EDUCATION COMMITTEE

OF THE

SOCIETY OF ACTUARIES

LONG-TERM ACTUARIAL MATHEMATICS STUDY NOTE

CHAPTERS 10–12 OF

LOSS MODELS, FROM DATA TO DECISIONS, FIFTH EDITION

by

Stuart A. Klugman, Harry H. Panjer and Gordon E. Willmot

The Education Committee provides study notes to persons preparing for the examinations of the Society of Actuaries. They are intended to acquaint candidates with some of the theoretical and practical considerations involved in the various subjects. While varying opinions are presented where appropriate, limits on the length of the material and other considerations sometimes prevent the inclusion of all possible opinions. These study notes do not, however, represent any official opinion, interpretations or endorsement of the Society of Actuaries or its Education Committee. The Society is grateful to the authors for their contributions in preparing the study notes.

LTAM-22-18 **Printed in U.S.A.**

CONTENTS

10

REVIEW OF MATHEMATICAL STATISTICS

10.1 Introduction and four data sets

Before studying empirical models and then parametric models, we review some concepts from mathematical statistics. Mathematical statistics is a broad subject that includes many topics not covered in this chapter. For those topics that are covered, it is assumed that the reader has had some prior exposure. The topics of greatest importance for constructing actuarial models are estimation and hypothesis testing. Because the Bayesian approach to statistical inference is often either ignored or treated lightly in introductory mathematical statistics texts and courses, it receives more in-depth coverage in this text in Chapter **??**. Bayesian methodology also provides the basis for the credibility methods covered in Chapter **??**.

To see the need for methods of statistical inference, consider the case where your supervisor needs a model for basic dental payments. One option is to simply announce the model. You proclaim that it is the lognormal distribution with $\mu = 5.1239$ and $\sigma = 1.0345$. (The many decimal places are designed to give your proclamation an aura of precision.) When your supervisor, a regulator, or an attorney who has put you on the witness stand asks you how you know that to be so, it will likely not be sufficient to answer that "I just know these things," "trust me, I am a trained statistician," "it is too complicated, you wouldn't understand," or "my friend at Gamma Dental uses that model."

Loss Models: From Data to Decisions, 5th edition.
By Stuart A. Klugman, Harry H. Panjer, Gordon E. Willmot

Table 10.1 Data Set A.

Number of accidents	Number of drivers
0	81,714
1	11,306
2	1,618
3	250
4	40
5 or more	7

Table 10.2 Data Set B.

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1,193	1,340	1,884	2,558	15,743

An alternative is to collect some data and use it to formulate a model. Most distributional models have two components. The first is a name, such as "Pareto." The second is the set of parameter values that complete the specification. Matters would be simpler if modeling could be done in that order. Most of the time we need to fix the parameters that go with a named model before we can decide if we want to use that model.

Because the parameter estimates are based on a sample from the population and not the entire population, the results will not be the true values. It is important to have an idea of the potential error. One way to express this error is with an interval estimate. That is, rather than announcing a particular value, a range of plausible values is presented.

When named parametric distributions are used, the parameterizations used are those from Appendixes **??** and **??**.

Alternatively, you may want to construct a nonparametric model (also called an empirical model) where the goal is to determine a model that essentially reproduces the data. Such models are discussed in Chapter 12

At this point we present four data sets, referred to as Data Sets A, B, C, and D. They will be used several times, some in this chapter and some in later chapters.

Data Set A This data set is well-known in the casualty actuarial literature. It was first analyzed in the paper [5] by Dropkin in 1959. He collected data from 1956–1958 on the number of accidents by one driver in one year. The results for 94,935 drivers are in Table 10.1.

Data Set B These numbers are artificial. They represent the amounts paid on workers compensation medical benefits but are not related to any particular policy or set of policyholders. These payments are the full amount of the loss. A random sample of 20 payments is given in Table 10.2.

Data Set C These observations represent payments on 227 claims from a general liability insurance policy. The data are in Table 10.3.

Data Set D This data set is from the experience of five-year term insurance policies. The study period is a fixed time period. The columns are interpreted as follows: (1) i is the policy number, 1–40. (2) d_i is the duration at which the insured was first observed. Thus, policies 1–30 were observed from when the policy was sold. The remining policies were

Table 10.3 Data Set C.

Payment range	Number of payments
0–7,500	99
7,500–17,500	42
17,500–32,500	29
32,500–67,500	28
67,500–125,000	17
125,000–300,000	9
Over 300,000	3

Table 10.4 Data Set D

i	d_i	x_i	u_i	i	d_i	x_i	u_i
1	0	–	0.1	16	0	4.8	–
2	0	–	0.5	17	0	–	4.8
3	0	–	0.8	18	0	–	4.8
4	0	0.8	–	19–30	0	–	5.0
5	0	–	1.8	31	0.3	–	5.0
6	0	–	1.8	32	0.7	–	5.0
7	0	–	2.1	33	1.0	4.1	–
8	0	–	2.5	34	1.8	3.1	–
9	0	–	2.8	35	2.1	–	3.9
10	0	2.9	–	36	2.9	–	5.0
11	0	2.9	–	37	2.9	–	4.8
12	0	–	3.9	38	3.2	4.0	–
13	0	4.0	–	39	3.4	–	5.0
14	0	–	4.0	40	3.9	–	5.0
15	0	–	4.1				

issued prior to the start of the observation period and were known to be alive at that duration. (3) x_i is the duration at which the insured was observed to die. Those who didn't die had "–" in that column. (4) u_i is the last duration at which those who didn't die were observed. That could be because they surrendered their policy before the five years elapsed, reached the end of the five-year term, or the study ended while the policy was still in force. The data are in Table 10.4.

10.2 Point estimation

10.2.1 Introduction

Regardless of how a model is estimated, it is extremely unlikely that the estimated model will exactly match the true distribution. Ideally, we would like to be able to measure the error we will be making when using the estimated model. But doing so is clearly impossible! If we knew the amount of error we had made, we could adjust our estimate

by that amount and then have no error at all. The best we can do is discover how much error is inherent in repeated use of the *procedure*, as opposed to how much error we made with our current estimate. Therefore, we are concerned about the quality of the ensemble of answers produced from the procedure, not about the quality of a particular answer.

This is a critical point with regard to actuarial practice. What is important is that an appropriate procedure be used, with everyone understanding that even the best procedure can lead to a poor result once the random future outcome has been revealed. This point is stated nicely in a Society of Actuaries principles draft [19, pp. 779–780] regarding the level of adequacy of a provision for a portfolio of life insurance risk obligations (i.e., the probability that the company will have enough money to meet its contractual obligations):

> The indicated level of adequacy is prospective, but the actuarial model is generally validated against past experience. It is incorrect to conclude on the basis of subsequent experience that the actuarial assumptions were inappropriate or that the indicated level of adequacy was overstated or understated.

When constructing models, there are several types of error. Some, such as model error (choosing the wrong model) and sampling frame error (trying to draw inferences about a population that differs from the one sampled), are not covered here. An example of model error is selecting a Pareto distribution when the true distribution is, or is close to, Weibull. An example of sampling frame error is sampling claims from insurance policies that were sold by independent agents to price policies that are to be sold over the internet.

The type of error that we can measure is that resulting from using a sample from the population to make inferences about the entire population. Errors occur when the items sampled do not represent the population. As noted earlier, we cannot know if the particular items sampled today do or do not represent the population. We can, however, estimate the extent to which estimators are affected by the possibility of a nonrepresentative sample.

The approach taken in this section is to consider all the samples that might be taken from the population. Each such sample leads to an estimated quantity (e.g., a probability, a parameter value, or a moment). We do not expect the estimated quantities to always match the true value. For a sensible estimation procedure, we do expect that for some samples the quantity will match the true value, for many it will be close, and for only a few it will be quite different. If we can construct a measure of how well the set of potential estimates matches the true value, we have a handle on the quality of our estimation procedure. The approach outlined here is often called the *classical* or *frequentist* approach to estimation.

Finally, we need a word about the difference between *estimate* and *estimator*. The former refers to the specific value obtained when applying an estimation procedure to a set of numbers. The latter refers to a rule or formula that produces the estimate. An estimate is a number or function, while an estimator is a random variable or a random function. Usually, both the words and the context will make clear which is being referred to.

10.2.2 Measures of quality

10.2.2.1 Introduction There are a variety of ways to measure the quality of an estimator. Three of them are discussed here. Two examples are used throughout to illustrate them.

■ **EXAMPLE 10.1**

A population contains the values 1, 3, 5, and 9. We want to estimate the population mean by taking a sample of size 2 with replacement. □

■ **EXAMPLE 10.2**

A population has the exponential distribution with a mean of θ. We want to estimate the population mean by taking a sample of size 3 with replacement. □

Both examples are clearly artificial in that we know the answers prior to sampling (4.5 and θ). However, that knowledge will make apparent the error in the procedure we select. For practical applications, we need to be able to estimate the error when we do not know the true value of the quantity being estimated.

10.2.2.2 Unbiasedness When constructing an estimator, it would be good if, on average, the errors we make cancel each other out. More formally, let θ be the quantity we want to estimate. Let $\hat{\theta}$ be the random variable that represents the estimator and let $\mathrm{E}(\hat{\theta}|\theta)$ be the expected value of the estimator $\hat{\theta}$ when θ is the true parameter value.

Definition 10.1 *An estimator, $\hat{\theta}$, is **unbiased** if $\mathrm{E}(\hat{\theta}|\theta) = \theta$ for all θ. The **bias** is $\mathrm{bias}_{\hat{\theta}}(\theta) = \mathrm{E}(\hat{\theta}|\theta) - \theta$.*

The bias depends on the estimator being used and may also depend on the particular value of θ.

■ **EXAMPLE 10.3**

For Example 10.1 determine the bias of the sample mean as an estimator of the population mean.

The population mean is $\theta = 4.5$. The sample mean is the average of the two observations. In all cases, we assume that sampling is random. In other words, every sample of size n has the same chance of being drawn. Such sampling also implies that any member of the population has the same chance of being observed as any other member. For this example, there are 16 equally likely ways the sample could have turned out:

1,1	1,3	1,5	1,9	3,1	3,3	3,5	3,9
5,1	5,3	5,5	5,9	9,1	9,3	9,5	9,9

These samples lead to the following 16 equally likely values for the sample mean:

1	2	3	5	2	3	4	6
3	4	5	7	5	6	7	9

Combining the common values, the sample mean, usually denoted $\bar{X}$, has the following probability distribution:

x	1	2	3	4	5	6	7	9
$p_{\bar{X}}(x)$	1/16	2/16	3/16	2/16	3/16	2/16	2/16	1/16

The expected value of the estimator is

$$\mathrm{E}(\bar{X}) = [1(1) + 2(2) + 3(3) + 4(2) + 5(3) + 6(2) + 7(2) + 9(1)]/16 = 4.5,$$

and so the sample mean is an unbiased estimator of the population mean for this example. □

EXAMPLE 10.4

For Example 10.2 determine the bias of the sample mean and the sample median as estimators of the population mean.

The sample mean is $\bar{X} = (X_1 + X_2 + X_3)/3$, where each X_j represents one of the observations from the exponential population. Its expected value is

$$\begin{aligned}\mathrm{E}(\bar{X}) &= \mathrm{E}\left[\frac{1}{3}(X_1 + X_2 + X_3)\right] = \tfrac{1}{3}\left[\mathrm{E}(X_1) + \mathrm{E}(X_2) + \mathrm{E}(X_3)\right] \\ &= \tfrac{1}{3}(\theta + \theta + \theta) = \theta\end{aligned}$$

and, therefore, the sample mean is an unbiased estimator of the population mean.

Investigating the sample median is a bit more difficult. The distribution function of the middle of three observations can be found as follows, using Y as the random variable of interest and X_j as the random variable for the jth observation from the population:

$$\begin{aligned}F_Y(y) = {} & \Pr(Y \le y) = \Pr(X_1, X_2, X_3 \le y) + \Pr(X_1, X_2 \le y, X_3 > y) \\ & + \Pr(X_1, X_3 \le y, X_2 > y) + \Pr(X_2, X_3 \le y, X_1 > y) \\ = {} & F_X(y)^3 + 3F_X(y)^2[1 - F_X(y)] \\ = {} & [1 - e^{-y/\theta}]^3 + 3[1 - e^{-y/\theta}]^2 e^{-y/\theta}.\end{aligned}$$

The first two lines follow because for the median to be less than or equal to y, either all three observations or exactly two of them must be less than or equal to y. The density function is

$$f_Y(y) = F_Y'(y) = \frac{6}{\theta}\left(e^{-2y/\theta} - e^{-3y/\theta}\right).$$

The expected value of this estimator is

$$\begin{aligned}\mathrm{E}(Y|\theta) &= \int_0^\infty y\frac{6}{\theta}\left(e^{-2y/\theta} - e^{-3y/\theta}\right)dy \\ &= \frac{5\theta}{6}.\end{aligned}$$

This estimator is clearly biased,[1] with

$$\mathrm{bias}_Y(\theta) = 5\theta/6 - \theta = -\theta/6.$$

On average, this estimator underestimates the true value. It is also easy to see that the sample median can be turned into an unbiased estimator by multiplying it by 1.2. □

[1]The sample median is unlikely to be selected as an estimator of the population mean. This example studies it for comparison purposes. Because the population median is $\theta \ln 2$, the sample median is also biased for the population median.

For Example 10.2 we have two estimators (the sample mean and 1.2 times the sample median) that are both unbiased. We will need additional criteria to decide which one we prefer.

Some estimators exhibit a small amount of bias, which vanishes as the sample size goes to infinity.

Definition 10.2 *Let $\hat{\theta}_n$ be an estimator of θ based on a sample size of n. The estimator is* ***asymptotically unbiased*** *if*

$$\lim_{n\to\infty} \mathrm{E}(\hat{\theta}_n|\theta) = \theta$$

for all θ.

■ **EXAMPLE 10.5**

Suppose a random variable has the uniform distribution on the interval $(0, \theta)$. Consider the estimator $\hat{\theta}_n = \max(X_1, \ldots, X_n)$. Show that this estimator is asymptotically unbiased.

Let Y_n be the maximum from a sample of size n. Then

$$\begin{aligned} F_{Y_n}(y) &= \Pr(Y_n \le y) = \Pr(X_1 \le y, \ldots, X_n \le y) \\ &= [F_X(y)]^n \\ &= (y/\theta)^n, \\ f_{Y_n}(y) &= \frac{ny^{n-1}}{\theta^n},\ 0 < y < \theta. \end{aligned}$$

The expected value is

$$\mathrm{E}(Y_n|\theta) = \int_0^\theta y(ny^{n-1}\theta^{-n})dy = \frac{n}{n+1}y^{n+1}\theta^{-n}\Big|_0^\theta = \frac{n\theta}{n+1}.$$

As $n \to \infty$, the limit is θ, showing that this estimator asymptotically unbiased. □

A drawback to unbiasedness as a measure of the quality of an estimator is that an unbiased estimator may often not be very close to the parameter, as would be the case if the estimator has a large variance. We will now demonstrate that there is a limit to the accuracy of an unbiased estimator in general, in the sense that there is a lower bound (called the Cramèr-Rao lower bound) on its variance.

In what follows, suppose that $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ has joint pf or pdf $g(\mathbf{x}; \theta)$ where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. In the i.i.d. special case $g(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta)$ where $f(x; \theta)$ is the common pf or pdf of the X_is. Of central importance in many discussions of parameter estimation is the **score function**, $U = \partial \ln g(\mathbf{X}; \theta)/\partial\theta$. We assume regularity conditions on g that will be discussed later in detail, but at this point we assume that g is twice differentiable with respect to θ and the order of differentiation and expectation may be interchanged. In particular, this excludes situations where an end point of the distribution depends on θ.

■ **EXAMPLE 10.6**

Determine the score function for the i.i.d. exponential case.

Let $f(x;\theta) = \theta^{-1}e^{-x/\theta}$. Then,

$$g(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \theta^{-n}\exp\left(-\sum_{i=1}^{n} x_i/\theta\right),$$

and thus,

$$U = \frac{\partial}{\partial\theta}\left(-n\ln\theta - \frac{1}{\theta}\sum_{i=1}^{n} X_i\right) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} X_i.$$

□

As is clear from the above example, U is a random function of θ (i.e., U is a random variable and a function of θ).

In the i.i.d. special case, let $W_i = \partial \ln f(x_i;\theta)/\partial\theta$ for $i = 1, 2, \ldots, n$, implying that $W_1, W_2, \ldots, W_n$ are i.i.d. Then $U = \sum_{i=1}^{n} W_i$.

We now turn to the evaluation of the mean of the score function. In the discrete case (the continuous case is similar),

$$\begin{aligned}
\mathrm{E}(U) &= \sum_{\text{all } \mathbf{x}} \left[\frac{\partial}{\partial\theta}\ln g(\mathbf{x};\theta)\right] g(\mathbf{x};\theta) \\
&= \sum_{\text{all } \mathbf{x}} \frac{\frac{\partial}{\partial\theta}g(\mathbf{x};\theta)}{g(\mathbf{x};\theta)} g(\mathbf{x};\theta) \\
&= \sum_{\text{all } \mathbf{x}} \frac{\partial}{\partial\theta} g(\mathbf{x};\theta) \\
&= \frac{\partial}{\partial\theta}\sum_{\text{all } \mathbf{x}} g(\mathbf{x};\theta) \\
&= \frac{\partial}{\partial\theta} 1 = 0.
\end{aligned}$$

The last step follows because the sum of the probabilities over all possible values must be one.

