for observations reported as exact values

Consider the leukemia remission data. Although the times were reported as exact times, these data (as with most data) are actually discrete. In this case, a time reported as $t$ weeks probably indicates that the patient came out of remission at a time that rounded to $t$ weeks after ending therapy. Thus the "correct likelihood" is one for interval-censored data (2.9). For example, with the exponential distribution, the likelihood contribution (probability) of the 21st observation recorded as 29 weeks is

$$L_{21}(\theta) = \Pr(28.5 < T \leq 29.5) = F(29.5;\theta) - F(28.5;\theta). \tag{2.19}$$

### 2.6.2 Using the density approximation for observations reported as exact values

The traditional and commonly used form of the likelihood for an observation, say the $i$th, reported as an "exact" time $t_i$, is

$$L_i(\theta) = f(t_i; \theta) \tag{2.20}$$

where $f(t_i; \theta) = dF(t_i; \theta)/dt$ is the assumed pdf for the random variable $T$. This density approximation in (2.20) is convenient, easy-to-use, and, in some simple special cases, yields closed form equations for ML estimates.

The use of the density approximation (2.20) instead of the correct discrete likelihood can be justified as follows. For most statistical models, the contribution to the likelihood (i.e., probability of the data) of observations reported as exact values can, for small $\Delta_i > 0$, be approximated by

$$[F(t_i; \theta) - F(t_i - \Delta_i; \theta)] \approx f(t_i; \theta)\Delta_i \tag{2.21}$$

where $\Delta_i$ does not depend on $\theta$. Because the right-hand sides of (2.21) and (2.20) differ by a factor of $\Delta_i$, when the density approximation is used, the approximate likelihood in (2.20) differs from the probability in (2.21) by a constant scale factor. As long as the approximation in (2.21) is adequate and because $\Delta_i$ does not depend on $\theta$, the general character of the likelihood (i.e., the shape and the location of the maximum) is not affected.

### 2.6.3 Right-censored observations

The observations in Example 2.13 that are marked with a * are right-censored observations because all we know is that the time to end of remission for these patients is beyond (to the right of) the recorded time. These right-censored observations can also be thought of as intervals running from from the recorded time to $\infty$. In general, for a right-censored observation $t_i$ for patient $i$, the probability of the observation is

$$L_i(\boldsymbol{\theta}) = \Pr(T > t_i) = F(\infty; \boldsymbol{\theta}) - F(t_i; \boldsymbol{\theta}) = 1 - F(t_i; \boldsymbol{\theta}).$$

This probability is illustrated by the right-censored observation in Figure 2.5.

### 2.6.4 Likelihood with exact and right-censored observations

The likelihood for a sample containing exact observations and right censored observations can be written as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \{f(t_i; \boldsymbol{\theta})\}^{\delta_i} \{1 - F(t_i; \boldsymbol{\theta})\}^{1-\delta_i}, \tag{2.22}$$

where $\delta_i = 1$ for an "exact" observation and $\delta_i = 0$ for a right-censored observation. Substituting in the exponential distribution pdf and cdf into (2.22) gives

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{\frac{1}{\theta} \exp\left[-\frac{t_i}{\theta}\right]\right\}^{\delta_i} \left\{\exp\left[-\frac{t_i}{\theta}\right]\right\}^{1-\delta_i}. \tag{2.23}$$

### 2.6.5 Assumptions concerning censoring

Use of the likelihood in (2.23) requires strong assumptions about the nature of the censoring. Consider a life time study. For each person in the study, we observe, at a given time, either a death or a survival. Simply put, to use (2.23) it is necessary that the censoring be "noninformative." For example, for a censored unit, what ever caused the unit to be removed from observation should not provide any information about the time that the unit would have failed. The assumption would be violated, for example, if a patient left a study, before death, be cause of deteriorating health.

### 2.6.6 ML estimates for the exponential $\theta$ based on the density approximation

The density approximation to the likelihood provides an adequate approximation for the exponential distribution. For a sample consisting of only right censored observations and observations reported as exact values (and no left-censored or interval-censored observations), one can show (see exercise 2.1) that the ML estimate of $\theta$ is computed as

$$\widehat{\theta} = \frac{TTT}{r} \tag{2.24}$$

where $r = \sum_{i=1}^{n} \delta_i$ is the number of exact values observed (as opposed to censored observations) and $TTT = \sum_{i=1}^{n} t_i$ is known as the "total time on test" and where $t_i$, $i = 1, n$ are the reported times for units with exact values and the censoring time for the right-censored observations. Note that the sum runs over *all* of the exact *and* each of the censoring values. In this case, an estimate of standard error of $\widehat{\theta}$ is computed as

$$\widehat{\mathrm{se}}_{\widehat{\theta}} = \sqrt{\left[-\frac{d^2\mathcal{L}(\theta)}{d\theta^2}\right]\Bigg|_{\widehat{\theta}}^{-1}} = \sqrt{\frac{\widehat{\theta}^2}{r}} = \frac{\widehat{\theta}}{\sqrt{r}}. \tag{2.25}$$

### 2.6.7 Confidence interval for the exponential $\theta$ with complete and failure-censored data.

If the data are complete or failure censored, a $100(1-\alpha)\%$ confidence interval for $\theta$ is

$$[\underset{\sim}{\theta}, \quad \tilde{\theta}] = \left[\frac{2(TTT)}{\chi^2_{(1-\alpha/2;2r)}}, \quad \frac{2(TTT)}{\chi^2_{(\alpha/2;2r)}}\right]. \tag{2.26}$$

This follows because the distribution of $\widehat{\theta}$ is proportional to a chisquare distribution with $2r$ degrees of freedom. That is, $2TTT/\theta \sim \chi^2_{2r}$. Then for other kinds of censoring this formula can be used to obtain an approximate confidence interval.