Also,

$$\frac{\partial^2}{\partial\theta^2}\ln g(\mathbf{x};\theta) = \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\ln g(\mathbf{x};\theta)\right] = \frac{\partial}{\partial\theta}\left[\frac{\frac{\partial}{\partial\theta}g(\mathbf{x};\theta)}{g(\mathbf{x};\theta)}\right],$$

and so, by the quotient rule for differentiation,

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\ln g(\mathbf{x};\theta) &= \frac{\frac{\partial^2}{\partial\theta^2}g(\mathbf{x};\theta)}{g(\mathbf{x};\theta)} - \left[\frac{\frac{\partial}{\partial\theta}g(\mathbf{x};\theta)}{g(\mathbf{x};\theta)}\right]^2 \\
&= \frac{\frac{\partial^2}{\partial\theta^2}g(\mathbf{x};\theta)}{g(\mathbf{x};\theta)} - \left[\frac{\partial}{\partial\theta}\ln g(\mathbf{x};\theta)\right]^2.
\end{aligned}$$

Taking expectations yields,

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln g(\mathbf{X};\theta)\right] &= \mathrm{E}\left[\frac{\frac{\partial^2}{\partial\theta^2}g(\mathbf{X};\theta)}{g(\mathbf{X};\theta)}\right] - \mathrm{E}(U^2) \\
&= \sum_{\text{all } \mathbf{x}} \frac{\frac{\partial^2}{\partial\theta^2}g(\mathbf{x};\theta)}{g(\mathbf{x};\theta)} g(\mathbf{x};\theta) - \mathrm{E}(U^2) \\
&= \sum_{\text{all } \mathbf{x}} \frac{\partial^2}{\partial\theta^2} g(\mathbf{x};\theta) - \mathrm{E}(U^2) \\
&= \frac{\partial^2}{\partial\theta^2}\left[\sum_{\text{all } \mathbf{x}} g(\mathbf{x};\theta)\right] - \mathrm{E}(U^2).
\end{aligned}$$

The first term is on the right hand side is zero and therefore

$$\mathrm{E}(U^2) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln g(\mathbf{X};\theta)\right].$$

Alternatively, using the definition of U we have

$$\mathrm{E}(U^2) = \mathrm{E}\left\{\left[\frac{\partial}{\partial\theta}\ln g(\mathbf{X};\theta)\right]^2\right\}.$$

Recall that $\mathrm{E}(U) = 0$. Then,

$$\mathrm{Var}(U) = \mathrm{E}(U^2) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln g(\mathbf{X};\theta)\right] = \mathrm{E}\left\{\left[\frac{\partial}{\partial\theta}\ln g(\mathbf{X};\theta)\right]^2\right\}.$$

Before proceeding, we digress to note that, for any two random variables Z_1 and Z_2,

$$\mathrm{Cov}(Z_1, Z_2) \leq \sqrt{\mathrm{Var}(Z_1)}\sqrt{\mathrm{Var}(Z_2)}.$$

To see that this is true, let $\sigma_1^2 = \mathrm{Var}(Z_1)$, $\sigma_2^2 = \mathrm{Var}(Z_2)$, $\sigma_{12} = \mathrm{Cov}(Z_1, Z_2)$, and $\rho = \sigma_{12}/(\sigma_1\sigma_2)$. Then,

$$\begin{aligned}
0 &\leq \mathrm{Var}\left(\frac{Z_1}{\sigma_1} - \rho\frac{Z_2}{\sigma_2}\right) \\
&= \frac{1}{\sigma_1^2}\mathrm{Var}(Z_1) + \left(\frac{\rho}{\sigma_2}\right)^2 \mathrm{Var}(Z_2) - \frac{2\rho}{\sigma_1\sigma_2}\mathrm{Cov}(Z_1, Z_2) \\
&= 1 + \rho^2 - 2\rho^2 = 1 - \rho^2.
\end{aligned}$$

Note that this development also proves that $-1 \leq \rho \leq 1$.

Now let $T = T(\mathbf{X})$ be an unbiased estimator of θ. Then, by the definition of unbiasedness,

$$\theta = \mathrm{E}(T) = \sum_{\text{all } \mathbf{x}} T(\mathbf{x}) g(\mathbf{x};\theta)$$

and differentiating with respect to θ yields (recalling that we are assuming that the order of differentiation and summation/integration may be interchanged)

$$\begin{aligned} 1 &= \sum_{\text{all } \mathbf{x}} T(\mathbf{x}) \frac{\partial}{\partial \theta} g(\mathbf{x};\theta) \\ &= \sum_{\text{all } \mathbf{x}} T(\mathbf{x}) \left[\frac{\partial}{\partial \theta} \ln g(\mathbf{x};\theta) \right] g(\mathbf{x};\theta) \\ &= \mathrm{E}(TU). \end{aligned}$$

Then,

$$\mathrm{Cov}(T,U) = \mathrm{E}(TU) - \mathrm{E}(T)\mathrm{E}(U) = 1 - (\theta)(0) = 1.$$

We next have

$$1 = \mathrm{Cov}(T,U) \le \sqrt{\mathrm{Var}(T)}\sqrt{\mathrm{Var}(U)}.$$

This implies that,

$$\mathrm{Var}(T) \ge \frac{1}{\mathrm{Var}(U)} = \frac{1}{-\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} \ln g(\mathbf{X};\theta)\right]} = \frac{1}{\mathrm{E}\left\{\left[\frac{\partial}{\partial \theta} \ln g(\mathbf{X};\theta)\right]^2\right\}} \tag{10.1}$$

In the i.i.d. case $\mathrm{Var}(U) = \sum_{i=1}^{n} \mathrm{Var}(W_i) = n\mathrm{Var}(W)$ where $W = \frac{\partial}{\partial \theta} \ln f(X;\theta)$ and X is a generic version of the X_is. Then (10.1) becomes

$$\mathrm{Var}(T) \ge \frac{1}{\mathrm{Var}(U)} = \frac{1}{-n\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} \ln f(X;\theta)\right]} = \frac{1}{n\mathrm{E}\left\{\left[\frac{\partial}{\partial \theta} \ln f(X;\theta)\right]^2\right\}} \tag{10.2}$$

Generally, the version using second partial derivatives (rather than the square of the first derivative) is easier to calculate.

The lower bounds (10.1) and (10.2) are often referred to as Cramèr-Rao lower bounds for the variance of unbiased estimators. This is extremely valuable for maximum likelihood and other estimation procedures. The denominators in each case are referred to as the *Fisher* or *expected* information.

EXAMPLE 10.7

Determine the Cramèr-Rao lower bound for an unbiased estimator of the sample mean from a population with an exponential distribution. Use both formulas from (10.2).

For the exponential distribution,

$$\begin{aligned}
f(X;\theta) &= \theta^{-1}e^{-X/\theta}\\
\ln f(X:\theta) &= -\ln\theta - X/\theta\\
\frac{\partial}{\partial\theta}\ln f(X:\theta) &= -\theta^{-1} + X\theta^{-2}\\
\frac{\partial^2}{\partial\theta^2}\ln f(X:\theta) &= \theta^{-2} - 2X\theta^{-3}\\
\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right] &= \mathrm{E}\left(\theta^{-2} - 2X\theta^{-3}\right) = \theta^{-2} - 2\theta^{-2} = -\theta^{-2}\\
\mathrm{E}\left\{\left[\frac{\partial}{\partial\theta}\ln f(X;\theta)\right]^2\right\} &= \mathrm{E}\left[\left(-\theta^{-1} + X\theta^{-2}\right)^2\right] = \mathrm{E}\left(\theta^{-2} - 2X\theta^{-3} + X^2\theta^{-4}\right)\\
&= \theta^{-2} - 2\theta^{-2} + 2\theta^{-2} = \theta^{-2}.
\end{aligned}$$

The lower bound is then $1/(n\theta^{-2}) = \theta^2/n$. □

10.2.2.3 Consistency Another desirable property of an estimator is that it works well for extremely large samples. Slightly more formally, as the sample size goes to infinity, the probability that the estimator is in error by more than a small amount goes to zero. A formal definition follows.

Definition 10.3 *An estimator is* ***consistent*** *(often called, in this context,* ***weakly consistent****) if, for all $\delta > 0$ and any θ,*

$$\lim_{n\to\infty}\Pr(|\hat{\theta}_n - \theta| > \delta) = 0.$$

A sufficient (although not necessary) condition for weak consistency is that the estimator be asymptotically unbiased and $\mathrm{Var}(\hat{\theta}_n) \to 0$ (equivalently, from (10.3), the mean-squared error goes to zero as $n \to \infty$).

■ **EXAMPLE 10.8**

Prove that, if the variance of a random variable is finite, the sample mean is a consistent estimator of the population mean.

From Exercise 10.2, the sample mean is unbiased. In addition,

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)\\
&= \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}(X_j)\\
&= \frac{\mathrm{Var}(X)}{n} \to 0.
\end{aligned}$$

The second step follows from assuming that the observations are independent. □

■ **EXAMPLE 10.9**

Show that the maximum observation from a uniform distribution on the interval $(0, \theta)$ is a consistent estimator of θ.

From Example 10.5, the maximum is asymptotically unbiased. The second moment is

$$\mathrm{E}(Y_n^2) = \int_0^\theta y^2(ny^{n-1}\theta^{-n})dy = \frac{n}{n+2}y^{n+2}\theta^{-n}\Big|_0^\theta = \frac{n\theta^2}{n+2},$$

and then

$$\mathrm{Var}(Y_n) = \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 = \frac{n\theta^2}{(n+2)(n+1)^2} \to 0.$$

□

10.2.2.4 Mean-squared error While consistency is nice, most estimators have this property. What would be truly impressive is an estimator that is not only correct on average but comes very close most of the time and, in particular, comes closer than rival estimators. One measure for a finite sample is motivated by the definition of consistency. The quality of an estimator could be measured by the probability that it gets within δ of the true value—that is, by measuring $\Pr(|\hat{\theta}_n - \theta| < \delta)$. But the choice of δ is arbitrary, and we prefer measures that cannot be altered to suit the investigator's whim. Then we might consider $\mathrm{E}(|\hat{\theta}_n - \theta|)$, the average absolute error. But we know that working with absolute values often presents unpleasant mathematical challenges, and so the following has become widely accepted as a measure of accuracy.

Definition 10.4 *The* ***mean-squared error*** **(*MSE*)** *of an estimator is*

$$\mathrm{MSE}_{\hat{\theta}}(\theta) = E[(\hat{\theta} - \theta)^2|\theta].$$

Note that the MSE is a function of the true value of the parameter. An estimator may perform extremely well for some values of the parameter but poorly for others.

■ **EXAMPLE 10.10**

Consider the estimator $\hat{\theta} = 5$ of an unknown parameter θ. The MSE is $(5 - \theta)^2$, which is very small when θ is near 5 but becomes poor for other values. Of course this estimate is both biased and inconsistent unless θ is exactly equal to 5. □

A result that follows directly from the various definitions is

$$\mathrm{MSE}_{\hat{\theta}}(\theta) = \mathrm{E}\{[\hat{\theta} - \mathrm{E}(\hat{\theta}|\theta) + \mathrm{E}(\hat{\theta}|\theta) - \theta]^2|\theta\} = \mathrm{Var}(\hat{\theta}|\theta) + [\mathrm{bias}_{\hat{\theta}}(\theta)]^2. \qquad (10.3)$$

If we restrict attention to only unbiased estimators, the best such estimator could be defined as follows.

Definition 10.5 *An estimator $\hat{\theta}$ is called a* ***uniformly minimum variance unbiased estimator*** **(*UMVUE*)** *if it is unbiased and for any true value of θ there is no other unbiased estimator that has a smaller variance.*

Because we are looking only at unbiased estimators, it would have been equally effective to make the definition in terms of MSE. We could also generalize the definition by looking for estimators that are uniformly best with regard to MSE, but the previous example indicates why that is not feasible. There are some results that can often assist with the determination of UMVUEs (e.g., [9, ch. 7]). However, such estimators are often difficult to determine. Nevertheless, MSE is still a useful criterion for comparing two alternative estimators.

EXAMPLE 10.11

For Example 10.2 compare the MSEs of the sample mean and 1.2 times the sample median. Demonstrate that for the exponential distribution the sample mean a UMVUE.

The sample mean has variance

$$\frac{\mathrm{Var}(X)}{3} = \frac{\theta^2}{3}.$$

When multiplied by 1.2, the sample median has second moment

$$\begin{aligned}
\mathrm{E}[(1.2Y)^2] &= 1.44\int_0^\infty y^2\frac{6}{\theta}\left(e^{-2y/\theta} - e^{-3y/\theta}\right)dy \\
&= 1.44\frac{6}{\theta}\left[y^2\left(\frac{-\theta}{2}e^{-2y/\theta} + \frac{\theta}{3}e^{-3y/\theta}\right)\right. \\
&\qquad - 2y\left(\frac{\theta^2}{4}e^{-2y/\theta} - \frac{\theta^2}{9}e^{-3y/\theta}\right) \\
&\qquad \left. +2\left(\frac{-\theta^3}{8}e^{-2y/\theta} + \frac{\theta^3}{27}e^{-3y/\theta}\right)\right]\Bigg|_0^\infty \\
&= \frac{8.64}{\theta}\left(\frac{2\theta^3}{8} - \frac{2\theta^3}{27}\right) = \frac{38\theta^2}{25}
\end{aligned}$$

for a variance of

$$\frac{38\theta^2}{25} - \theta^2 = \frac{13\theta^2}{25} > \frac{\theta^2}{3}.$$

The sample mean has the smaller MSE regardless of the true value of θ. Therefore, for this problem, it is a superior estimator of θ.

From Example 10.7 we see that the minimum possible variance for an unbiased estimator is, for a sample of size 3, $\theta^2/3$. This matches the variance of the sample mean and therefore no other unbiased estimator can have a smaller variance. Hence the sample mean is a UMVUE. □

EXAMPLE 10.12

For the uniform distribution on the interval $(0, \theta)$ compare the MSE of the estimators $2\bar{X}$ and $[(n+1)/n]\max(X_1, \ldots, X_n)$. Also evaluate the MSE of $\max(X_1, \ldots, X_n)$.

The first two estimators are unbiased, so it is sufficient to compare their variances. For twice the sample mean,

$$\mathrm{Var}(2\bar{X}) = \frac{4}{n}\mathrm{Var}(X) = \frac{4\theta^2}{12n} = \frac{\theta^2}{3n}.$$

For the adjusted maximum, the second moment is

$$\mathrm{E}\left[\left(\frac{n+1}{n}Y_n\right)^2\right] = \frac{(n+1)^2}{n^2}\frac{n\theta^2}{n+2} = \frac{(n+1)^2\theta^2}{(n+2)n}$$

for a variance of

$$\frac{(n+1)^2\theta^2}{(n+2)n} - \theta^2 = \frac{\theta^2}{n(n+2)}.$$

Except for the case $n = 1$ (and then the two estimators are identical), the one based on the maximum has the smaller MSE. The third estimator is biased. For it, the MSE is

$$\frac{n\theta^2}{(n+2)(n+1)^2} + \left(\frac{n\theta}{n+1} - \theta\right)^2 = \frac{2\theta^2}{(n+1)(n+2)},$$

which is also larger than that for the adjusted maximum. □

For this example the regulatory conditions underlying the derivation of the Cramèr-Rao lower bound do not hold and so (10.2) cannot be used to set a minimum possible value.

10.2.3 Exercises

10.1 For Example 10.1, show that the mean of three observations drawn without replacement is an unbiased estimator of the population mean, while the median of three observations drawn without replacement is a biased estimator of the population mean.

10.2 Prove that for random samples the sample mean is always an unbiased estimator of the population mean.

10.3 Let X have the uniform distribution over the range $(\theta - 2, \theta + 2)$. That is, $f_X(x) = 0.25$, $\theta - 2 < x < \theta + 2$. Show that the median from a sample of size 3 is an unbiased estimator of θ.

10.4 Explain why the sample mean may not be a consistent estimator of the population mean for a Pareto distribution.

10.5 For the sample of size 3 in Exercise 10.3, compare the MSE of the sample mean and median as estimates of θ.

10.6 (*) You are given two independent estimators of an unknown quantity θ. For estimator A, $\mathrm{E}(\hat{\theta}_A) = 1{,}000$ and $\mathrm{Var}(\hat{\theta}_A) = 160{,}000$, while for estimator B, $\mathrm{E}(\hat{\theta}_B) = 1{,}200$ and $\mathrm{Var}(\hat{\theta}_B) = 40{,}000$. Estimator C is a weighted average, $\hat{\theta}_C = w\hat{\theta}_A + (1-w)\hat{\theta}_B$. Determine the value of w that minimizes $\mathrm{Var}(\hat{\theta}_C)$.

10.7 (*) A population of losses has the Pareto distribution (see Appendix **??**) with $\theta = 6{,}000$ and α unknown. Simulation of the results from maximum likelihood estimation based on samples of size 10 has indicated that $\mathrm{E}(\hat{\alpha}) = 2.2$ and $\mathrm{MSE}(\hat{\alpha}) = 1$. Determine $\mathrm{Var}(\hat{\alpha})$ if it is known that $\alpha = 2$.

10.8 (*) Two instruments are available for measuring a particular nonzero distance. The random variable X represents a measurement with the first instrument and the random

variable Y with the second instrument. Assume X and Y are independent with $\mathrm{E}(X) = 0.8m$, $\mathrm{E}(Y) = m$, $\mathrm{Var}(X) = m^2$, and $\mathrm{Var}(Y) = 1.5m^2$, where m is the true distance. Consider estimators of m that are of the form $Z = \alpha X + \beta Y$. Determine the values of α and β that make Z a UMVUE within the class of estimators of this form.

10.9 A population contains six members, with values 1, 1, 2, 3, 5, and 10. A random sample of size 3 is drawn without replacement. In each case the objective is to estimate the population mean. *Note:* A spreadsheet with an optimization routine may be the best way to solve this problem.

(a) Determine the bias, variance, and MSE of the sample mean.

(b) Determine the bias, variance, and MSE of the sample median.

(c) Determine the bias, variance, and MSE of the sample midrange (the average of the largest and smallest observations).

(d) Consider an arbitrary estimator of the form $aX_{(1)} + bX_{(2)} + cX_{(3)}$, where $X_{(1)} \le X_{(2)} \le X_{(3)}$ are the sample order statistics.

i. Determine a restriction on the values of a, b, and c that will assure that the estimator is unbiased.

ii. Determine the values of a, b, and c that will produce the unbiased estimator with the smallest variance.

iii. Determine the values of a, b, and c that will produce the (possibly biased) estimator with the smallest MSE.

10.10 (*) Two different estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, are being considered. To test their performance, 75 trials have been simulated, each with the true value set at $\theta = 2$. The following totals were obtained:

$$\sum_{j=1}^{75} \hat{\theta}_{1j} = 165, \sum_{j=1}^{75} \hat{\theta}_{1j}^2 = 375, \sum_{j=1}^{75} \hat{\theta}_{2j} = 147, \sum_{j=1}^{75} \hat{\theta}_{2j}^2 = 312,$$

where $\hat{\theta}_{ij}$ is the estimate based on the jth simulation using estimator $\hat{\theta}_i$. Estimate the MSE for each estimator and determine the **relative efficiency** (the ratio of the MSEs).

10.3 Interval estimation

All of the estimators discussed to this point have been **point estimators**. That is, the estimation process produces a single value that represents our best attempt to determine the value of the unknown population quantity. While that value may be a good one, we do not expect it to exactly match the true value. A more useful statement is often provided by an **interval estimator**. Instead of a single value, the result of the estimation process is a range of possible numbers, any of which is likely to be the true value. A specific type of interval estimator is the confidence interval.

Definition 10.6 *A* $100(1-\alpha)\%$ ***confidence interval*** *for a parameter* θ *is a pair of random values,* L *and* U*, computed from a random sample such that* $\Pr(L \le \theta \le U) \ge 1 - \alpha$ *for all* θ*.*

Note that this definition does not uniquely specify the interval. Because the definition is a probability statement and must hold for all θ, it says nothing about whether or not a particular interval encloses the true value of θ from a particular population. Instead, the **level of confidence**, $1-\alpha$, is a property of the method used to obtain L and U and not of the particular values obtained. The proper interpretation is that, if we use a particular interval estimator over and over on a variety of samples, at least $100(1-\alpha)\%$ of the time our interval will enclose the true value. Keep in mind that it is the interval end points that are random.

Constructing confidence intervals is usually very difficult. For example, we know that, if a population has a normal distribution with unknown mean and variance, a $100(1-\alpha)\%$ confidence interval for the mean uses

$$L = \bar{X} - t_{\alpha/2,n-1}s/\sqrt{n}, \quad U = \bar{X} + t_{\alpha/2,n-1}s/\sqrt{n}, \tag{10.4}$$

where $s = \sqrt{\sum_{j=1}^{n}(X_j - \bar{X})^2/(n-1)}$ and $t_{\alpha/2,b}$ is the $100(1-\alpha/2)$th percentile of the t distribution with b degrees of freedom. But it takes a great deal of effort to verify that (10.4) is correct (see, e.g., [9, p. 186]).

However, there is a method for constructing approximate confidence intervals that is often accessible. Suppose we have a point estimator $\hat{\theta}$ of parameter θ such that $\mathrm{E}(\hat{\theta}) \doteq \theta$, $\mathrm{Var}(\hat{\theta}) \doteq v(\theta)$, and $\hat{\theta}$ has approximately a normal distribution. Theorem 11.4 shows that these three properties are often the case. With all these approximations, we have that approximately

$$1-\alpha \doteq \Pr\left(-z_{\alpha/2} \le \frac{\hat{\theta}-\theta}{\sqrt{v(\theta)}} \le z_{\alpha/2}\right), \tag{10.5}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution. Solving for θ produces the desired interval. Sometimes obtaining the solution is difficult to do (due to the appearance of θ in the denominator), and so, if necessary, replace $v(\theta)$ in (10.5) with $v(\hat{\theta})$ to obtain a further approximation:

$$1-\alpha \doteq \Pr\left(\hat{\theta} - z_{\alpha/2}\sqrt{v(\hat{\theta})} \le \theta \le \hat{\theta} + z_{\alpha/2}\sqrt{v(\hat{\theta})}\right). \tag{10.6}$$

EXAMPLE 10.13

Use (10.6) to construct an approximate 95% confidence interval for the mean of a normal population with unknown variance.

Use $\hat{\theta} = \bar{X}$ and then note that $\mathrm{E}(\hat{\theta}) = \theta$, $\mathrm{Var}(\hat{\theta}) = \sigma^2/n$, and $\hat{\theta}$ does have a normal distribution. The confidence interval is then $\bar{X} \pm 1.96s/\sqrt{n}$. Because $t_{0.025,n-1} > 1.96$, this approximate interval must be narrower than the exact interval given by (10.4), which implies our level of confidence is something less than 95%. □

EXAMPLE 10.14

Use (10.5) and (10.6) to construct approximate 95% confidence intervals for the mean of a Poisson distribution. Obtain intervals for the particular case where $n = 25$ and $\bar{x} = 0.12$.

Here $\theta = \lambda$, the mean of the Poisson distribution. Let $\hat{\theta} = \bar{X}$, the sample mean. For the Poisson distribution, $\mathrm{E}(\hat{\theta}) = \mathrm{E}(X) = \theta$ and $v(\theta) = \mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n = \theta/n$. For the first interval,

$$0.95 \doteq \Pr\left(-1.96 \le \frac{\bar{X} - \theta}{\sqrt{\theta/n}} \le 1.96\right)$$

is true if and only if

$$|\bar{X} - \theta| \le 1.96\sqrt{\frac{\theta}{n}},$$

which is equivalent to

$$(\bar{X} - \theta)^2 \le \frac{3.8416\theta}{n}$$

or

$$\theta^2 - \theta\left(2\bar{X} + \frac{3.8416}{n}\right) + \bar{X}^2 \le 0.$$

Solving the quadratic produces the interval

$$\bar{X} + \frac{1.9208}{n} \pm \frac{1}{2}\sqrt{\frac{15.3664\bar{X} + 3.8416^2/n}{n}},$$

and for this problem the interval is 0.197 ± 0.156.

For the second approximation, the interval is $\bar{X} \pm 1.96\sqrt{\bar{X}/n}$, and for the example, it is 0.12 ± 0.136. This interval extends below zero (which is not possible for the true value of θ) because (10.6) is too crude an approximation in this case. □

10.3.1 Exercises

10.11 Let $x_1, \ldots, x_n$ be a random sample from a population with pdf $f(x) = \theta^{-1}e^{-x/\theta}$, $x > 0$. This exponential distribution has a mean of θ and a variance of θ^2. Consider the sample mean, $\bar{X}$, as an estimator of θ. It turns out that $\bar{X}/\theta$ has a gamma distribution with $\alpha = n$ and $\theta = 1/n$, where in the second expression the "θ" on the left is the parameter of the gamma distribution. For a sample of size 50 and a sample mean of 275, develop 95% confidence intervals by each of the following methods. In each case, if the formula requires the true value of θ, substitute the estimated value.

(a) Use the gamma distribution to determine an exact interval.

(b) Use a normal approximation, estimating the variance prior to solving the inequalities as in (10.6).

(c) Use a normal approximation, estimating θ after solving the inequalities as in Example 10.14.

10.12 (*) A sample of 2,000 policies had 1,600 with no claims and 400 with one or more claims. Using the normal approximation, determine the symmetric 95% confidence interval for the probability that a single policy has one or more claims.

10.4 Construction of Parametric Estimators

In previous sections we developed methods for assessing the quality of an estimator. In all the exmaples, the estimators being evaluated were arbitrary, though reasonable. This section reviews two methods for constructing estimators. A third is covered in Chapter 11. In this section, we assume that n independent observations from the same parametric distribution have been collected. There are two, essentially incompatible approaches to estimating parameters. This section and Chapter 11 cover the frequentist approach to estimation introduced in Section 10.2. An alternative estimation approach, known as Bayesian estimation, is covered in Chapter **??**.

The methods introduced in Section 10.4.1 are relatively easy to implement but tend to give poor results. Chapter 11 covers maximum likelihood estimation. This method is more difficult to use but has superior statistical properties and is considerably more flexible.

10.4.1 Method of moments and percentile matching

Let the distribution function for an individual observation be given by

$$F(x|\theta), \quad \theta^T = (\theta_1, \theta_2, \ldots, \theta_p),$$

where θ^T is the transpose of θ. That is, θ is a column vector containing the p parameters to be estimated. Furthermore, let $\mu'_k(\theta) = \mathrm{E}(X^k|\theta)$ be the kth raw moment, and let $\pi_g(\theta)$ be the $100g$th percentile of the random variable. That is, $F[\pi_g(\theta)|\theta] = g$. If the distribution function is continuous, there will be at least one solution to that equation.

For a sample of n independent observations from this random variable, let $\hat{\mu}'_k = \frac{1}{n}\sum_{j=1}^{n} x_j^k$ be the empirical estimate of the kth moment and let $\hat{\pi}_g$ be the empirical estimate of the $100g$th percentile

Definition 10.7 *A **method-of-moments estimate** of θ is any solution of the p equations*

$$\mu'_k(\theta) = \hat{\mu}'_k, \; k = 1, 2, \ldots, p.$$

The motivation for this estimator is that it produces a model that has the same first p raw moments as the data (as represented by the empirical distribution). The traditional definition of the method of moments uses positive integers for the moments. Arbitrary negative or fractional moments could also be used. In particular, when estimating parameters for inverse distributions, matching negative moments may be a superior approach.[2]

■ **EXAMPLE 10.15**

Use the method of moments to estimate parameters for the exponential, gamma, and Pareto distributions for Data Set B.

The first two sample moments are

$$\hat{\mu}'_1 = \tfrac{1}{20}(27 + \cdots + 15{,}743) = 1{,}424.4,$$
$$\hat{\mu}'_2 = \tfrac{1}{20}(27^2 + \cdots + 15{,}743^2) = 13{,}238{,}441.9.$$

[2]One advantage is that, with appropriate moments selected, the equations may have a solution within the range of allowable parameter values.

For the exponential distribution, the equation is

$$\theta = 1{,}424.4$$

with the obvious solution $\hat{\theta} = 1{,}424.4$.

For the gamma distribution, the two equations are

$$\begin{aligned} \mathrm{E}(X) &= \alpha\theta = 1{,}424.4, \\ \mathrm{E}(X^2) &= \alpha(\alpha+1)\theta^2 = 13{,}238{,}441.9. \end{aligned}$$

Dividing the second equation by the square of the first equation yields

$$\frac{\alpha+1}{\alpha} = 6.52489, \quad 1 = 5.52489\alpha,$$

and so $\hat{\alpha} = 1/5.52489 = 0.18100$ and $\hat{\theta} = 1{,}424.4/0.18100 = 7{,}869.61$.

For the Pareto distribution, the two equations are

$$\begin{aligned} \mathrm{E}(X) &= \frac{\theta}{\alpha-1} = 1{,}424.4, \\ \mathrm{E}(X^2) &= \frac{2\theta^2}{(\alpha-1)(\alpha-2)} = 13{,}238{,}441.9. \end{aligned}$$

Dividing the second equation by the square of the first equation yields

$$\frac{2(\alpha-1)}{\alpha-2} = 6.52489,$$

with a solution of $\hat{\alpha} = 2.442$ and then $\hat{\theta} = 1{,}424.4(1.442) = 2{,}053.985$. □

There is no guarantee that the equations will have a solution or, if there is a solution, that it will be unique.

Definition 10.8 *A **percentile matching estimate** of θ is any solution of the p equations*

$$\pi_{g_k}(\theta) = \hat{\pi}_{g_k},\ k = 1, 2, \ldots, p,$$

where $g_1, g_2, \ldots, g_p$ are p arbitrarily chosen percentiles. From the definition of percentile, the equations can also be written as

$$F(\hat{\pi}_{g_k}|\theta) = g_k,\ k = 1, 2, \ldots, p.$$

The motivation for this estimator is that it produces a model with p percentiles that match the data (as represented by the empirical distribution). As with the method of moments, there is no guarantee that the equations will have a solution or, if there is a solution, that it will be unique. One problem with this definition is that percentiles for discrete random variables (such as the empirical distribution) are not always well defined. For example, Data Set B has 20 observations. Any number between 384 and 457 has 10 observations below and 10 above and so could serve as the median. The convention is to use the midpoint. However, for other percentiles, there is no "official" interpolation scheme.[3] The following definition is used here.

[3]Hyndman and Fan [**?**] present nine different methods. They recommend a slight modification of the one presented here using $j = \lfloor g(n+\frac{1}{3})+\frac{1}{3} \rfloor$ and $h = g(n+\frac{1}{3})+\frac{1}{3}-j$.

Definition 10.9 *The **smoothed empirical estimate** of a percentile is calculated as*

$$\hat{\pi}_g = (1-h)x_{(j)} + hx_{(j+1)},$$

where

$$j = \lfloor (n+1)g \rfloor \text{ and } h = (n+1)g - j.$$

Here $\lfloor \cdot \rfloor$ indicates the greatest integer function and $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ are the order statistics from the sample.

Unless there are two or more data points with the same value, no two percentiles will have the same value. One feature of this definition is that $\hat{\pi}_g$ cannot be obtained for $g < 1/(n+1)$ or $g > n/(n+1)$. This seems reasonable as we should not expect to be able to infer the value of very large or small percentiles from small samples. We use the smoothed version whenever an empirical percentile estimate is needed.

■ **EXAMPLE 10.16**

Use percentile matching to estimate parameters for the exponential and Pareto distributions for Data Set B.

For the exponential distribution, select the 50th percentile. The empirical estimate is the traditional median of $\hat{\pi}_{0.5} = (384+457)/2 = 420.5$ and the equation to solve is

$$\begin{aligned} 0.5 &= F(420.5|\theta) = 1 - e^{-420.5/\theta}, \\ \ln 0.5 &= \frac{-420.5}{\theta}, \\ \hat{\theta} &= \frac{-420.5}{\ln 0.5} = 606.65. \end{aligned}$$

For the Pareto distribution, select the 30th and 80th percentiles. The smoothed empirical estimates are found as follows:

$$\begin{aligned} \text{30th:} \quad j &= \lfloor 21(0.3) \rfloor = \lfloor 6.3 \rfloor = 6, h = 6.3 - 6 = 0.3, \\ \hat{\pi}_{0.3} &= 0.7(161) + 0.3(243) = 185.6, \\ \text{80th:} \quad j &= \lfloor 21(0.8) \rfloor = \lfloor 16.8 \rfloor = 16, h = 16.8 - 16 = 0.8, \\ \hat{\pi}_{0.8} &= 0.2(1{,}193) + 0.8(1{,}340) = 1{,}310.6. \end{aligned}$$

The equations to solve are

$$0.3 = F(185.6) = 1 - \left(\frac{\theta}{185.6+\theta}\right)^{\alpha},$$

$$0.8 = F(1{,}310.6) = 1 - \left(\frac{\theta}{1{,}310.6+\theta}\right)^{\alpha},$$

$$\ln 0.7 = -0.356675 = \alpha \ln\left(\frac{\theta}{185.6+\theta}\right),$$

$$\ln 0.2 = -1.609438 = \alpha \ln\left(\frac{\theta}{1{,}310.6+\theta}\right),$$

$$\frac{-1.609438}{-0.356675} = 4.512338 = \frac{\ln\left(\frac{\theta}{1{,}310.6+\theta}\right)}{\ln\left(\frac{\theta}{185.6+\theta}\right)}.$$

Any of the methods from Appendix **??** can be used to solve this equation for $\hat{\theta} = 715.03$. Then, from the first equation,

$$0.3 = 1 - \left(\frac{715.03}{185.6+715.03}\right)^{\alpha},$$

which yields $\hat{\alpha} = 1.54559$. □

The estimates are much different from those obtained in Example 10.15, which is one indication that these methods may not be particularly reliable.

10.4.2 Exercises

10.13 Determine the method-of-moments estimate for a lognormal model for Data Set B.

10.14 (*) The 20th and 80th percentiles from a sample are 5 and 12, respectively. Using the percentile matching method, estimate $S(8)$ assuming the population has a Weibull distribution.