**Example 2.14 Exponential distribution fit to the remission times.** Figure 2.8 is an exponential probability plot of the remission time data with approximate 95% simultaneous confidence bands for the cdf. The points show deviation from a straight line, but due to the width of the bands, it is not possible to rule out the exponential distributions as an appropriate model for these data. Figure 2.8 shows the (straight line) exponential distribution ML estimate and confidence intervals plotted on an exponential probability plot, making it easier to see deviation from linearity. ∎

### 2.6.8 Fitting the lognormal distribution to right-censored data

Although the exponential distribution could not be ruled out as a model for the remission times, the lack of fit suggested that there might be an alternative distribution that would provide a better description of the distribution. The following example considers the lognormal distribution.

**Example 2.15 Lognormal probability plot of the remission times.** Figure 2.10 is a lognormal probability plot of the remission data. The lack of an important deviation from linearity shows that the lognormal distribution provides a much better description of these data. ∎

### 2.6.9 Joint confidence region for $\boldsymbol{\theta}$

For a distribution with two-parameters (say $\theta_1$ and $\theta_2$), a level-curve on a relative likelihood contour plot defines an approximate joint confidence region for $\theta_1$ and $\theta_2$. The contour line can be accurately calibrated, even in moderately small samples, by using the large sample $\chi^2$ approximation for the distribution of the corresponding likelihood-ratio statistic, analogous to the approach used in Section 2.4.2 for a single parameter distribution. In particular, the region $R(\theta_1, \theta_2) > \exp(-\chi^2_{(1-\alpha;2)}/2) = \alpha$ provides an approximate $100(1-\alpha)\%$ joint confidence region for $\boldsymbol{\theta}$.

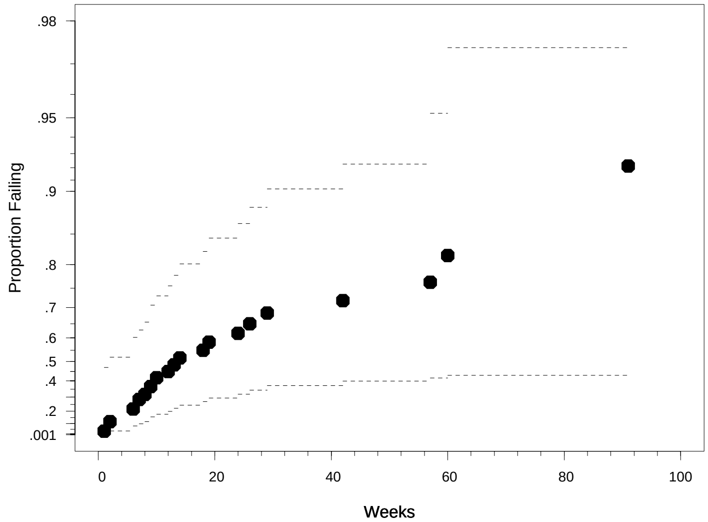


Figure 2.8: Exponential probability plot of the remission times with approximate 95% simultaneous confidence bands.

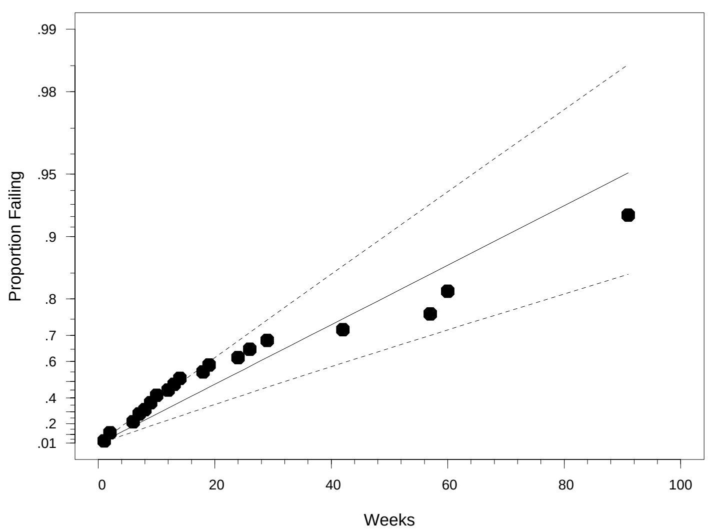


Figure 2.9: Exponential probability plot of remission times with maximum likelihood estimates and approximate 95% pointwise confidence intervals for $F(t; \mu, \sigma)$.

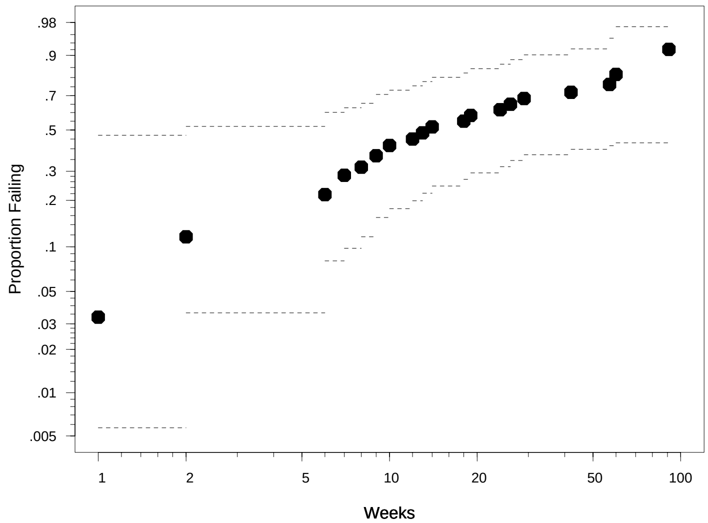


Figure 2.10: Lognormal probability plot of the remission times with approximate 95% simultaneous confidence bands for $F(t;\mu,\sigma)$.

**Example 2.16 Lognormal fit to the remission times.** The likelihood for the lognormal distribution with right-censored data is obtained by substituting the lognormal pdf and cdf into (2.22), where $\boldsymbol{\theta} = (\mu, \sigma)$, giving

$$L(\mu,\sigma) = \prod_{i=1}^{n} \left\{ \frac{1}{\sigma}\phi_{\text{nor}}\left[\frac{\log(t_i)-\mu}{\sigma}\right]\right\}^{\delta_i} \left\{1-\Phi_{\text{nor}}\left[\frac{\log(t_i)-\mu}{\sigma}\right]\right\}^{1-\delta_i}.$$

Figure 2.11 shows the lognormal likelihood contours for $\sigma$ and $t_{.5} = \exp(\mu)$. Figure 2.12 shows the shows the corresponding approximate joint confidence regions for $\sigma$ and $t_{.5} = \exp(\mu)$. Figure 2.13 shows the lognormal distribution fit to the remission times, showing that the fit is excellent. ∎

### 2.6.10 Confidence intervals for elements of $\boldsymbol{\theta}$ and for scalar functions of $\boldsymbol{\theta}$

Often interest focuses on a single parameter or a scalar function of a set of parameters. In order to do this, we use a one-dimensional "profile" relative likelihood function. The profile likelihood looks and is used like the one-dimensional relative likelihood in a single-parameter distribution (as described and illustrated in Section 2.3.2). If the two-parameter relative function can be viewed as a mountain, the profile likelihood can be thought of as the profile of the mountain against the horizon, as viewed from the coordinate axis of the parameter of interest. Profile likelihoods for functions of $\boldsymbol{\theta}$ can be obtained through reparameterization so that the function of interest replaces one of the parameters. Details on and examples of the computation of a profile likelihoods are given in Chapter 8 of Meeker and Escobar (1998).

**Example 2.17 Profile likelihoods for the lognormal distribution parameters from the remission time data.** Figure 2.14 gives a profile plots for $t_{.5}$, the median of the remission time distribution. Note that one can read the likelihood-based confidence interval from this plot. Numerically, the approximate 95% confidence interval is $[8.9, \quad 27.3]$. Figure 2.15 is a profile plot for $\sigma$, the lognormal scale parameter. Numerically, the approximate 95% confidence interval for $\sigma$ is $[1.13, \quad 2.02]$. ∎

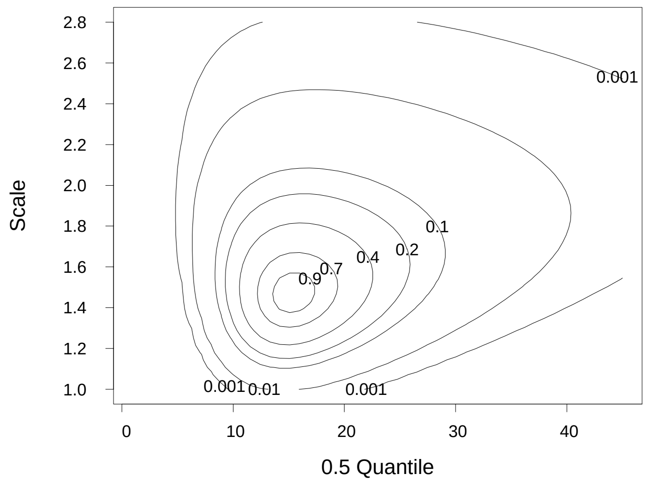


Figure 2.11: Relative likelihood for the remission data and the lognormal distribution.

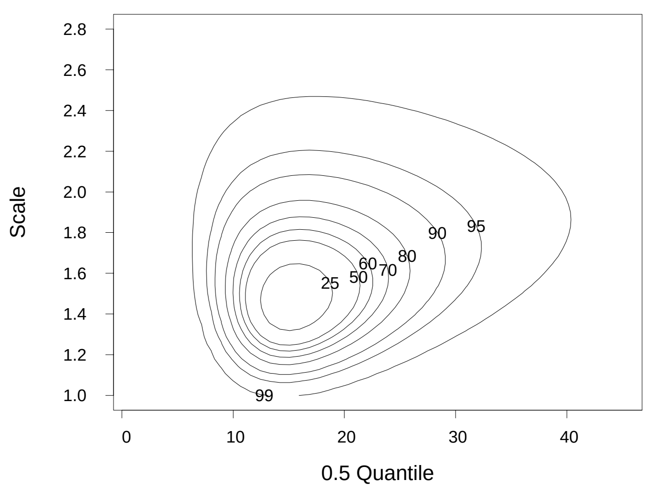


Figure 2.12: Lognormal distribution parameter joint confidence regions for the remission data.

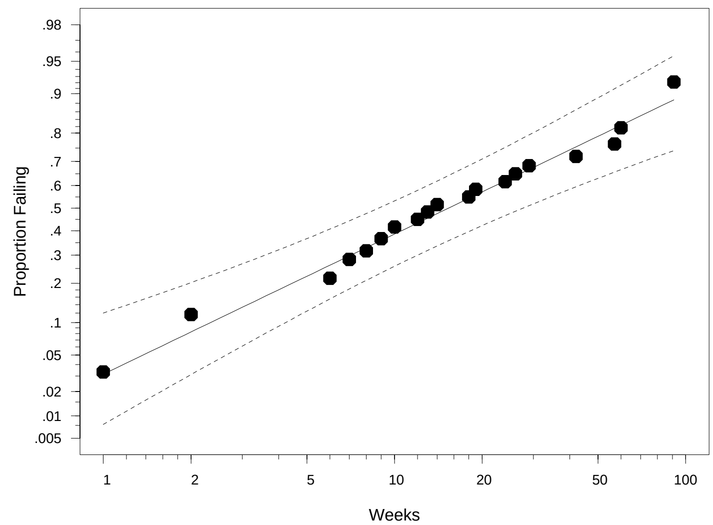


Figure 2.13: Lognormal probability plot of remission times with maximum likelihood estimates and approximate 95% pointwise confidence intervals for $F(t; \mu, \sigma)$.