10.15 (*) From a sample you are given that the mean is 35,000, the standard deviation is 75,000, the median is 10,000, and the 90th percentile is 100,000. Using the percentile matching method, estimate the parameters of a Weibull distribution.

10.16 (*) A sample of size 5 produced the values 4, 5, 21, 99, and 421. You fit a Pareto distribution using the method of moments. Determine the 95th percentile of the fitted distribution.

10.17 (*) In year 1 there are 100 claims with an average size of 10,000 and in year 2 there are 200 claims with an average size of 12,500. Inflation increases the size of all claims by 10% per year. A Pareto distribution with $\alpha = 3$ and θ unknown is used to model the claim size distribution. Estimate θ for year 3 using the method of moments.

10.18 (*) From a random sample the 20th percentile is 18.25 and the 80th percentile is 35.8. Estimate the parameters of a lognormal distribution using percentile matching and then use these estimates to estimate the probability of observing a value in excess of 30.

10.19 (*) A claim process is a mixture of two random variables A and B, where A has an exponential distribution with a mean of 1 and B has an exponential distribution with a mean of 10. A weight of p is assigned to distribution A and $1-p$ to distribution B. The standard deviation of the mixture is 2. Estimate p by the method of moments.

10.20 (*) A random sample of 20 observations has been ordered as follows:

$$\begin{array}{cccccccccc} 12 & 16 & 20 & 23 & 26 & 28 & 30 & 32 & 33 & 35 \\ 36 & 38 & 39 & 40 & 41 & 43 & 45 & 47 & 50 & 57 \end{array}$$

Determine the 60th sample percentile using the smoothed empirical estimate.

10.21 (*) The following 20 wind losses (in millions of dollars) were recorded in one year:

$$\begin{array}{cccccccccc} 1 & 1 & 1 & 1 & 1 & 2 & 2 & 3 & 3 & 4 \\ 6 & 6 & 8 & 10 & 13 & 14 & 15 & 18 & 22 & 25 \end{array}$$

Determine the sample 75th percentile using the smoothed empirical estimate.

10.22 (*) The observations 1,000, 850, 750, 1,100, 1,250, and 900 were obtained as a random sample from a gamma distribution with unknown parameters α and θ. Estimate these parameters by the method of moments.

10.23 (*) A random sample of claims has been drawn from a loglogistic distribution. In the sample, 80% of the claims exceed 100 and 20% exceed 400. Estimate the loglogistic parameters by percentile matching.

10.24 (*) Let $x_1, \ldots, x_n$ be a random sample from a population with cdf $F(x) = x^p,\ 0 < x < 1$. Determine the method-of-moments estimate of p.

10.25 (*) A random sample of 10 claims obtained from a gamma distribution is given as follows:

$$\begin{array}{cccccccccc} 1{,}500 & 6{,}000 & 3{,}500 & 3{,}800 & 1{,}800 & 5{,}500 & 4{,}800 & 4{,}200 & 3{,}900 & 3{,}000 \end{array}$$

Estimate α and θ by the method of moments.

10.26 (*) A random sample of five claims from a lognormal distribution is given as follows:

$$\begin{array}{ccccc} 500 & 1{,}000 & 1{,}500 & 2{,}500 & 4{,}500 \end{array}$$

Estimate μ and σ by the method of moments. Estimate the probability that a loss will exceed 4,500.

10.27 (*) The random variable X has pdf $f(x) = \beta^{-2} x \exp(-0.5x^2/\beta^2),\ x, \beta > 0$. For this random variable, $\mathrm{E}(X) = (\beta/2)\sqrt{2\pi}$ and $\mathrm{Var}(X) = 2\beta^2 - \pi\beta^2/2$. You are given the following five observations:

$$\begin{array}{ccccc} 4.9 & 1.8 & 3.4 & 6.9 & 4.0 \end{array}$$

Determine the method-of-moments estimate of β.

Table 10.5 Data for Exercise 10.29.

No. of claims	No. of policies
0	9,048
1	905
2	45
3	2
4+	0

Table 10.6 Data for Exercise 10.30.

No. of claims	No. of policies
0	861
1	121
2	13
3	3
4	1
5	0
6	1
7+	0

10.28 The random variable X has pdf $f(x) = \alpha\lambda^{\alpha}(\lambda+x)^{-\alpha-1},\ x, \alpha, \lambda > 0$. It is known that $\lambda = 1{,}000$. You are given the following five observations:

$$43 \quad 145 \quad 233 \quad 396 \quad 775$$

Determine the method-of-moments estimate of α.

10.29 Use the data in Table 10.5 to determine the method-of-moments estimate of the parameters of the negative binomial model.

10.30 Use the data in Table 10.6 to determine the method-of-moments estimate of the parameters of the negative binomial model.

10.31 (*) Losses have a Burr distribution with $\alpha = 2$. A random sample of 15 losses is 195, 255, 270, 280, 350, 360, 365, 380, 415, 450, 490, 550, 575, 590, and 615. Use the smoothed empirical estimates of the 30th and 65th percentiles and percentile matching to estimate the parameters γ and θ.

10.32 (*) Losses have a Weibull distribution. A random sample of 16 losses is 54, 70, 75, 81, 84, 88, 97, 105, 109, 114, 122, 125, 128, 139, 146, and 153. Use the smoothed empirical estimates of the 20th and 70th percentiles and percentile matching to estimate the parameters τ and θ.

10.33 (*) Losses follow a distribution with pdf $f(x) = \theta^{-1}\exp[-(x-\delta)/\theta], x > \delta$. The sample mean is 300 and the sample median is 240. Estimate δ and θ by matching these two quantities.

10.5 Tests of hypotheses

Hypothesis testing is covered in detail in most mathematical statistics texts. This review is fairly straightforward and does not address philosophical issues or consider alternative approaches. A hypothesis test begins with two hypotheses, one called the *null* and one called the *alternative*. The traditional notation is H_0 for the null hypothesis and H_1 for the alternative hypothesis. The two hypotheses are not treated symmetrically. Reversing them may alter the results. To illustrate this process, a simple example is used.

■ **EXAMPLE 10.17**

Your company has been basing its premiums on an assumption that the average claim is 1,200. You want to raise the premium, and a regulator has insisted that you provide evidence that the average now exceeds 1,200. Let Data Set B be a sample of 20 claims. What are the hypotheses for this problem?

Let μ be the population mean. One hypothesis (the one you claim is true) is that $\mu > 1{,}200$. Because hypothesis tests must present an either/or situation, the other hypothesis must be $\mu \leq 1{,}200$. The only remaining task is to decide which of them is the null hypothesis. Whenever the universe of continuous possibilities is divided in two, there is likely to be a boundary that needs to be assigned to one hypothesis or the other. The hypothesis that includes the boundary must be the null hypothesis. Therefore, the problem can be succinctly stated as:

$$H_0 : \mu \leq 1{,}200,$$
$$H_1 : \mu > 1{,}200.$$

□

The decision is made by calculating a quantity called a *test statistic*. It is a function of the observations and is treated as a random variable. That is, in designing the test procedure, we are concerned with the samples that might have been obtained and not with the particular sample that was obtained. The test specification is completed by constructing a *rejection region*. It is a subset of the possible values of the test statistic. If the value of the test statistic for the observed sample is in the rejection region, the null hypothesis is rejected and the alternative hypothesis is announced as the result that is supported by the data. Otherwise, the null hypothesis is not rejected (more on this later). The boundaries of the rejection region (other than plus or minus infinity) are called the *critical values*.

■ **EXAMPLE 10.18**

(*Example 10.17 continued*) Complete the test using the test statistic and rejection region that is promoted in most statistics books. Assume that the population has a normal distribution with standard deviation 3,435.

The traditional test statistic for this problem (normal population and standard deviation known) is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1{,}424.4 - 1{,}200}{3{,}435/\sqrt{20}} = 0.292,$$

where μ_0 is the value that separates the null and alternative hypotheses. The null hypothesis is rejected if $z > 1.645$. Because 0.292 is less than 1.645, the null hypothesis

is not rejected. The data do not support the assertion that the average claim exceeds 1,200. □

The test in the previous example was constructed to meet certain objectives. The first objective is to control what is called the *Type I error*. It is the error made when the test rejects the null hypothesis in a situation where it happens to be true. In the example, the null hypothesis can be true in more than one way. As a result, a measure of the propensity of a test to make a Type I error must be carefully defined.

Definition 10.10 *The* ***significance level*** *of a hypothesis test is the probability of making a Type I error given that the null hypothesis is true. If it can be true in more than one way, the level of significance is the maximum of such probabilities. The significance level is usually denoted by* α.

This is a conservative definition in that it looks at the worst case. It is typically a case that is on the boundary between the two hypotheses.

■ **EXAMPLE 10.19**

Determine the level of significance for the test in Example 10.18.

Begin by computing the probability of making a Type I error when the null hypothesis is true with $\mu = 1{,}200$. Then,

$$\Pr(Z > 1.645 | \mu = 1{,}200) = 0.05$$

because the assumptions imply that Z has a standard normal distribution.

Now suppose μ has a value that is below 1,200. Then

$$\Pr\left(\frac{\bar{X} - 1{,}200}{3{,}435/\sqrt{20}} > 1.645\right) = \Pr\left(\frac{\bar{X} - \mu + \mu - 1{,}200}{3{,}435/\sqrt{20}} > 1.645\right)$$
$$= \Pr\left(\frac{\bar{X} - \mu}{3{,}435/\sqrt{20}} > 1.645 - \frac{\mu - 1{,}200}{3{,}435/\sqrt{20}}\right).$$

The random variable on the left has a standard normal distribution. Because μ is known to be less than 1,200, the right-hand side is always greater than 1.645. Therefore the probability is less than 0.05, and so the significance level is 0.05. □

The significance level is usually set in advance and is often between 1% and 10%. The second objective is to keep the *Type II error* (not rejecting the null hypothesis when the alternative is true) probability small. Generally, attempts to reduce the probability of one type of error increase the probability of the other. The best we can do once the significance level has been set is to make the Type II error as small as possible, though there is no assurance that the probability will be a small number. The best test is one that meets the following requirement.

Definition 10.11 *A hypothesis test is* ***uniformly most powerful*** *if no other test exists that has the same or lower significance level and, for a particular value within the alternative hypothesis, has a smaller probability of making a Type II error.*

EXAMPLE 10.20

(*Example 10.19 continued*) Determine the probability of making a Type II error when the alternative hypothesis is true with $\mu = 2{,}000$.

$$\begin{aligned}
&\Pr\left(\frac{\bar{X} - 1{,}200}{3{,}435/\sqrt{20}} < 1.645 | \mu = 2{,}000\right) \\
&\quad = \Pr(\bar{X} - 1{,}200 < 1{,}263.51 | \mu = 2{,}000) \\
&\quad = \Pr(\bar{X} < 2{,}463.51 | \mu = 2{,}000) \\
&\quad = \Pr\left(\frac{\bar{X} - 2{,}000}{3{,}435/\sqrt{20}} < \frac{2{,}463.51 - 2{,}000}{3{,}435/\sqrt{20}} = 0.6035\right) = 0.7269.
\end{aligned}$$

For this value of μ, the test is not very powerful, having over a 70% chance of making a Type II error. Nevertheless (though this is not easy to prove), the test used is the most powerful test for this problem. □

Because the Type II error probability can be high, it is customary to not make a strong statement when the null hypothesis is not rejected. Rather than say we choose or accept the null hypothesis, we say that we fail to reject it. That is, there was not enough evidence in the sample to make a strong argument in favor of the alternative hypothesis, so we take no stand at all.

A common criticism of this approach to hypothesis testing is that the choice of the significance level is arbitrary. In fact, by changing the significance level, any result can be obtained.

EXAMPLE 10.21

(*Example 10.20 continued*) Complete the test using a significance level of $\alpha = 0.45$. Then determine the range of significance levels for which the null hypothesis is rejected and for which it is not rejected.

Because $\Pr(Z > 0.1257) = 0.45$, the null hypothesis is rejected when

$$\frac{\bar{X} - 1{,}200}{3{,}435/\sqrt{20}} > 0.1257.$$

In this example, the test statistic is 0.292, which is in the rejection region, and thus the null hypothesis is rejected. Of course, few people would place confidence in the results of a test that was designed to make errors 45% of the time. Because $\Pr(Z > 0.292) = 0.3851$, the null hypothesis is rejected for those who select a significance level that is greater than 38.51% and is not rejected by those who use a significance level that is less than 38.51%. □

Few people are willing to make errors 38.51% of the time. Announcing this figure is more persuasive than the earlier conclusion based on a 5% significance level. When a significance level is used, readers are left to wonder what the outcome would have been with other significance levels. The value of 38.51% is called a p-value. A working definition follows.

Definition 10.12 *For a hypothesis test, the* ***p-value*** *is the probability that the test statistic takes on a value that is less in agreement with the null hypothesis than the value obtained from the sample. Tests conducted at a significance level that is greater than the p-value will lead to a rejection of the null hypothesis, while tests conducted at a significance level that is smaller than the p-value will lead to a failure to reject the null hypothesis.*

Also, because the p-value must be between 0 and 1, it is on a scale that carries some meaning. The closer to zero the value is, the more support the data give to the alternative hypothesis. Common practice is that values above 10% indicate that the data provide no evidence in support of the alternative hypothesis, while values below 1% indicate strong support for the alternative hypothesis. Values in between indicate uncertainty as to the appropriate conclusion and may call for more data or a more careful look at the data or the experiment that produced it.

This approach to hypothesis testing has some consequences that can create difficulties when answering actuarial questions. The following example illustrate these problems.

■ **EXAMPLE 10.22**

You believe that the lognormal model is appropriate for the problem you are investigating. You have collected some data and would like to test this hypothesis. What are the null and alternative hypotheses and what will you learn after completing the test?

Methods for conducting this test are presented in Section **??**. One hypothesis is that the population has the lognormal distribution and the other is that it does not. The first one is the statement of equality and so must be the null hypothesis. The problem is that while data can confirm that the population is *not* lognormal, the method does not allow you to assert that the population *is* lognormal. A second problem is that often the null hypothesis is known to be false. In this case we know that the population is unlikely to be exactly lognormal. If our sample size is large enough, the hypothesis test will discover this, and it is likely that all models will be rejected. □

It is important to keep in mind that hypothesis testing was invented for situations where collecting data was either expensive or inconvenient. For example, in deciding if a new drug cures a disease, it is important to confirm this fact with the smallest possible sample so that, if the results are favorable, the drug can be approved and made available. Or, consider testing a new crop fertilizer. Every test acre planted costs time and money. In contrast, in many types of actuarial problems, there is a large amount of data available from historical records. In this case, unless the data follow a parametric model extremely closely, almost any model can be rejected by using a sufficiently large set of data.

10.5.1 Exercise

10.34 (*Exercise 10.11 continued*) Test $H_0 : \theta \geq 325$ versus $H_1 : \theta < 325$ using a significance level of 5% and the sample mean as the test statistic. Also, compute the p-value. Do this using the exact distribution of the test statistic and a normal approximation.

10.6 Solutions to Exercises

10.1 When three observations are taken without replacement, there are only four possible results. They are 1,3,5; 1,3,9; 1,5,9; and 3,5,9. The four sample means are 9/3, 13/3, 15/3, and 17/3. The expected value (each has probability 1/4) is 54/12 or 4.5, which equals the population mean. The four sample medians are 3, 3, 5, and 5. The expected value is 4, and so the median is biased.

10.2

$$\begin{aligned} E(\bar{X}) &= E\left[\frac{1}{n}(X_1 + \cdots + X_n)\right] = \frac{1}{n}\left[E(X_1) + \cdots + E(X_n)\right] \\ &= \frac{1}{n}(\mu + \cdots + \mu) = \mu. \end{aligned}$$

10.3 For a sample of size 3 from a continuous distribution, the density function of the median is $6f(x)F(x)[1-F(x)]$. For this exercise, $F(x) = (x-\theta+2)/4, \theta-2 < x < \theta+2$. The density function for the median is

$$f_{med}(x) = 6(0.25)\frac{x-\theta+2}{4}\frac{2-x+\theta}{4}$$

and the expected value is

$$\begin{aligned} &6\int_{\theta-2}^{\theta+2} \frac{x(x-\theta+2)(2-x+\theta)}{64}dx = \frac{6}{64}\int_0^4 (y+\theta-2)y(4-y)dy \\ &= \frac{6}{64}\int_0^4 -y^3 + (6-\theta)y^2 + 4(\theta-2)ydy \\ &= \frac{6}{64}\left[-\frac{y^4}{4} + \frac{(6-\theta)y^3}{3} + \frac{4(\theta-2)y^2}{2}\right]\Bigg|_0^4 \\ &= \frac{6}{64}\left[-64 + \frac{(6-\theta)64}{3} + 32(\theta-2)\right] \\ &= \theta, \end{aligned}$$

where the first line used the substitution $y = x - \theta + 2$.

10.4 Because the mean of a Pareto distribution does not always exist, it is not reasonable to discuss unbiasedness or consistency. Had the problem been restricted to Pareto distributions with $\alpha > 1$, then consistency can be established. It turns out that for the sample mean to be consistent, only the first moment needs to exist (the variance having a limit of zero is a sufficient, but not a necessary, condition for consistency).

10.5 The mean is unbiased, so its MSE is its variance. It is

$$MSE_{\text{mean}}(\theta) = \text{Var}(X)/3 = \frac{4^2}{12(3)} = \frac{4}{9}.$$

The median is also unbiased. The variance is

$$\begin{aligned}
&6\int_{\theta-2}^{\theta+2}\frac{(x-\theta)^2(x-\theta+2)(2-x+\theta)}{64}dx = \frac{6}{64}\int_0^4 (y-2)^2y(4-y)dy\\
&\quad=\frac{6}{64}\int_0^4 -y^4+8y^3-20y^2+16ydy\\
&\quad=\frac{6}{64}\left[-\frac{4^5}{5}+\frac{8(4)^4}{4}-\frac{20(4)^3}{3}+\frac{16(4)^2}{2}\right]\\
&\quad=4/5,
\end{aligned}$$

and so the sample mean has the smaller MSE.

10.6 We have

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_C) &= \mathrm{Var}[w\hat{\theta}_A+(1-w)\hat{\theta}_B]\\
&= w^2(160,000)+(1-w)^2(40,000)\\
&= 200,000w^2-80,000w+40,000.
\end{aligned}$$

The derivative is $400,000w-80,000$, and setting it equal to zero provides the solution, $w=0.2$.

10.7 MSE $=$ Var $+$bias^2. $1=$ Var $+(0.2)^2$, Var $=0.96$.

10.8 To be unbiased,

$$m=\mathrm{E}(Z)=\alpha(0.8m)+\beta m=(0.8\alpha+\beta)m,$$

and so $1=0.8\alpha+\beta$ or $\beta=1-0.8\alpha$. Then

$$\mathrm{Var}(Z)=\alpha^2m^2+\beta^2 1.5m^2=[\alpha^2+(1-0.8\alpha)^2 1.5]m^2,$$

which is minimized when $\alpha^2+1.5-2.4\alpha+0.96\alpha^2$ is minimized. This result occurs when $3.92\alpha-2.4=0$ or $\alpha=0.6122$. Then $\beta=1-0.8(0.6122)=0.5102$.

10.9 One way to solve this problem is to list the 20 possible samples of size 3 and assign probability 1/20 to each. The population mean is $(1+1+2+3+5+10)/6=11/3$.

(a) The 20 sample means have an average of $11/3$, and so the bias is zero. The variance of the sample means (dividing by 20 because this is the population of sample means) is 1.9778, which is also the MSE.

(b) The 20 sample medians have an average of 2.7, and so the bias is $2.7-11/3=-0.9667$. The variance is 1.81 and the MSE is 2.7444.

(c) The 20 sample midranges have an average of 4.15, and so the bias is $4.15-11/3=0.4833$. The variance is 2.65 and the MSE is 2.8861.

(d) $E(aX_{(1)} + bX_{(2)} + cX_{(3)}) = 1.25a + 2.7b + 7.05c$, where the expected values of the order statistics can be found by averaging the 20 values from the enumerated population. To be unbiased, the expected value must be $11/3$, and so the restriction is $1.25a + 2.7b + 7.05c = 11/3$. With this restriction, the MSE is minimized at $a = 1.445337$, $b = 0.043733$, and $c = 0.247080$ with an MSE of 1.620325. With no restrictions, the minimum is at $a = 1.289870$, $b = 0.039029$, and $c = 0.220507$ with an MSE of 1.446047 (and a bias of -0.3944).

10.10 bias$(\hat{\theta}_1) = 165/75 - 2 = 0.2$, $\text{Var}(\hat{\theta}_1) = 375/75 - (165/75)^2 = 0.16$, MSE$(\hat{\theta}_1) = 0.16 + (0.2)^2 = 0.2$. bias$(\hat{\theta}_2) = 147/75 - 2 = -0.04$, $\text{Var}(\hat{\theta}_2) = 312/75 - (147/75)^2 = 0.3184$, MSE$(\hat{\theta}_2) = 0.3184 + (-0.04)^2 = 0.32$. The relative efficiency is $0.2/0.32 = 0.625$, or $0.32/0.20 = 1.6$.

10.11 (a) From the information given in the problem, we can begin with

$$0.95 = \Pr(a \le \bar{X}/\theta \le b),$$

where $\bar{X}/\theta$ is known to have the gamma distribution with $\alpha = 50$ and $\theta = 0.02$. This does not uniquely specify the endpoints. However, if 2.5% probability is allocated to each side, then $a = 0.7422$ and $b = 1.2956$. Inserting these values in the inequality, taking reciprocals, and multiplying through by $\bar{X}$ gives

$$0.95 = \Pr(0.7718\bar{X} \le \theta \le 1.3473\bar{X}).$$

Inserting the sample mean gives an interval of 212.25 to 370.51.

(b) The sample mean has $\text{E}(\bar{X}) = \theta$ and $\text{Var}(\bar{X}) = \theta^2/50$. Then, using $275^2/50$ as the approximate variance,

$$\begin{aligned} n0.95 &\doteq \Pr\left(-1.96 \le \frac{\bar{X} - \theta}{275/\sqrt{50}} \le 1.96\right) \\ &= \Pr(-76.23 \le \bar{X} - \theta \le 76.23). \end{aligned}$$

Inserting the sample mean of 275 gives the interval 275 ± 76.23, or 198.77 to 351.23.

(c) Leaving the θ in the variance alone,

$$\begin{aligned} 0.95 &\doteq \Pr\left(-1.96 \le \frac{\bar{X} - \theta}{\theta/\sqrt{50}} \le 1.96\right) \\ &= \Pr(-0.2772\theta \le \bar{X} - \theta \le 0.2772\theta) \\ &= \Pr(\bar{X}/1.2772 \le \theta \le \bar{X}/0.7228) \end{aligned}$$

for an interval of 215.31 to 380.46.

10.12 The estimated probability of one or more claims is $400/2{,}000 = 0.2$. For this binomial distribution, the variance is estimated as $0.2(0.8)/2{,}000 = 0.00008$. The upper bound is $0.2 + 1.96\sqrt{0.00008} = 0.21753$.

10.13 The equations to solve are

$$\exp(\mu + \sigma^2/2) = 1{,}424.4,$$
$$\exp(2\mu + 2\sigma^2) = 13{,}238{,}441.9.$$

Taking logarithms yields

$$\mu + \sigma^2/2 = 7.261506,$$
$$2\mu + 2\sigma^2 = 16.398635.$$

The solution of these two equations is $\hat{\mu} = 6.323695$ and $\hat{\sigma}^2 = 1.875623$, and then $\hat{\sigma} = 1.369534$.

10.14 The two equations to solve are

$$0.2 = 1 - e^{-(5/\theta)^\tau},$$
$$0.8 = 1 - e^{-(12/\theta)^\tau}.$$

Moving the 1 to the left-hand side and taking logarithms produces

$$0.22314 = (5/\theta)^\tau,$$
$$1.60944 = (12/\theta)^\tau.$$

Dividing the second equation by the first equation produces

$$7.21269 = 2.4^\tau.$$

Taking logarithms again produces

$$1.97584 = 0.87547\tau,$$

and so $\hat{\tau} = 2.25689$. Using the first equation,

$$\hat{\theta} = 5/(0.22314^{1/2.25689}) = 9.71868.$$

Then $\hat{S}(8) = e^{-(8/9.71868)^{2.25689}} = 0.52490$.

10.15 The equations to solve are

$$0.5 = 1 - \exp[-(10{,}000/\theta)^\tau],$$
$$0.9 = 1 - \exp[-(100{,}000/\theta)^\tau].$$

Then

$$\ln 0.5 = -0.69315 = -(10{,}000/\theta)^\tau,$$
$$\ln 0.1 = -2.30259 = -(100{,}000/\theta)^\tau.$$

Dividing the second equation by the first gives

$$3.32192 = 10^\tau,$$
$$\ln 3.32192 = 1.20054 = \tau \ln 10 = 2.30259\tau,$$
$$\hat{\tau} = 1.20054/2.30259 = 0.52139.$$

Then

$$\begin{aligned} 0.69315 &= (10{,}000/\theta)^{0.52139}, \\ 0.69315^{1/0.52139} &= 0.49512 = 10{,}000/\theta, \\ \hat{\theta} &= 20{,}197. \end{aligned}$$

10.16 The two moment equations are

$$\begin{aligned} \frac{4+5+21+99+421}{5} &= 110 = \frac{\theta}{\alpha-1}, \\ \frac{4^2+5^2+21^2+99^2+421^2}{5} &= 37,504.8 = \frac{2\theta^2}{(\alpha-1)(\alpha-2)}. \end{aligned}$$

Dividing the second equation by the square of the first equation gives

$$\frac{37,504.8}{110^2} = 3.0996 = \frac{2(\alpha-1)}{\alpha-2}.$$

The solution is $\hat{\alpha} = 3.8188$. From the first equation, $\hat{\theta} = 110(2.8188) = 310.068$. For the 95th percentile,

$$0.95 = 1 - \left(\frac{310.068}{310.068 + \pi_{0.95}}\right)^{3.8188}$$

for $\hat{\pi}_{0.95} = 369.37$.

10.17 After inflation, the 100 claims from year 1 total $100(10{,}000)(1.1)^2 = 1{,}210{,}000$, and the 200 claims from year 2 total $200(12{,}500)(1.1) = 2{,}750{,}000$. The average of the 300 inflated claims is $3{,}960{,}000/300 = 13{,}200$. The moment equation is $13{,}200 = \theta/(3-1)$, which yields $\hat{\theta} = 26{,}400$.

10.18 The equations to solve are

$$\begin{aligned} 0.2 &= F(18.25) = \Phi\left(\frac{\ln 18.25 - \mu}{\sigma}\right), \\ 0.8 &= F(35.8) = \Phi\left(\frac{\ln 35.8 - \mu}{\sigma}\right). \end{aligned}$$

The 20th and 80th percentiles of the normal distribution are -0.842 and 0.842, respectively. The equations become

$$\begin{aligned} -0.842 &= \frac{2.904 - \mu}{\sigma}, \\ 0.842 &= \frac{3.578 - \mu}{\sigma}. \end{aligned}$$

Dividing the first equation by the second yields

$$-1 = \frac{2.904 - \mu}{3.578 - \mu}.$$

The solution is $\hat{\mu} = 3.241$ and substituting in either equation yields $\hat{\sigma} = 0.4$. The probability of exceeding 30 is

$$\begin{aligned}\Pr(X > 30) &= 1 - F(30) = 1 - \Phi\left(\frac{\ln 30 - 3.241}{0.4}\right) = 1 - \Phi(0.4)\\ &= 1 - \Phi(0.4) = 1 - 0.6554 = 0.3446.\end{aligned}$$

10.19 For a mixture, the mean and second moment are a combination of the individual moments. The first two moments are

$$\begin{aligned}\mathrm{E}(X) &= p(1) + (1-p)(10) = 10 - 9p,\\ \mathrm{E}(X^2) &= p(2) + (1-p)(200) = 200 - 198p,\\ \mathrm{Var}(X) &= 200 - 198p - (10 - 9p)^2 = 100 - 18p - 81p^2 = 4.\end{aligned}$$

The only positive root of the quadratic equation is $\hat{p} = 0.983$.

10.20 We need the $0.6(21) = 12.6th$ smallest observation. It is $0.4(38) + 0.6(39) = 38.6$.

10.21 We need the $0.75(21) = 15.75th$ smallest observation. It is $0.25(13) + 0.75(14) = 13.75$.

10.22 $\hat{\mu} = 975$, $\hat{\mu}_2' = 977{,}916\frac{2}{3}$, $\hat{\sigma}^2 = 977{,}916\frac{2}{3} - 975^2 = 27{,}291\frac{2}{3}$. The moment equations are $975 = \alpha\theta$ and $27{,}291\frac{2}{3} = \alpha\theta^2$. The solutions are $\hat{\alpha} = 34.8321$ and $\hat{\theta} = 27.9915$.

10.23 $F(x) = (x/\theta)^\gamma/[1 + (x/\theta)^\gamma]$. The equations are $0.2 = (100/\theta)^\gamma/[1 + (100/\theta)^\gamma]$ and $0.8 = (400/\theta)^\gamma/[1 + (400/\theta)^\gamma]$. From the first equation $0.2 = 0.8(100/\theta)^\gamma$ or $\theta^\gamma = 4(100)^\gamma$. Insert this result in the second equation to get $0.8 = 4^{\gamma-1}/(1 + 4^{\gamma-1})$, and so $\hat{\gamma} = 2$ and then $\hat{\theta} = 200$.

10.24 $\mathrm{E}(X) = \int_0^1 px^p dx = p/(1+p) = \bar{x}$. $\hat{p} = \bar{x}/(1 - \bar{x})$.

10.25 $\hat{\mu} = 3{,}800 = \alpha\theta$. $\mu_2' = 16{,}332{,}000$, $\hat{\sigma}^2 = 1{,}892{,}000 = \alpha\theta^2$. $\hat{\alpha} = 7.6321$, $\hat{\theta} = 497.89$.

10.26 $\hat{\mu} = 2{,}000 = \exp(\mu + \sigma^2/2)$, $\hat{\mu}_2' = 6{,}000{,}000 = \exp(2\mu + 2\sigma^2)$. $7.690090 = \mu + \sigma^2/2$ and $15.60727 = 2\mu + 2\sigma^2$. The solutions are $\hat{\mu} = 7.39817$ and $\hat{\sigma} = 0.636761$.

$$\begin{aligned}\Pr(X > 4{,}500) &= 1 - \Phi[(\ln 4{,}500 - 7.39817)/0.636761]\\ &= 1 - \Phi(1.5919) = 0.056.\end{aligned}$$

10.27 $\hat{\mu} = 4.2 = (\beta/2)\sqrt{2\pi}$. $\hat{\beta} = 3.35112$.

10.28 X is Pareto, and so $\mathrm{E}(X) = 1{,}000/(\alpha - 1) = \bar{x} = 318.4$. $\hat{\alpha} = 4.141$.

10.29 $r\beta = 0.1001$, $r\beta(1+\beta) = 0.1103 - 0.1001^2 = 0.10027999$. $1+\beta = 1.0017981$, $\hat{\beta} = 0.0017981$, $\hat{r} = 55.670$.

10.30 $r\beta = 0.166$, $r\beta(1+\beta) = 0.252 - 0.166^2 = 0.224444$. $1+\beta = 1.352072$, $\hat{\beta} = 0.352072$, $\hat{r} = 0.47149$.

10.31 With $n+1 = 16$, we need the $0.3(16) = 4.8$ and $0.65(16) = 10.4$ smallest observations. They are $0.2(280) + 0.8(350) = 336$ and $0.6(450) + 0.4(490) = 466$. The equations are

$$\begin{aligned}
0.3 = 1 - \left(\frac{\theta^\gamma}{\theta^\gamma + 336^\gamma}\right)^2 &\text{ and } 0.65 = 1 - \left(\frac{\theta^\gamma}{\theta^\gamma + 466^\gamma}\right)^2,\\
(0.7)^{-1/2} = 1 + (336/\theta)^\gamma &\text{ and } (0.35)^{-1/2} = 1 + (466/\theta)^\gamma,\\
\frac{(0.7)^{-1/2} - 1}{(0.35)^{-1/2} - 1} = 0.282814 &= \left(\frac{336}{466}\right)^\gamma,\\
\ln(0.282814) &= \gamma \ln(336/466),\ \gamma = 3.8614,\\
(0.7)^{-1/2} - 1 &= (336/\theta)^{3.8614},\\
(0.19523)^{1/3.8614} &= 336/\theta,\ \theta = 558.74.
\end{aligned}$$

10.32 With $n+1 = 17$, we need the $0.2(17) = 3.4$ and $0.7(17) = 11.9$ smallest observations. They are $0.6(75)+0.4(81) = 77.4$ and $0.1(122)+0.9(125) = 124.7$. The equations are

$$\begin{aligned}
0.2 = 1 - e^{-(77.4/\theta)^\tau} &\text{ and } 0.7 = 1 - e^{-(124/7/\theta)^\tau},\\
-\ln 0.8 = 0.22314 = (77.4/\theta)^\tau &\text{ and } -\ln 0.3 = 1.20397 = (124.7/\theta)^\tau,\\
1.20397/0.22314 = 5.39558 &= (124.7/77.4)^\tau,\\
\tau &= \ln(5.39558)/\ln(124.7/77.4) = 3.53427,\\
\theta &= 77.4/0.22314^{1/3.53427} = 118.32.
\end{aligned}$$

10.33 Shifting adds δ to the mean and median. The median of the unshifted distribution is the solution to $0.5 = S(m) = e^{-m/\theta}$ for $m = \theta\ln(2)$. The equations to solve are

$$300 = \theta + \delta \text{ and } 240 = \theta\ln(2) + \delta.$$

Subtracting the second equation from the first gives $60 = \theta[1 - \ln(2)]$ for $\theta = 195.53$. From the first equation, $\delta = 104.47$.

10.34 For the exact test, null should be rejected if $\bar{X} \le c$. The value of c comes from

$$\begin{aligned}
0.05 &= \Pr(\bar{X} \le c|\theta = 325)\\
&= \Pr(\bar{X}/325 \le c/325),
\end{aligned}$$

where $\bar{X}/325$ has the gamma distribution with parameters 50 and 0.02. From that distribution, $c/325 = 0.7793$ for $c = 253.27$. Because the sample mean of 275 is not below this

value, the null hypothesis is not rejected. The p-value is obtained from

$$\Pr(\bar{X} \leq 275|\theta = 325) = \Pr(\bar{X}/325 \leq 275/325).$$

From the gamma distribution with parameters 50 and 0.02, this probability is 0.1353.