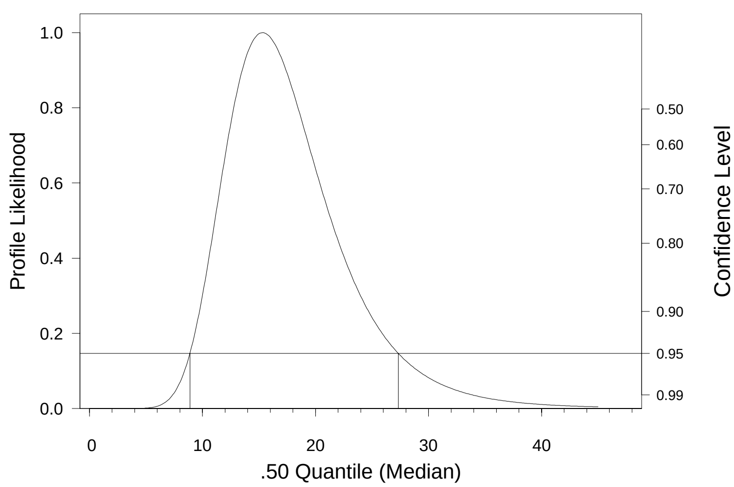


Figure 2.14: Lognormal profile likelihood $R[\exp(\mu)]$ $(\exp(\mu) = t_{.5})$ for the remission time data.

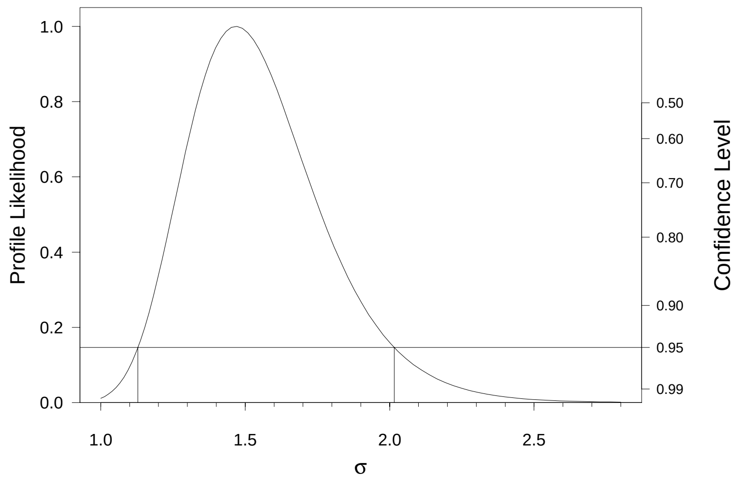


Figure 2.15: Lognormal profile likelihood $R[\exp(\sigma)]$ for the remission time data.

## 2.7 Fitting Distributions with Left-Censored Observations

### 2.7.1 Left-censored observations

**Example 2.18 X-ray flux-ratio data from a sample of active galaxies** Table 2.3 gives the ratio of optical to X-ray flux for 107 active galaxies. The data first reported in Zamorani et al. (1981). The data were also analyzed by Feigelson and Nelson (1985). The numbers marked with a $<$ are left-censored (upper bounds on the actual values) that resulted from limitations in observation sensitivity. Investigators were interested in quantifying the distribution of this ratio among galaxies in a larger population, in order to gain potential insight into possible physical causes for the different types of radiation. The analysis of the data was complicated by the left censoring. ∎

Left-censored observations arise when we have an upper bound on a response. Such observations are common when an instrument lacks the sensitivity needed to measure observations below a threshold, the value of which is known.

In general, if we have an upper bound $t_i$ for observation $i$, the probability of the observation is

$$L_i(\boldsymbol{\theta}) = \Pr(-\infty < T \leq t_i) = F(t_i; \boldsymbol{\theta}) - F(-\infty; \boldsymbol{\theta}) = F(t_i; \boldsymbol{\theta}). \tag{2.27}$$

If the lower limit of support for $F(t; \boldsymbol{\theta})$ is 0 (e.g., when $T$ has to be positive as in a concentration level, tensile strength, or waiting time), we replace $-\infty$ with 0 in equation (2.27).

**Example 2.19 Assessing the adequacy of the Weibull distribution to describe the X-ray flux data.** Figure 2.16 is a Weibull probability plot of the X-ray flux-ratio data, constructed using methods similar to those described in Section 2.2.5 (and described in detail in Chapter 6 of Meeker and Escobar 1998). The plotted points in the lower tail of the distribution seem to depart somewhat from linearity, indicating a small departure from the Weibull distribution. Even if the distribution of galaxy flux-ratios could be described by a Weibull distribution, it is possible that such a departure could arise from the particular sample of galaxies that were chosen for observation (assuming that the sample was, in fact "random" from a list of such galaxies that could have been observed). The departure from linearity in Figure 2.16 is not pronounced. Figure 2.17 is a similar lognormal probability plot. Although the points in the lower tail seem to fit better in the lognormal probability plot, the Weibull distribution provides a better description over most of the range of the distribution.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1.52 | 1.39 | 1.50 | $<$1.50 | $<$1.57 | 1.21 | 1.38 | $<$1.63 |
| 1.41 | 1.06 | 1.00 | 1.54 | 1.68 | 1.48 | $<$1.59 | $<$1.68 |
| $<$1.82 | 1.60 | 1.58 | 1.26 | 1.48 | $<$1.63 | 1.31 | 1.47 |
| 1.56 | 1.66 | 1.15 | 1.28 | 1.34 | 1.57 | 1.00 | 1.46 |
| 1.25 | 1.44 | 1.67 | 1.38 | 1.43 | 1.26 | 1.46 | 1.22 |
| 1.42 | 1.15 | 1.14 | $<$1.38 | $<$1.82 | $<$1.38 | 1.12 | 1.71 |
| 1.44 | 1.30 | 1.30 | 1.51 | 1.14 | 1.25 | 1.53 | 1.62 |
| 1.22 | 1.40 | $<$1.25 | $<$1.24 | $<$1.17 | $<$1.23 | $<$1.56 | $<$1.50 |
| $<$1.15 | $<$1.18 | $<$1.40 | $<$1.33 | 1.38 | 1.70 | 1.43 | 0.97 |
| $<$1.47 | 1.54 | 1.55 | $<$1.71 | 1.71 | $<$1.32 | 1.48 | $<$1.57 |
| 1.87 | 1.59 | 1.52 | 1.33 | 1.38 | 1.45 | $<$1.04 | 1.21 |
| $<$1.12 | 1.42 | 1.65 | 1.21 | 1.46 | 1.66 | 1.55 | 1.36 |
| 1.30 | 1.22 | 1.30 | 1.68 | 1.61 | 0.91 | $<$1.36 | 1.18 |
| 1.55 | 1.44 | 1.59 | | | | | |