For the normal approximation,

$$\begin{aligned} 0.05 &= \Pr(\bar{X} \leq c|\theta = 325) \\ &= \Pr\left(Z \leq \frac{c - 325}{325/\sqrt{50}}\right). \end{aligned}$$

Solving $(c-325)/(325/\sqrt{50}) = -1.645$ produces $c = 249.39$. Again, the null hypothesis is not rejected. For the p-value,

$$\Pr\left(Z \leq \frac{275 - 325}{325/\sqrt{50}} = -1.0879\right) = 0.1383.$$

11

MAXIMUM LIKELIHOOD ESTIMATION

11.1 Introduction

Estimation by the method of moments and percentile matching is often easy to do, but these estimators tend to perform poorly mainly because they use a few features of the data, rather than the entire set of observations. It is particularly important to use as much information as possible when the population has a heavy right tail. For example, when estimating parameters for the normal distribution, the sample mean and variance are sufficient.[1] However, when estimating parameters for a Pareto distribution, it is important to know all the extreme observations in order to successfully estimate α. Another drawback of these methods is that they require that all the observations are from the same random variable. Otherwise, it is not clear what to use for the population moments or percentiles. For example, if half the observations have a deductible of 50 and half have a deductible of 100, it is not clear to what the sample mean should be equated.[2] Finally, these methods allow the analyst to make arbitrary decisions regarding the moments or percentiles to use.

[1]This applies both in the formal statistical definition of sufficiency (not covered here) and in the conventional sense. If the population has a normal distribution, the sample mean and variance convey as much information as the original observations.

[2]One way to rectify that drawback is to first determine a data-dependent model such as the Kaplan–Meier estimate introduced in Section 12.3. Then use percentiles or moments from that model.

Loss Models: From Data to Decisions, 5th edition.
By Stuart A. Klugman, Harry H. Panjer, Gordon E. Willmot

There are a variety of estimators that use the individual data points. All of them are implemented by setting an objective function and then determining the parameter values that optimize that function. Of the many possibilities, the only one presented here is the maximum likelihood estimator.

To define the maximum likelihood estimator, let the data set consist of n events $A_1, \ldots, A_n$, where A_j is whatever was observed for the jth observation. For example, A_j may consist of a single point or an interval. The latter arises, for example, with grouped data . Further assume that the event A_j results from observing the random variable X_j. The random variables $X_1, \ldots, X_n$ need not have the same probability distribution, but their distributions must depend on the same parameter vector, θ. In addition, the random variables are assumed to be independent.

Definition 11.1 *The **likelihood function** is*

$$L(\theta) = \prod_{j=1}^{n} \Pr(X_j \in A_j|\theta)$$

*and the **maximum likelihood estimate** of θ is the vector that maximizes the likelihood function.*[3]

In the definition, if A_j is a single point and the distribution is continuous, then $\Pr(X_j \in A_j|\theta)$ is interpreted as the probability density function evaluated at that point. In all other cases it is the probability of that event.

There is no guarantee that the function has a maximum at eligible parameter values. It is possible that as various parameters become zero or infinite, the likelihood function will continue to increase. Care must be taken when maximizing this function because there may be local maxima in addition to the global maximum. Often, it is not possible to analytically maximize the likelihood function (by setting partial derivatives equal to zero). Numerical approaches, such as those outlined in Appendix **??**, will usually be needed.

Because the observations are assumed to be independent, the product in the definition represents the joint probability $\Pr(X_1 \in A_1, \ldots, X_n \in A_n|\theta)$, that is, the likelihood function is the probability of obtaining the sample results that were obtained, given a particular parameter value. The estimate is then the parameter value that produces the model under which the actual observations are most likely to be observed. One of the major attractions of this estimator is that it is almost always available. That is, if you can write an expression for the desired probabilities, you can execute this method. If you cannot write and evaluate an expression for probabilities using your model, there is no point in postulating that model in the first place because you will not be able to use it to solve the larger actuarial problem using your model.

EXAMPLE 11.1

Suppose the data in Data Set B, introduced in Chapter 10 and reproduced here as Table 11.1 were such that the exact value of all observations above 250 was unknown. All that is known is that the value was greater than 250. Determine the maximum likelihood estimate of θ for an exponential distribution.

[3]Some authors write the likelihood function as $L(\theta|\mathbf{x})$, where the vector $\mathbf{x}$ represents the observed data. Because observed data can take many forms, the dependence of the likelihood function on the data is suppressed in the notation.

Table 11.1 Data Set B.

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1,193	1,340	1,884	2,558	15,743

For the first seven values, the set A_j contains the single point equal to the observation x_j. Thus the first seven terms of the product are

$$f(27)f(82)\cdots f(243) = \theta^{-1}e^{-27/\theta}\theta^{-1}e^{-82/\theta}\cdots\theta^{-1}e^{-243/\theta} = \theta^{-7}e^{-909/\theta}.$$

For each of the final 13 terms, the set A_j is the interval from 250 to infinity and, therefore, $\Pr(X_j \in A_j) = \Pr(X_j > 250) = e^{-250/\theta}$. There are 13 such factors making the likelihood function

$$L(\theta) = \theta^{-7}e^{-909/\theta}(e^{-250/\theta})^{13} = \theta^{-7}e^{-4{,}159/\theta}.$$

It is easier to maximize the logarithm of the likelihood function. Because it occurs so often, we denote the **loglikelihood function** as $l(\theta) = \ln L(\theta)$. Then

$$\begin{aligned} l(\theta) &= -7\ln\theta - 4{,}159\theta^{-1}, \\ l'(\theta) &= -7\theta^{-1} + 4{,}159\theta^{-2} = 0, \\ \hat{\theta} &= \frac{4{,}159}{7} = 594.14. \end{aligned}$$

In this case, the calculus technique of setting the first derivative equal to zero is easy to do. Also, evaluating the second derivative at this solution produces a negative number, verifying that this solution is a maximum. □

11.2 Individual data

Consider the special case where the value of each observation is recorded. It is easy to write the loglikelihood function:

$$L(\theta) = \prod_{j=1}^{n} f_{X_j}(x_j|\theta), \quad l(\theta) = \sum_{j=1}^{n} \ln f_{X_j}(x_j|\theta).$$

The notation indicates that it is not necessary for each observation to come from the same distribution, but we continue to assume the parameter vector is common to each distribution.

EXAMPLE 11.2

Using Data Set B, determine the maximum likelihood estimates for an exponential distribution, for a gamma distribution where α is known to equal 2, and for a gamma distribution where both parameters are unknown.

For the exponential distribution, the general solution is

$$\begin{aligned} l(\theta) &= \sum_{j=1}^{n} \left(-\ln\theta - x_j\theta^{-1}\right) = -n\ln\theta - n\bar{x}\theta^{-1}, \\ l'(\theta) &= -n\theta^{-1} + n\bar{x}\theta^{-2} = 0, \\ n\theta &= n\bar{x}, \\ \hat{\theta} &= \bar{x}. \end{aligned}$$

For Data Set B, $\hat{\theta} = \bar{x} = 1{,}424.4$. The value of the loglikelihood function is -165.23. For this situation the method-of-moments and maximum likelihood estimates are identical. See Exercise 11.16 for further insight.

For the gamma distribution with $\alpha = 2$,

$$\begin{aligned} f(x|\theta) &= \frac{x^{2-1}e^{-x/\theta}}{\Gamma(2)\theta^2} = x\theta^{-2}e^{-x/\theta}, \\ \ln f(x|\theta) &= \ln x - 2\ln\theta - x\theta^{-1}, \\ l(\theta) &= \sum_{j=1}^{n} \ln x_j - 2n\ln\theta - n\bar{x}\theta^{-1}, \\ l'(\theta) &= -2n\theta^{-1} + n\bar{x}\theta^{-2} = 0, \\ \hat{\theta} &= \tfrac{1}{2}\bar{x}. \end{aligned}$$

For Data Set B, $\hat{\theta} = 1{,}424.4/2 = 712.2$ and the value of the loglikelihood function is -179.98. Again, this estimate is the same as the method-of-moments estimate.

For the gamma distribution with unknown parameters, the function is not as simple:

$$\begin{aligned} f(x|\alpha,\theta) &= \frac{x^{\alpha-1}e^{-x/\theta}}{\Gamma(\alpha)\theta^{\alpha}}, \\ \ln f(x|\alpha,\theta) &= (\alpha-1)\ln x - x\theta^{-1} - \ln\Gamma(\alpha) - \alpha\ln\theta. \end{aligned}$$

The partial derivative with respect to α requires the derivative of the gamma function. The resulting equation cannot be solved analytically. Using numerical methods, the estimates are $\hat{\alpha} = 0.55616$ and $\hat{\theta} = 2{,}561.1$ and the value of the loglikelihood function is -162.29. These do not match the method-of-moments estimates. □

11.2.1 Exercises

11.1 Repeat Example 11.2 using the inverse exponential, inverse gamma with $\alpha = 2$, and inverse gamma distributions. Compare your estimates with the method-of-moments estimates.

11.2 (*) You are given the five observations 521, 658, 702, 819, and 1,217. Your model is the single-parameter Pareto distribution with distribution function

$$F(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \ x > 500, \ \alpha > 0.$$

Determine the maximum likelihood estimate of α.

11.3 (*) You have observed the following five claim severities: 11.0, 15.2, 18.0, 21.0, and 25.8. Determine the maximum likelihood estimate of μ for the following model (which is the reciprocal inverse Gaussian distribution, see Exercise 5.20a of [13]):

$$f(x) = \frac{1}{\sqrt{2\pi x}} \exp\left[-\frac{1}{2x}(x-\mu)^2\right], \ x, \mu > 0.$$

11.4 (*) The following values were calculated from a random sample of 10 losses:

$$\sum_{j=1}^{10} x_j^{-2} = 0.00033674, \quad \sum_{j=1}^{10} x_j^{-1} = 0.023999,$$

$$\sum_{j=1}^{10} x_j^{-0.5} = 0.34445, \quad \sum_{j=1}^{10} x_j^{0.5} = 488.97$$

$$\sum_{j=1}^{10} x_j = 31{,}939, \quad \sum_{j=1}^{10} x_j^2 = 211{,}498{,}983.$$

Losses come from a Weibull distribution with $\tau = 0.5$ [so $F(x) = 1 - e^{-(x/\theta)^{0.5}}$]. Determine the maximum likelihood estimate of θ.

11.5 (*) A sample of n independent observations $x_1, \ldots, x_n$ came from a distribution with a pdf of $f(x) = 2\theta x \exp(-\theta x^2), \ x > 0$. Determine the maximum likelihood estimator (mle) of θ.

11.6 (*) Let $x_1, \ldots, x_n$ be a random sample from a population with cdf $F(x) = x^p, \ 0 < x < 1$. Determine the mle of p.

11.7 A random sample of 10 claims obtained from a gamma distribution is given as follows:

1,500 6,000 3,500 3,800 1,800 5,500 4,800 4,200 3,900 3,000

(a) (*) Suppose it is known that $\alpha = 12$. Determine the maximum likelihood estimate of θ.

(b) Determine the maximum likelihood estimates of α and θ.

11.8 A random sample of five claims from a lognormal distribution is given as follows:

500 1,000 1,500 2,500 4,500

Estimate μ and σ by maximum likelihood. Estimate the probability that a loss will exceed 4,500.

11.9 (*) Let $x_1, \ldots, x_n$ be a random sample from a random variable with pdf $f(x) = \theta^{-1} e^{-x/\theta}, \ x > 0$. Determine the mle of θ.

11.10 (*) The random variable X has pdf $f(x) = \beta^{-2}x\exp(-0.5x^2/\beta^2),\ x, \beta > 0$. For this random variable, $\mathrm{E}(X) = (\beta/2)\sqrt{2\pi}$ and $\mathrm{Var}(X) = 2\beta^2 - \pi\beta^2/2$. You are given the following five observations:

$$4.9 \quad 1.8 \quad 3.4 \quad 6.9 \quad 4.0$$

Determine the maximum likelihood estimate of β.

11.11 (*) Let $x_1, \ldots, x_n$ be a random sample from a random variable with cdf $F(x) = 1 - x^{-\alpha},\ x > 1,\ \alpha > 0$. Determine the mle of α.

11.12 (*) The random variable X has pdf $f(x) = \alpha\lambda^\alpha(\lambda + x)^{-\alpha-1},\ x, \alpha, \lambda > 0$. It is known that $\lambda = 1{,}000$. You are given the following five observations:

$$43 \quad 145 \quad 233 \quad 396 \quad 775$$

Determine the maximum likelihood estimate of α.

11.13 The following 20 observations were collected. It is desired to estimate $\Pr(X > 200)$. When a parametric model is called for, use the single-parameter Pareto distribution for which $F(x) = 1 - (100/x)^\alpha,\ x > 100,\ \alpha > 0$.

132	149	476	147	135	110	176	107	147	165
135	117	110	111	226	108	102	108	227	102

(a) Determine the empirical estimate of $\Pr(X > 200)$.

(b) Determine the method-of-moments estimate of the single-parameter Pareto parameter α and use it to estimate $\Pr(X > 200)$.

(c) Determine the maximum likelihood estimate of the single-parameter Pareto parameter α and use it to estimate $\Pr(X > 200)$.

11.14 Consider the inverse Gaussian distribution with density given by

$$f_X(x) = \left(\frac{\theta}{2\pi x^3}\right)^{1/2} \exp\left[-\frac{\theta}{2x}\left(\frac{x-\mu}{\mu}\right)^2\right], \quad x > 0.$$

(a) Show that

$$\sum_{j=1}^{n} \frac{(x_j - \mu)^2}{x_j} = \mu^2 \sum_{j=1}^{n} \left(\frac{1}{x_j} - \frac{1}{\overline{x}}\right) + \frac{n}{\overline{x}}(\overline{x} - \mu)^2,$$

where $\overline{x} = (1/n)\sum_{j=1}^{n} x_j$.

(b) For a sample $(x_1, \ldots, x_n)$, show that the maximum likelihood estimates of μ and θ are

$$\hat{\mu} = \bar{x}$$

and

$$\hat{\theta} = \frac{n}{\sum_{j=1}^{n}\left(\frac{1}{x_j} - \frac{1}{\bar{x}}\right)}.$$

11.15 Suppose that $X_1, \ldots, X_n$ are independent and normally distributed with mean $\mathrm{E}(X_j) = \mu$ and $\mathrm{Var}(X_j) = (\theta m_j)^{-1}$, where $m_j > 0$ is a known constant. Prove that the maximum likelihood estimates of μ and θ are

$$\hat{\mu} = \bar{X}$$

and

$$\hat{\theta} = n \left[\sum_{j=1}^{n} m_j (X_j - \bar{X})^2 \right]^{-1},$$

where $\bar{X} = (1/m) \sum_{j=1}^{n} m_j X_j$ and $m = \sum_{j=1}^{n} m_j$.

11.16 Suppose $X_1, \ldots, X_n$ are i.i.d. with distribution (**??**). Prove that the maximum likelihood estimate of the mean is the sample mean. In other words, if $\hat{\theta}$ is the mle of θ, prove that

$$\widehat{\mu(\theta)} = \mu(\hat{\theta}) = \bar{X}.$$

11.3 Grouped data

When data are grouped and the groups span the range of possible observations, the observations may be summarized as follows. Begin with a set of numbers $c_0 < c_1 < \cdots < c_k$, where c_0 is the smallest possible observation (often zero) and c_k is the largest possible observation (often infinity). From the sample, let n_j be the number of observations in the interval $(c_{j-1}, c_j]$. For such data, the likelihood function is

$$L(\theta) = \prod_{j=1}^{k} [F(c_j|\theta) - F(c_{j-1}|\theta)]^{n_j},$$

and its logarithm is

$$l(\theta) = \sum_{j=1}^{k} n_j \ln[F(c_j|\theta) - F(c_{j-1}|\theta)].$$

EXAMPLE 11.3

Using Data Set C from Chapter 10 and reproduced here as Table 11.2, determine the maximum likelihood estimate for an exponential distribution.

The loglikelihood function is

$$\begin{aligned} l(\theta) &= 99 \ln[F(7{,}500) - F(0)] + 42 \ln[F(17{,}500) - F(7{,}500)] + \cdots \\ &\quad + 3 \ln[1 - F(300{,}000)] \\ &= 99 \ln(1 - e^{-7{,}500/\theta}) + 42 \ln(e^{-7{,}500/\theta} - e^{-17{,}500/\theta}) + \cdots \\ &\quad + 3 \ln e^{-300{,}000/\theta}. \end{aligned}$$

A numerical routine is needed to produce $\hat{\theta} = 29{,}721$, and the value of the loglikelihood function is -406.03. □

Table 11.2 Data Set C.

Payment range	Number of payments
0–7,500	99
7,500–17,500	42
17,500–32,500	29
32,500–67,500	28
67,500–125,000	17
125,000–300,000	9
Over 300,000	3

Table 11.3 Data for Exercise 11.19.