Table 2.3: Ratio of optical to X-ray flux for 107 active galaxies.

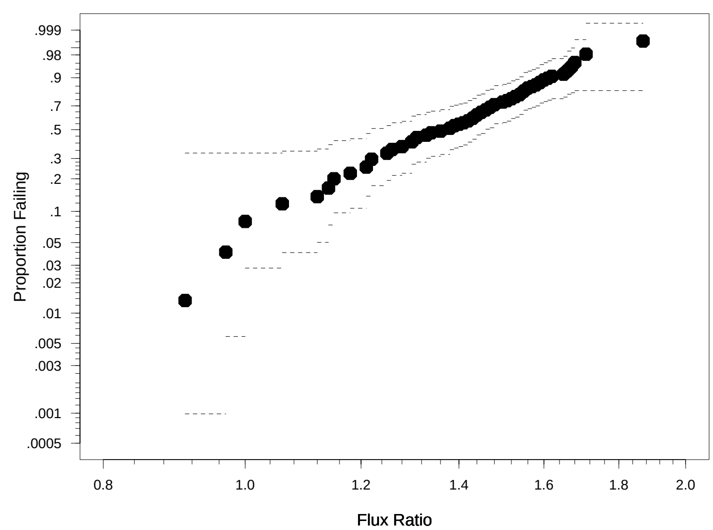


Figure 2.16: Weibull probability plot of X-ray flux-ratio data with maximum likelihood estimates and approximate 95% pointwise confidence intervals for $F(t; \mu, \sigma)$.

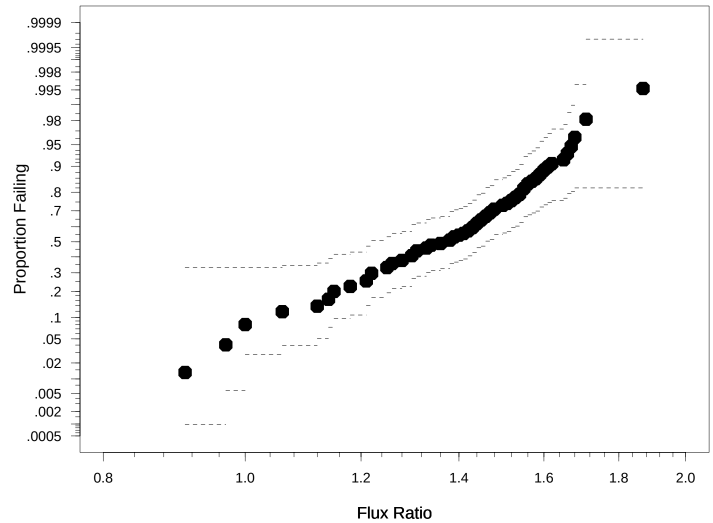


Figure 2.17: Lognormal probability plot of X-ray flux-ratio data with maximum likelihood estimates and approximate 95% pointwise confidence intervals for $F(t; \mu, \sigma)$.

In the next section, the 2-parameter Weibull and lognormal distributions are fit to the flux-ratio data using ML. This is equivalent to fitting a straight line through the data on the Weibull probability paper to estimate the cdf, using the ML criterion to choose the line. ∎

### 2.7.2 The likelihood function and its maximum

With exact and left-censored observations, the likelihood is:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \{F(t_i; \boldsymbol{\theta})\}^{1-\delta_i} \{f(t_i; \boldsymbol{\theta})\}^{\delta_i},$$

where $\delta_i = 1$ for an "exact" observation and $\delta_i = 0$ for a left-censored observation. For the Weibull distribution

$$L(\mu, \sigma) = \prod_{i=1}^{n} \left\{ \Phi_{\text{sev}} \left[ \frac{\log(t_i) - \mu}{\sigma} \right] \right\}^{1-\delta_i} \left\{ \frac{1}{\sigma} \phi_{\text{sev}} \left[ \frac{\log(t_i) - \mu}{\sigma} \right] \right\}^{\delta_i}, \tag{2.28}$$

where $\phi_{\text{sev}}(z) = d\Phi_{\text{sev}}(z)/dz = \exp[z - \exp(z)]$ is the standardized smallest extreme value density. ML estimates computed with the density approximation generally agreed with those computed with the "correct" likelihood to within plus or minus one in the third significant digit

**Example 2.20 Fitting the Weibull distribution to the X-ray flux-ratio data.** Figure 2.18 provides a contour plot of the relative likelihood function $R[\mu, \log(\sigma)] = L(\mu, \sigma)/L(\widehat{\mu}, \widehat{\sigma})$ for the Weibull distribution. The surface is well behaved with a unique maximum defining the ML estimates. The straight line on Figure 2.19 is $F(t; \widehat{\mu}, \widehat{\sigma})$, the ML estimate of the Weibull $F(t; \mu, \sigma)$. The curved line going through the points is the corresponding ML estimate of the lognormal $F(t; \mu, \sigma)$. The dotted lines are drawn through a set of pointwise normal-theory confidence intervals for the Weibull $F(t; \mu, \sigma)$; the corresponding set of intervals for the lognormal $F(t; \mu, \sigma)$ (not shown on the figure) were similar in width. As is frequently the case, there is good agreement for inferences from these two distributions *within the range of the data*. Without more information, however, one cannot make inferences in the lower tail of the distribution because the estimates are importantly different *and* the data do not strongly suggest one distribution over the other. In general, it is useful and important to fit different distributions to compare results on questions of interest. ∎

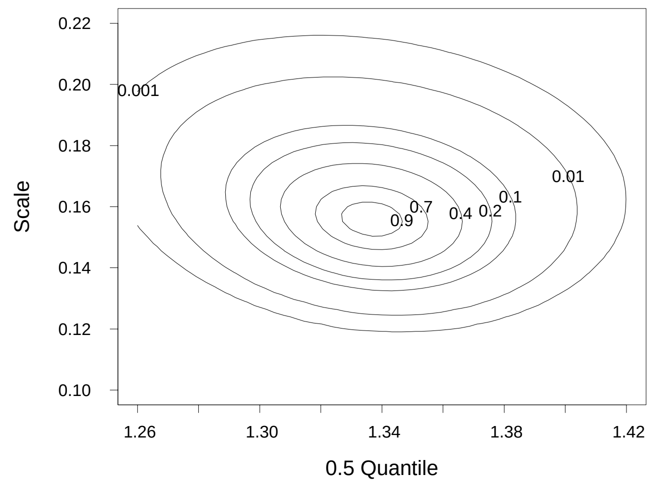


Figure 2.18: Relative likelihood for the X-ray flux-ratio data and the Weibull distribution.

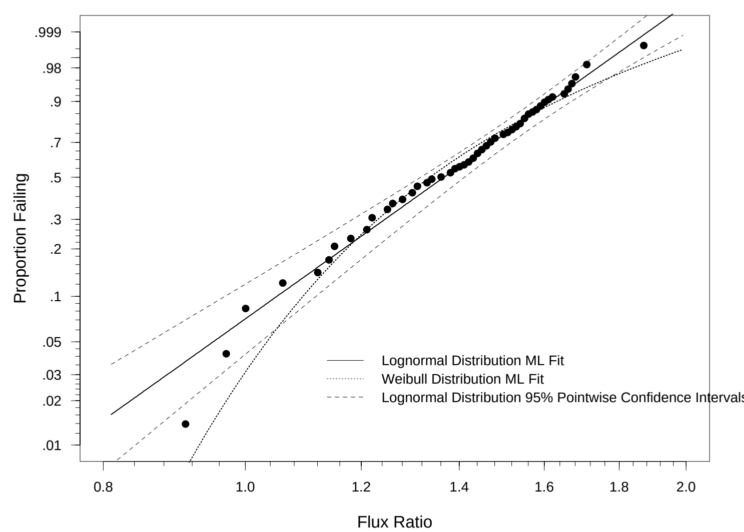


Figure 2.19: Weibull probability plot of the X-ray flux-ratio data with Weibull and lognormal maximum likelihood estimates and approximate 95% pointwise confidence intervals for the Weibull $F(t; \mu, \sigma)$.

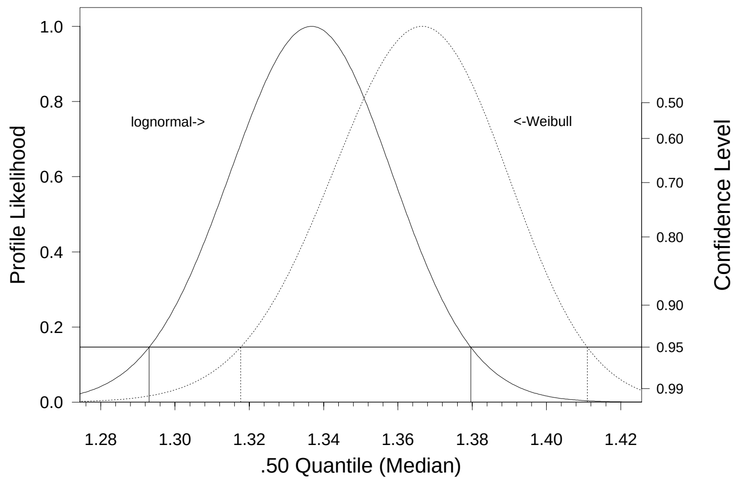


Figure 2.20: Comparison profile likelihoods for the median of the X-ray flux-ratio distribution using Weibull and lognormal distributions.

**Example 2.21 Joint confidence region for the Weibull $\mu$ and $\sigma$ for the X-ray flux ratio data.** A contour line on Figure 2.18 defines an approximate joint confidence region for $\mu$ and $\sigma$ that can be accurately calibrated, even in moderately small samples, by using the large sample $\chi^2$ approximation for the distribution of the likelihood-ratio statistic. For example, the region $R[\mu, \log(\sigma)] > \exp(-\chi^2_{(.90;2)}/2) = .100$ provides an approximate 90% joint confidence region for $\mu$ and $\sigma$. ∎

As shown in Section 2.6.10 profile likelihood functions for a particular quantity of interest are useful for visualizing the information provided by one's data for that particular quantity.