Loss	No. of observations	Loss	No. of observations
0–25	5	350–500	17
25–50	37	500–750	13
50–75	28	750–1000	12
75–100	31	1,000–1,500	3
100–125	23	1,500–2,500	5
125–150	9	2,500–5,000	5
150–200	22	5,000–10,000	3
200–250	17	10,000–25,000	3
250–350	15	25,000–	2

11.3.1 Exercises

11.17 From Data Set C, determine the maximum likelihood estimates for gamma, inverse exponential, and inverse gamma distributions.

11.18 (*) Losses follow a distribution with cdf $F(x) = 1 - \theta/x$, $x > \theta$. A sample of 20 losses contained 9 below 10, 6 between 10 and 25, and 5 in excess of 25. Determine the maximum likelihood estimate of θ.

11.19 The data in Table 11.3 presents the results of a sample of 250 losses. Consider the inverse exponential distribution with cdf $F(x) = e^{-\theta/x}$, $x > 0$, $\theta > 0$. Determine the maximum likelihood estimate of θ.

11.4 Truncated or censored data

The definition of right censoring is:

Definition 11.2 *An observation is* ***censored from above*** *(also called* ***right censored****) at u if when it is at or above u it is recorded as being equal to u, but when it is below u it is recorded at its observed value.*

A similar definition applies to left censoring, which is uncommon in insurance settings.

Data Set D in Chapter 10 illustrates how censoring can occur in mortality data. If observation of an insured ends before death, all we know is that death occurs sometime after the time of the last observation. Another common situation a policy limit where, if the actual loss exceeds the limit, all that is known is that the limit was exceeded.

When data are censored, there is no additional complication. Right censoring simply creates an interval running from the censoring point to infinity. data below the censoring point are individual data, and so the likelihood function contains both density and distribution function terms.

The definition of trunctation is:

Definition 11.3 *An observation is* ***truncated from below*** (*also called* ***left truncated***) *at d if when it is at or below d it is not recorded, but when it is above d it is recorded at its observed value.*

A similar definition applies to right truncation, which is uncommon in insurance settings.

Data Set D also illustrates left truncation. For policies sold before the observation period begins, some will die while others will be alive to enter observation. Not only will their death times be unrecorded, we will not even know how many there were. Another common situation is a deductible. Losses below the deductible are not recorded and there is no count of how many losses were below the deductible.

Truncated data present more of a challenge. There are two ways to proceed. One is to shift the data by subtracting the truncation point from each observation. The other is to accept the fact that there is no information about values below the truncation point but then attempt to fit a model for the original population.

■ **EXAMPLE 11.4**

Assume the values in Data Set B had been truncated from below at 200. Using both methods, estimate the value of α for a Pareto distribution with $\theta = 800$ known. Then use the model to estimate the cost per payment with deductibles of 0, 200, and 400.

Using the shifting approach, the values become 43, 94, 140, 184, 257, 480, 655, 677, 774, 993, 1,140, 1,684, 2,358, and 15,543. The likelihood function is

$$
\begin{aligned}
L(\alpha) &= \prod_{j=1}^{14} \frac{\alpha(800^{\alpha})}{(800+x_j)^{\alpha+1}},\\
l(\alpha) &= \sum_{j=1}^{14} [\ln\alpha + \alpha\ln 800 - (\alpha+1)\ln(x_j+800)]\\
&= 14\ln\alpha + 93.5846\alpha - 103.969(\alpha+1)\\
&= 14\ln\alpha - 103.969 - 10.384\alpha,\\
l'(\alpha) &= 14\alpha^{-1} - 10.384,\\
\hat{\alpha} &= \frac{14}{10.384} = 1.3482.
\end{aligned}
$$

Because the data have been shifted, it is not possible to estimate the cost with no deductible. With a deductible of 200, the expected cost is the expected value of the estimated Pareto distribution, $800/0.3482 = 2{,}298$. Raising the deductible to 400 is

equivalent to imposing a deductible of 200 on the modeled distribution. From Theorem **??**, the expected cost per payment is

$$\frac{\mathrm{E}(X)-\mathrm{E}(X\wedge 200)}{1-F(200)}=\frac{\frac{800}{0.3482}\left(\frac{800}{200+800}\right)^{0.3482}}{\left(\frac{800}{200+800}\right)^{1.3482}}=\frac{1{,}000}{0.3482}=2{,}872.$$

For the unshifted approach, we need to ask the key question required when constructing the likelihood function. That is, what is the probability of observing each value knowing that values under 200 are omitted from the data set? This becomes a conditional probability and therefore the likelihood function is (where the x_j values are now the original values)

$$\begin{aligned}
L(\alpha)&=\prod_{j=1}^{14}\frac{f(x_j|\alpha)}{1-F(200|\alpha)}=\prod_{j=1}^{14}\left[\frac{\alpha(800^\alpha)}{(800+x_j)^{\alpha+1}}\Big/\left(\frac{800}{800+200}\right)^{\alpha}\right]\\
&=\prod_{j=1}^{14}\frac{\alpha(1{,}000^\alpha)}{(800+x_j)^{\alpha+1}},\\
l(\alpha)&=14\ln\alpha+14\alpha\ln 1{,}000-(\alpha+1)\sum_{j=1}^{14}\ln(800+x_j),\\
&=14\ln\alpha+96.709\alpha-(\alpha+1)105.810,\\
l'(\alpha)&=14\alpha^{-1}-9.101,\\
\hat{\alpha}&=1.5383.
\end{aligned}$$

This model is for losses with no deductible, and therefore the expected payment without a deductible is $800/0.5383=1{,}486$. Imposing deductibles of 200 and 400 produces the following results:

$$\begin{aligned}
\frac{\mathrm{E}(X)-\mathrm{E}(X\wedge 200)}{1-F(200)}&=\frac{1{,}000}{0.5383}=1{,}858,\\
\frac{\mathrm{E}(X)-\mathrm{E}(X\wedge 400)}{1-F(400)}&=\frac{1{,}200}{0.5383}=2{,}229.
\end{aligned}$$

□

It should now be clear that the contribution to the likelihood function can be written for most any type of observation. In general, the likelihood is always (proportional to) the probability of observing the data under the model, with the understanding for this purpose that pdfs for continuous variables should be viewed as probabilities. The following two steps summarize the process:

1. For the numerator, use $f(x)$ if the exact value, x, of the observation is known. If it is only known that the observation is between y and z, use $F(z)-F(y)$.

2. For the denominator, if there is no trunction, the denominator is 1. Else, let d be the truncation point, in which case the denominator is $1-F(d)$.

Table 11.4 Likelihood function for Example 11.5.

Obs.	x, y	d	L	Obs.	x, y	d	L
1	$y = 0.1$	0	$1 - F(0.1)$	16	$x = 4.8$	0	$f(4.8)$
2	$y = 0.5$	0	$1 - F(0.5)$	17	$y = 4.8$	0	$1 - F(4.8)$
3	$y = 0.8$	0	$1 - F(0.8)$	18	$y = 4.8$	0	$1 - F(4.8)$
4	$x = 0.8$	0	$f(0.8)$	19–30	$y = 5.0$	0	$1 - F(5.0)$
5	$y = 1.8$	0	$1 - F(1.8)$	31	$y = 5.0$	0.3	$\frac{1-F(5.0)}{1-F(0.3)}$
6	$y = 1.8$	0	$1 - F(1.8)$	32	$y = 5.0$	0.7	$\frac{1-F(5.0)}{1-F(0.7)}$
7	$y = 2.1$	0	$1 - F(2.1)$	33	$x = 4.1$	1.0	$\frac{f(4.1)}{1-F(1.0)}$
8	$y = 2.5$	0	$1 - F(2.5)$	34	$x = 3.1$	1.8	$\frac{f(3.1)}{1-F(1.8)}$
9	$y = 2.8$	0	$1 - F(2.8)$	35	$y = 3.9$	2.1	$\frac{1-F(3.9)}{1-F(2.1)}$
10	$x = 2.9$	0	$f(2.9)$	36	$y = 5.0$	2.9	$\frac{1-F(5.0)}{1-F(2.9)}$
11	$x = 2.9$	0	$f(2.9)$	37	$y = 4.8$	2.9	$\frac{1-F(4.8)}{1-F(2.9)}$
12	$y = 3.9$	0	$1 - F(3.9)$	38	$x = 4.0$	3.2	$\frac{f(4.0)}{1-F(3.2)}$
13	$x = 4.0$	0	$f(4.0)$	39	$y = 5.0$	3.4	$\frac{1-F(5.0)}{1-F(3.4)}$
14	$y = 4.0$	0	$1 - F(4.0)$	40	$y = 5.0$	3.9	$\frac{1-F(5.0)}{1-F(3.9)}$
15	$y = 4.1$	0	$1 - F(4.1)$				

EXAMPLE 11.5

Determine Pareto and gamma models for the time to death for Data Set D, introduced in Chapter 10.

Table 11.4 shows how the likelihood function is constructed for these observations. For observed deaths, the time is known, and so the exact value of x is available. For surrenders or those reaching time 5, the observation is censored, and therefore death is known to be some time in the interval from the surrender time, y, to infinity. In the table, $z = \infty$ is not noted because all interval observations end at infinity. The likelihood function must be maximized numerically. For the Pareto distribution, there is no solution. The likelihood function keeps getting larger as α and θ get larger.[4] For the gamma distribution, the maximum is at $\hat{\alpha} = 2.617$ and $\hat{\theta} = 3.311$. □

Discrete data present no additional problems.

[4]For a Pareto distribution, the limit as the parameters α and θ become infinite with the ratio being held constant is an exponential distribution. Thus, for this example, the exponential distribution is a better model (as measured by the likelihood function) than any Pareto model.

Table 11.5 Data Set A.

Number of accidents	Number of drivers
0	81,714
1	11,306
2	1,618
3	250
4	40
5 or more	7

EXAMPLE 11.6

For Data Set A, which was introduced in Chapter 10 and reproduced here as Table 11.5, assume that the seven drivers with five or more accidents all had exactly five accidents. Determine the maximum likelihood estimate for a Poisson distribution and for a binomial distribution with $m = 8$.

In general, for a discrete distribution with complete data, the likelihood function is

$$L(\theta) = \prod_{j=1}^{\infty} [p(x_j|\theta)]^{n_j},$$

where x_j is one of the observed values, $p(x_j|\theta)$ is the probability of observing x_j, and n_x is the number of times x was observed in the sample. For the Poisson distribution

$$L(\lambda) = \prod_{x=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^x}{x!} \right)^{n_x} = \prod_{x=0}^{\infty} \frac{e^{-n_x\lambda}\lambda^{xn_x}}{(x!)^{n_x}},$$

$$l(\lambda) = \sum_{x=0}^{\infty} (-n_x\lambda + xn_x \ln\lambda - n_x \ln x!) = -n\lambda + n\bar{x}\ln\lambda - \sum_{x=0}^{\infty} n_x \ln x!,$$

$$l'(\lambda) = -n + \frac{n\bar{x}}{\lambda} = 0,$$

$$\hat{\lambda} = \bar{x}.$$

For the binomial distribution

$$L(q) = \prod_{x=0}^{m} \left[\binom{m}{x} q^x (1-q)^{m-x} \right]^{n_x} = \prod_{x=0}^{m} \frac{m!^{n_x} q^{xn_x} (1-q)^{(m-x)n_x}}{(x!)^{n_x}[(m-x)!]^{n_x}},$$

$$l(q) = \sum_{x=0}^{m} [n_x \ln m! + xn_x \ln q + (m-x)n_x \ln(1-q)]$$
$$- \sum_{x=0}^{m} [n_x \ln x! + n_x \ln(m-x)!],$$

$$l'(q) = \sum_{x=0}^{m} \frac{xn_x}{q} - \frac{(m-x)n_x}{1-q} = \frac{n\bar{x}}{q} - \frac{mn - n\bar{x}}{1-q} = 0,$$

$$\hat{q} = \frac{\bar{x}}{m}.$$

For this problem, $\bar{x} = [81{,}714(0) + 11{,}306(1) + 1{,}618(2) + 250(3) + 40(4) + 7(5)]/94{,}935 = 0.16313$. Therefore, for the Poisson distribution, $\hat{\lambda} = 0.16313$, and for the binomial distribution, $\hat{q} = 0.16313/8 = 0.02039$. □

In Exercise 11.23 you are asked to estimate the Poisson parameter when the actual values for those with five or more accidents are not known.

11.4.1 Exercises

11.20 Determine maximum likelihood estimates for Data Set B using the inverse exponential, gamma, and inverse gamma distributions. Assume the data have been censored at 250 and then compare your answers to those obtained in Example 11.2 and Exercise 11.1.

11.21 Repeat Example 11.4 using a Pareto distribution with both parameters unknown.

11.22 Repeat Example 11.5, this time finding the distribution of the time to surrender.

11.23 Repeat Example 11.6, but this time assume that the actual values for the seven drivers who have five or more accidents are unknown. Note that this is a case of censoring.

11.24 (*) Five hundred losses are observed. Five of the losses are 1,100, 3,200, 3,300, 3,500, and 3,900. All that is known about the other 495 losses is that they exceed 4,000. Determine the maximum likelihood estimate of the mean of an exponential model.

11.25 (*) Ten claims were observed. The values of seven of them (in thousands) were 3, 7, 8, 12, 12, 13, and 14. The remaining three claims were all censored at 15. The proposed model has a hazard rate function given by

$$h(t) = \begin{cases} \lambda_1, & 0 < t < 5, \\ \lambda_2, & 5 \le t < 10, \\ \lambda_3, & t \ge 10. \end{cases}$$

Determine the maximum likelihood estimates of the three parameters.

11.26 (*) A random sample of size 5 is taken from a Weibull distribution with $\tau = 2$. Two of the sample observations are known to exceed 50 and the three remaining observations are 20, 30, and 45. Determine the maximum likelihood estimate of θ.

11.27 (*) Phil and Sylvia are competitors in the light bulb business. Sylvia advertises that her light bulbs burn twice as long as Phil's. You were able to test 20 of Phil's bulbs and 10 of Sylvia's. You assumed that both of their bulbs have an exponential distribution with time measured in hours. You have separately estimated the parameters as $\hat{\theta}_P = 1{,}000$ and $\hat{\theta}_S = 1{,}500$ for Phil and Sylvia, respectively, using maximum likelihood. Using all 30 observations, determine $\hat{\theta}^*$, the maximum likelihood estimate of θ_P restricted by Sylvia's claim that $\theta_S = 2\theta_P$.

11.28 (*) A sample of 100 losses revealed that 62 were below 1,000 and 38 were above 1,000. An exponential distribution with mean θ is considered. Using only the given information, determine the maximum likelihood estimate of θ. Now suppose you are also given that the 62 losses that were below 1,000 totalled 28,140, while the total for the 38

above 1,000 remains unknown. Using this additional information, determine the maximum likelihood estimate of θ.

11.29 (*) For claims reported in 1997, the number settled in 1997 (year 0) was unknown, the number settled in 1998 (year 1) was 3, and the number settled in 1999 (year 2) was 1. The number settled after 1999 is unknown. For claims reported in 1998, there were 5 settled in year 0, 2 settled in year 1, and the number settled after year 1 is unknown. For claims reported in 1999, there were 4 settled in year 0 and the number settled after year 0 is unknown. Let N be the year in which a randomly selected claim is settled and assume that it has probability function $\Pr(N = n) = p_n = (1-p)p^n$, $n = 0, 1, 2, \ldots$. Determine the maximum likelihood estimate of p.

11.30 (*) Losses have a uniform distribution on the interval $(0, w)$. Five losses are observed, all with a deductible of 4. Three losses are observed with values of 5, 9, and 13. The other two losses are censored at a value of $4 + p$. The maximum likelihood estimate of w is 29. Determine the value of p.

11.31 (*) Three losses are observed with values 66, 91, and 186. Seven other losses are known to be less than or equal to 60. Losses have an inverse exponential distribution with cdf $F(x) = e^{-\theta/x}$, $x > 0$. Determine the maximum likelihood estimate of the population mode.

11.32 (*) Policies have a deductible of 100. Seven losses are observed, with values 120, 180, 200, 270, 300, 1,000, and 2,500. Ground-up losses have a Pareto distribution with $\theta = 400$ and α unknown. Determine the maximum likelihood estimate of α.

11.5 Variance and interval estimation for maximum likelihood estimators

In general, it is not easy to determine the variance of complicated estimators such as the mle. However, it is possible to approximate the variance. The key is a theorem that can be found in most mathematical statistics books. The particular version stated here and its multiparameter generalization is taken from [17] and stated without proof. Recall that $L(\theta)$ is the likelihood function and $l(\theta)$ its logarithm. All of the results assume that the population has a distribution that is a member of the chosen parametric family.

Theorem 11.4 *Assume that the pdf (pf in the discrete case) $f(x;\theta)$ satisfies the following for θ in an interval containing the true value (replace integrals by sums for discrete variables):*

(i) $\ln f(x;\theta)$ *is three times differentiable with respect to* θ.

(ii) $\int \frac{\partial}{\partial\theta} f(x;\theta)\,dx = 0$. *This formula implies that the derivative may be taken outside the integral and so we are just differentiating the constant* 1.[5]

(iii) $\int \frac{\partial^2}{\partial\theta^2} f(x;\theta)\,dx = 0$. *This formula is the same concept for the second derivative.*

[5]The integrals in *(ii)* and *(iii)* are to be evaluated over the range of x values for which $f(x;\theta) > 0$.