**Example 2.22 Profile likelihoods comparing Weibull and lognormal distribution estimates of the X-ray flux ratio distribution median.** Figure 2.20 compares profile plots for the $t_{.5}$, the median of the respective distributions. The ML estimates differ somewhat, but not importantly, relative to the width of the confidence intervals. This indicates some robustness with respect to the choice of a distribution to use. ∎

# Bibliographic Notes

The ML method dates back to early work by Fisher (1925) and has been used as a method for constructing estimators ever since. Statistical theory for ML is briefly covered in most textbooks on mathematical statistics. Some particularly useful references for the approach used here are Kulldorff (1961), Kempthorne and Folks (1971), Rao (1973), and Cox and Hinkley (1974), Lehmann (1983), Nelson (1982), Lawless (1982), Casella and Berger (1990), and Lindsey (1996). Epstein and Sobel (1953) described the fundamental ideas of using the exponential distribution as a model for life data.

For examples of failures of the "density approximation" to the likelihood, see Le Cam (1990), Friedman and Gertsbakh (1980) and Meeker and Escobar (1994). Each of these references describes a model or models for which the "density approximation" likelihood has a path or paths in the parameter space along which $L(\theta)$ approaches $\infty$. In these cases, using the "correct" likelihood based on discrete data eliminates the singularity in $L(\theta)$.

## References


Berkson, J. (1966), Examination of randomness of $\alpha$-particle emissions, in *Festschrift for J. Neyman, Research Papers in Statistics*, F. N. David, Editor. New York: John Wiley & Sons.

Casella, G., and Berger, R. L. (1990), *Statistical Inference*, Belmont, CA: Wadsworth.

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

Epstein, B., and Sobel, M. (1953), Life testing, *Journal of the American Statistical Association*, **48**, 486–502.

E. D. Feigelson and P. I. Nelson (1985), "Statistical methods for astronomical data with upper limits. I. univariate distributions," *The Astronomical Journal* **293**, 192-206.

R. A. Fisher (1925), "Theory of statistical estimation," *Proceedings of the Cambridge Philosophical Society* **22**, 700-725.

Friedman, L. B., and Gertsbakh, I. (1980), Maximum likelihood estimation in a minimum-type model with exponential and Weibull failure modes, *Journal of the American Statistical Association*, **75**, 460–465.

Hahn, G. J., and Meeker, W. Q. (1991), *Statistical Intervals: A Guide for Practitioners*, New York: John Wiley & Sons.

Kempthorne, O., and Folks, L. (1971), *Probability, Statistics, and Data Analysis*, Ames, IA: Iowa State University Press.

Kulldorff, G. (1961), *Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples*, Stockholm: Almqvist and Wiksell.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, New York: John Wiley & Sons.

Le Cam, L. (1990), Maximum likelihood: an introduction, *International Statistical Review*, **58**, 153–171.

Lindsey, J. K. (1996), *Parametric Statistical Inference*, Oxford: Clarendon Press.

Meeker, W. Q., and Escobar, L. A. (1994), Maximum likelihood methods for fitting parametric statistical models to censored and truncated data, in *Probabilistic and Statistical Methods in the Physical Sciences*, J. Stanford and S. Vardeman, Editors. New York: Academic Press.

Meeker, W. Q., and Escobar, L. A. (1995), Teaching about approximate confidence regions based on maximum likelihood estimation, *The American Statistician*, **49**, 48–53.

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

W. Nelson (1982), *Applied Life Data Analysis,* New York: John Wiley.

G. Ostrouchov and W. Q. Meeker (1988), "Accuracy of approximate confidence bounds computed from interval censored Weibull and lognormal data," *Journal of Statistical Computation and Simulation* **29**, 43-76.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons.

Zamorani et al. (1981), X-ray studies of quasars with the Einstein Observatory. II, *The Astrophysical Journal*, **245**, 357-374.

## Exercises

**2.1** Consider the case of $n$ observations reported as "exact" values (e.g., times) from an EXP($\theta$) distribution.

(a). Write down an expression for the likelihood.

(b). Write down an expression for the loglikelihood. Take the first derivative of this expression.

(c). Show that the ML estimate, $\widehat{\theta}$, of $\theta$ is the sample mean, say $\bar{t} = \sum_{i=1}^{n} t_i/n$.

(d). Show that the relative likelihood has the simple expression

$$R(\theta) = \exp(n)\left(\frac{\bar{t}}{\theta}\right)^{n} \exp\left(-\frac{n\bar{t}}{\theta}\right).$$

(e). Explain how to use $R(\theta)$ to obtain an approximate confidence interval for $\theta$ based on inverting a likelihood ratio test [i.e., assume that when evaluated at the true value of $\theta$, $R(\theta) \mathrel{\dot\sim} \chi^2_1$].

(f). Suppose that $n = 4$ and $\bar{t} = 0.87$. Obtain the ML estimate and an approximate 90% confidence interval for $\theta$ using the method in part (e). Plot $R(\theta)$ to facilitate your work.

**2.2** Start with equation (2.1) and derive the expression for the exponential distribution quantile $t_p = -\theta\log(1-p)$.

**2.3** The following table gives survival times for 81 patients that had been diagnosed with melanoma. The observations marked with a > represent patients who were lost to follow-up (e.g., some patients dropped out of the study because they had moved or for some other reasons) or who were still alive at the time the data were analyzed.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 136 | 58 | >55 | >181 | 21 | 23 | >190 | 65 |
| 234 | >194 | 14 | 90 | 20 | 130 | >213 | >215 |
| 124 | >108 | 54 | 98 | >193 | 138 | 141 | 110 |
| >67 | 50 | 26 | 103 | 59 | >134 | >147 | >162 |
| 65 | 40 | 34 | 57 | >81 | >152 | >125 | >151 |
| 34 | 158 | 27 | >148 | 27 | >132 | >140 | 32 |
| >130 | 38 | 85 | >129 | >100 | 19 | 118 | 53 |
| >140 | 66 | 46 | 37 | >50 | >114 | >124 | 26 |
| 102 | >93 | >80 | 60 | >86 | >21 | >44 | 23 |
| 80 | >73 | 19 | 38 | 31 | 25 | >76 | >13 |
| >16 | | | | | | | |

(a). Attempt to find an adequate distributional model to describe these data. Suppose that the purpose of the analysis is to present a survival curve to doctors who need to advice their patients. List the steps that you followed in your analysis and present some of the graphical displays that you used along the way.

(b). Describe what assumptions one needs to make about the censored observations. Discuss the reasonableness of those assumptions for this application.

**2.4** Nelson (1982) gives failure and censoring times for a turbine fans. The data were collected in order to estimate the failure-time distribution of the fans so that appropriate maintenance policies could be developed. The data are available as a SPLIDA object `Fan.ld`. Attempt to find an adequate distributional model to describe these data. List the steps that you followed in your analysis and present some of the graphical displays that you used along the way.

**2.5** Provide an intuitive explanation for the reason that likelihood-based approximate confidence intervals are generally better than normal-approximation approximate confidence intervals.

**2.6** Let $t_1, \ldots, t_r$ be the observed values (e.g., times) in a singly time-censored sample of size $n$ from an EXP($\theta$) distribution. Suppose that the prespecified censoring time is $t_c$. The total time on test is $TTT = \sum_{i=1}^{r} t_i + (n-r)t_c$.

(a). Write an expression for $\mathcal{L}(\theta)$, the log likelihood using the density approximation for the observations reported as an exact value.

(b). Show that the ML estimate of the exponential mean is $\widehat{\theta} = TTT/r$.

(c). Use the result of part (a) to show that the ML estimate of $\theta$ is equal to $\infty$ (or "does not exist") when all the observations are censored.

(d). Derive an expression for the relative likelihood similar to the one given in part (d) of Problem 2.1 (note that in this case $TTT/r$ plays the role of $\bar{t}$ and $r$ the role of $n$).

**2.7** A large electronic system contains thousands of copies of a particular component, which we will refer to as Component-A (each system is custom-made to order and the actual number of components varies with the size of the particular system). The failure rate for Component-A is small, but because there are so many of the components in operation, failures are reported from time to time. A field tracking study was conducted to estimate the failure rate of a new design for Component-A (which was intended to provide better reliability at lower cost), to see if there was any real improvement. A number of systems were put into service simultaneously and monitored for 1000 hours. The total number of copies of Component-A in all of the systems was 9432. Failures of Component-A were observed at 345, 556, 712, and 976 hours. Failure mode analysis suggested that the failures were due to random shocks rather than any kind of quality defects or wearout. This, along with past experience with simpler components, suggested that an exponential distribution might be an appropriate model for the lifetime of Component-A.

(a). Compute the ML estimate for the exponential mean $\theta$ for Component-A.

(b). Compute an approximate 95% confidence interval for $\theta$.

(c). Explain the interpretation of the confidence interval obtained in part (b).

**2.8** Example 2.3 describes and illustrates the use of an exponential probability plot.

(a). Show that solving for $p = F(t)$ in (2.1) gives the needed relationship to linearize $F(t)$ as a function of $t$.

(b). Explain how one could construct a probability plot given the observed fraction failing and a piece of regular linear graph paper.

(c). Use the method that you gave in part (b) to make, by hand, a probability plot of the data in exercise 2.7.

**2.9** Refer to Section 2.4.3.

(a). Show how equations (2.14) and (2.15) lead to equation (2.16).

(b). Explain the implication of this result. As part of your explanation, note which of the quantities in these equations are fixed and which are random.

(c). Explain the proper interpretation of a confidence interval (or a confidence interval procedure).

**2.10** Give an intuitive explanation for why the relative likelihood should play such an important role in statistical inference.

**2.11** Coverage probability is defined as the probability that a confidence interval procedure will contain the quantity that it is supposed to contain. Refer to Example 2.11 in the text. Explain why it is that the coverage probability for the confidence interval for $\lambda$ is the same as the coverage probability for the confidence interval for $\theta$.

**2.12** Equation (2.19) gives the correct likelihood assuming that the remission times were rounded to the nearest week.

(a). How should the "correct" likelihood term be written if the reported value of 29 indicated that the patient came out of remission in the 29th week? Draw a simple picture to illustrate.

(b). For which recording method is using the density approximation $f(29; \boldsymbol{\theta})$ more accurate? Explain.

(c). Explain the justification for the approximation in (2.21). Draw a simple picture to illustrate.

(d). Suggest a "refined" density approximation that might be used in cases where the density approximation cannot be used. Draw a simple picture to illustrate.

**2.13** Consider the remission data given in Example 2.13.

(a). Compute $\widehat{\theta}$ the ML estimate of the mean of the exponential distribution. Compare with output from an appropriate computer program.

(b). Compute and estimate of the standard error of $\widehat{\theta}$.

(c). Compute and compare 95% confidence intervals using the normal approximation methods given in Sectionr̃efsection:normal.theory.exponential [use,specifically, equations (2.17) and (2.12)] and the chisquare method in (2.26). Can you determine which approximation is being used by the computer?

**2.14** Give an intuitive explanation as to why the normal approximation method in (2.17) would be expected to be better than that in (2.12).

**2.15** Example 2.14 uses both the exponential and lognormal distributions to describe the remission time data. Try fitting some other distributions. Compare the estimates of coming out of remission after one year using these different distributions. Can you find any other adequate distributions? Are the numerical differences among the answers given by the different distributions important, relative to statistical uncertainty? How about from a clinical point of view?

**2.16** Find a likelihood confidence interval for the median (.5 quantile) of the $\alpha$-particle interarrival time distribution based on the $n = 200$ sample.