(iv) $-\infty < \int f(x;\theta)\dfrac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\,dx < 0$. *This inequality establishes that the indicated integral exists and that the location where the derivative is zero is a maximum.*

(v) There exists a function $H(x)$ *such that*

$$\int H(x)f(x;\theta)\,dx < \infty \text{ with } \left|\frac{\partial^3}{\partial\theta^3}\ln f(x;\theta)\right| < H(x).$$

This inequality makes sure that the population is not overpopulated with regard to extreme values.

Then the following results hold:

(a) As $n \to \infty$, *the probability that the likelihood equation* $[L'(\theta) = 0]$ *has a solution goes to* 1.

(b) As $n \to \infty$, *the distribution of the mle* $\hat{\theta}_n$ *converges to a normal distribution with mean* θ *and variance such that* $I(\theta)\operatorname{Var}(\hat{\theta}_n) \to 1$, *where*

$$\begin{aligned} I(\theta) &= -n\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right] = -n\int f(x;\theta)\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\,dx \\ &= n\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right] = n\int f(x;\theta)\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2 dx. \end{aligned}$$

For any z, part (b) is to be interpreted as

$$\lim_{n\to\infty}\Pr\left(\frac{\hat{\theta}_n - \theta}{[I(\theta)]^{-1/2}} < z\right) = \Phi(z),$$

and therefore $[I(\theta)]^{-1}$ is a useful approximation for $\operatorname{Var}(\hat{\theta}_n)$. The quantity $I(\theta)$ is called the *information* (sometimes more specifically, *Fisher's information*). It follows from this result that the mle is asymptotically unbiased and consistent. The conditions in statements *(i)–(v)* are often referred to as "mild regularity conditions." A skeptic would translate this statement as "conditions that are almost always true but are often difficult to establish, so we'll just assume they hold in our case." Their purpose is to ensure that the density function is fairly smooth with regard to changes in the parameter and that there is nothing unusual about the density itself.[6]

The preceding results assume that the sample consists of i.i.d. random observations. A more general version of the result uses the logarithm of the likelihood function:

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}l(\theta)\right] = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}l(\theta)\right)^2\right].$$

The only requirement here is that the same parameter value apply to each observation.

[6]For an example of a situation where these conditions do not hold, see Exercise 11.34.

If there is more than one parameter, the only change is that the vector of maximum likelihood estimates now has an asymptotic multivariate normal distribution. The covariance matrix[7] of this distribution is obtained from the inverse of the matrix with (r, s)th element,

$$\begin{aligned}\mathbf{I}(\theta)_{rs} &= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_s\,\partial\theta_r}l(\theta)\right] = -n\mathrm{E}\left[\frac{\partial^2}{\partial\theta_s\,\partial\theta_r}\ln f(X;\theta)\right]\\ &= \mathrm{E}\left[\frac{\partial}{\partial\theta_r}l(\theta)\frac{\partial}{\partial\theta_s}l(\theta)\right] = n\mathrm{E}\left[\frac{\partial}{\partial\theta_r}\ln f(X;\theta)\frac{\partial}{\partial\theta_s}\ln f(X;\theta)\right].\end{aligned}$$

The first expression on each line is always correct. The second expression assumes that the likelihood is the product of n identical densities. This matrix is often called the *information matrix*. The information matrix also forms the Cramér–Rao lower bound as developed in Section 10.2.2.2. That is, under the usual conditions, no unbiased estimator has a smaller variance than that given by the inverse of the information. Therefore, at least asymptotically, no unbiased estimator is more accurate than the mle.

EXAMPLE 11.7

Estimate the covariance matrix of the mle for the lognormal distribution. Then apply this result to Data Set B.

The likelihood function and its logarithm are

$$\begin{aligned}L(\mu,\sigma) &= \prod_{j=1}^{n}\frac{1}{x_j\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln x_j-\mu)^2}{2\sigma^2}\right],\\ l(\mu,\sigma) &= \sum_{j=1}^{n}\left[-\ln x_j-\ln\sigma-\frac{1}{2}\ln(2\pi)-\frac{1}{2}\left(\frac{\ln x_j-\mu}{\sigma}\right)^2\right].\end{aligned}$$

The first partial derivatives are

$$\frac{\partial l}{\partial\mu} = \sum_{j=1}^{n}\frac{\ln x_j-\mu}{\sigma^2} \text{ and } \frac{\partial l}{\partial\sigma} = -\frac{n}{\sigma}+\sum_{j=1}^{n}\frac{(\ln x_j-\mu)^2}{\sigma^3}.$$

The second partial derivatives are

$$\begin{aligned}\frac{\partial^2 l}{\partial\mu^2} &= -\frac{n}{\sigma^2},\\ \frac{\partial^2 l}{\partial\sigma\,\partial\mu} &= -2\sum_{j=1}^{n}\frac{\ln x_j-\mu}{\sigma^3},\\ \frac{\partial^2 l}{\partial\sigma^2} &= \frac{n}{\sigma^2}-3\sum_{j=1}^{n}\frac{(\ln x_j-\mu)^2}{\sigma^4}.\end{aligned}$$

[7]For any multivariate random variable, the covariance matrix has the variances of the individual random variables on the main diagonal and covariances in the off-diagonal positions.

The expected values are ($\ln X_j$ has a normal distribution with mean μ and standard deviation σ)

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial \mu^2}\right) = -\frac{n}{\sigma^2},$$
$$\mathrm{E}\left(\frac{\partial^2 l}{\partial \mu\, \partial \sigma}\right) = 0,$$
$$\mathrm{E}\left(\frac{\partial^2 l}{\partial \sigma^2}\right) = -\frac{2n}{\sigma^2}.$$

Changing the signs and inverting produce an estimate of the covariance matrix (it is an estimate because Theorem 11.4 only provides the covariance matrix in the limit). It is

$$\begin{bmatrix} \dfrac{\sigma^2}{n} & 0 \\ 0 & \dfrac{\sigma^2}{2n} \end{bmatrix}.$$

For the lognormal distribution, the maximum likelihood estimates are the solutions to the two equations

$$\sum_{j=1}^{n} \frac{\ln x_j - \mu}{\sigma^2} = 0 \text{ and } -\frac{n}{\sigma} + \sum_{j=1}^{n} \frac{(\ln x_j - \mu)^2}{\sigma^3} = 0.$$

From the first equation $\hat{\mu} = (1/n)\sum_{j=1}^{n} \ln x_j$, and from the second equation, $\hat{\sigma}^2 = (1/n)\sum_{j=1}^{n} (\ln x_j - \hat{\mu})^2$. For Data Set B, the values are $\hat{\mu} = 6.1379$ and $\hat{\sigma}^2 = 1.9305$, or $\hat{\sigma} = 1.3894$. With regard to the covariance matrix, the true values are needed. The best we can do is substitute the estimated values to obtain

$$\widehat{\mathrm{Var}}(\hat{\mu}, \hat{\sigma}) = \begin{bmatrix} 0.0965 & 0 \\ 0 & 0.0483 \end{bmatrix}. \tag{11.1}$$

The multiple "hats" in the expression indicate that this is an estimate of the variance of the estimators. □

The zeros off the diagonal indicate that the two parameter estimators are asymptotically uncorrelated. For the particular case of the lognormal distribution, the estimators are uncorrelated for any sample size. One thing we could do with this information is construct approximate 95% confidence intervals for the true parameter values. These would be 1.96 standard deviations on either side of the estimate:

$$\mu\text{:} \quad 6.1379 \pm 1.96(0.0965)^{1/2} = 6.1379 \pm 0.6089,$$
$$\sigma\text{:} \quad 1.3894 \pm 1.96(0.0483)^{1/2} = 1.3894 \pm 0.4308.$$

To obtain the information matrix, it is necessary to take both derivatives and expected values, which is not always easy to do. A way to avoid this problem is to simply not take the expected value. Rather than working with the number that results from the expectation, use the observed data points. The result is called the **observed information**.

EXAMPLE 11.8

Estimate the covariance in Example 11.7 using the observed information.

Substituting the observations into the second derivatives produces

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} = -\frac{20}{\sigma^2},$$

$$\frac{\partial^2 l}{\partial \sigma\, \partial \mu} = -2\sum_{j=1}^{n} \frac{\ln x_j - \mu}{\sigma^3} = -2\frac{122.7576 - 20\mu}{\sigma^3},$$

$$\frac{\partial^2 l}{\partial \sigma^2} = \frac{n}{\sigma^2} - 3\sum_{j=1}^{n} \frac{(\ln x_j - \mu)^2}{\sigma^4} = \frac{20}{\sigma^2} - 3\frac{792.0801 - 245.5152\mu + 20\mu^2}{\sigma^4}.$$

Inserting the parameter estimates produces the negatives of the entries of the observed information,

$$\frac{\partial^2 l}{\partial \mu^2} = -10.3600, \quad \frac{\partial^2 l}{\partial \sigma\, \partial \mu} = 0, \quad \frac{\partial^2 l}{\partial \sigma^2} = -20.7190.$$

Changing the signs and inverting produce the same values as in (11.1). This is a feature of the lognormal distribution that need not hold for other models. □

Sometimes it is not even possible to take the derivative. In that case, an approximate second derivative can be used. A reasonable approximation is

$$\frac{\partial^2 f(\theta)}{\partial \theta_i\, \partial \theta_j} \doteq \frac{1}{h_i h_j}[f(\theta + \tfrac{1}{2}h_i\mathbf{e}_i + \tfrac{1}{2}h_j\mathbf{e}_j) - f(\theta + \tfrac{1}{2}h_i\mathbf{e}_i - \tfrac{1}{2}h_j\mathbf{e}_j) - f(\theta - \tfrac{1}{2}h_i\mathbf{e}_i + \tfrac{1}{2}h_j\mathbf{e}_j) + f(\theta - \tfrac{1}{2}h_i\mathbf{e}_i - \tfrac{1}{2}h_j\mathbf{e}_j)],$$

where $\mathbf{e}_i$ is a vector with all zeros except for a 1 in the ith position and $h_i = \theta_i/10^v$, where v is one-third the number of significant digits used in calculations.

EXAMPLE 11.9

Repeat Example 11.8 using approximate derivatives.

Assume that there are 15 significant digits being used. Then $h_1 = 6.1379/10^5$ and $h_2 = 1.3894/10^5$. Reasonably close values are 0.00006 and 0.00001. The first approximation is

$$\frac{\partial^2 l}{\partial \mu^2} \doteq \frac{l(6.13796, 1.3894) - 2l(6.1379, 1.3894) + l(6.13784, 1.3894)}{(0.00006)^2}$$

$$= \frac{-157.71389308198 - 2(-157.71389304968) + (-157.71389305468)}{(0.00006)^2}$$

$$= -10.3604.$$

The other two approximations are

$$\frac{\partial^2 l}{\partial \sigma\, \partial \mu} \doteq 0.0003 \text{ and } \frac{\partial^2 l}{\partial \sigma^2} \doteq -20.7208.$$

We see that here the approximation works very well. □

11.5.1 Exercises

11.33 Determine 95% confidence intervals for the parameters of exponential and gamma models for Data Set B. The likelihood function and maximum likelihood estimates were determined in Example 11.2.

11.34 Let X have a uniform distribution on the interval from 0 to θ. Show that the maximum likelihood estimator is $\hat{\theta} = \max(X_1, \ldots, X_n)$. Use Examples 10.5 and 10.9 to show that this estimator is asymptotically unbiased and to obtain its variance. Show that Theorem 11.4 yields a negative estimate of the variance and that item (ii) in the conditions does not hold.

11.35 (*) A distribution has two parameters, α and β. A sample of size 10 produced the following loglikelihood function:

$$l(\alpha, \beta) = -2.5\alpha^2 - 3\alpha\beta - \beta^2 + 50\alpha + 2\beta + k,$$

where k is a constant. Estimate the covariance matrix of the mle $(\hat{\alpha}, \hat{\beta})$.

11.36 (*) A sample of size 40 has been taken from a population with pdf

$$f(x) = (2\pi\theta)^{-1/2} e^{-x^2/(2\theta)}, \ -\infty < x < \infty, \ \theta > 0.$$

The mle of θ is $\hat{\theta} = 2$. Approximate the MSE of $\hat{\theta}$.

11.37 Four observations were made from a random variable having the density function $f(x) = 2\lambda x e^{-\lambda x^2}, \ x, \lambda > 0$. Exactly one of the four observations was less than 2.

(a) (*) Determine the mle of λ.

(b) Approximate the variance of the mle of λ.

11.38 Consider a random sample of size n from a Weibull distribution. For this exercise, write the Weibull survival function as

$$S(x) = \exp\left\{-\left[\frac{\Gamma(1+\tau^{-1})x}{\mu}\right]^{\tau}\right\}.$$

For this exercise, assume that τ is known and that only μ is to be estimated.

(a) Show that $\mathrm{E}(X) = \mu$.

(b) Show that the maximum likelihood estimate of μ is

$$\hat{\mu} = \Gamma(1+\tau^{-1}) \left(\frac{1}{n}\sum_{j=1}^{n} x_j^{\tau}\right)^{1/\tau}.$$

(c) Show that using the observed information produces the variance estimate

$$\hat{Var}(\hat{\mu}) = \frac{\hat{\mu}^2}{n\tau^2},$$

where μ is replaced by $\hat{\mu}$.

(d) Show that using the information (again replacing μ with $\hat{\mu}$) produces the same variance estimate as in part (c).

(e) Show that $\hat{\mu}$ has a transformed gamma distribution with $\alpha = n, \theta = \mu n^{-1/\tau}$, and $\tau = \tau$. Use this result to obtain the exact variance of $\hat{\mu}$ (as a function of μ). *Hint*: The variable X^τ has an exponential distribution, and so the variable $\sum_{j=1}^{n} X_j^\tau$ has a gamma distribution with first parameter equal to n and second parameter equal to the mean of the exponential distribution.

11.6 Functions of aymptotically normal estimators

We are often more interested in a quanity that is a function of the parameters. For example, we might be interested in the lognormal mean as an estimate of the population mean. That is, we want to use $\exp(\hat{\mu} + \hat{\sigma}^2/2)$ as an estimate of the population mean, where the maximum likelihood estimates of the parameters are used. It not trivial to evaluate the mean and variance of this random variable because it is a function of two variables with different distributions. The following theorem (from [16]) provides an approximate solution. The method is often called the *delta method*.

Theorem 11.5 *Let $\mathbf{X}_n = (X_{1n}, \ldots, X_{kn})^T$ be a multivariate random variable of dimension k based on a sample of size n. Assume that $\mathbf{X}$ is asymptotically normal with mean θ and covariance matrix $\boldsymbol{\Sigma}/n$, where neither θ nor $\boldsymbol{\Sigma}$ depend on n. Let g be a function of k variables that is totally differentiable. Let $G_n = g(X_{1n}, \ldots, X_{kn})$. Then G_n is asymptotically normal with mean $g(\theta)$ and variance $(\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A})/n$, where $\mathbf{A}$ is the vector of first derivatives of g, that is, $\mathbf{A} = (\partial g/\partial\theta_1, \ldots, \partial g/\partial\theta_k)^T$ and it is to be evaluated at θ, the true parameters of the original random variable.*

The statement of the theorem is hard to decipher. The Xs are the estimators and g is the function of the parameters that are being estimated. For a model with one parameter, the theorem reduces to the following statement: Let $\hat{\theta}$ be an estimator of θ that has an asymptotic normal distribution with mean θ and variance $\text{Var}(\hat{\theta})$. Then $g(\hat{\theta})$ has an asymptotic normal distribution with mean $g(\theta)$ and asymptotic variance $[g'(\theta)]\text{Var}(\hat{\theta})[g'(\theta)] = g'(\theta)^2\text{Var}(\hat{\theta})$. To obtain a useful numerical result, any parameters that appear will usually be replaced by an estimate.

Note that the theorem does not specify the type of estimator used. All that is required is that the estimator have an asymptotic multivariate normal distribution. For maximum likelihood estimators, under the usual regularity conditions, this holds and the information matrix can be used to estimate the asymptotic variance.

EXAMPLE 11.10

Use the delta method to approximate the variance of the mle of the probability that an observation from an exponential distribution exceeds 200. Apply this result to Data Set B.

From Example 11.2 we know that the maximum likelihood estimate of the exponential parameter is the sample mean. We are asked to estimate $p = \Pr(X > 200) = \exp(-200/\theta)$. The maximum likelihood estimate is $\hat{p} = \exp(-200/\hat{\theta}) = \exp(-200/\bar{x})$. Determining the mean and variance of this quantity is not easy. But

we do know that $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n = \theta^2/n$. Furthermore,

$$g(\theta) = e^{-200/\theta}, \quad g'(\theta) = 200\theta^{-2}e^{-200/\theta},$$

and therefore the delta method gives

$$\mathrm{Var}(\hat{p}) \doteq \frac{(200\theta^{-2}e^{-200/\theta})^2\theta^2}{n} = \frac{40{,}000\theta^{-2}e^{-400/\theta}}{n}.$$

For Data Set B,

$$\begin{aligned}
\bar{x} &= 1{,}424.4,\\
\hat{p} &= \exp\left(-\frac{200}{1{,}424.4}\right) = 0.86900,\\
\widehat{\mathrm{Var}}(\hat{p}) &= \frac{40{,}000(1,424.4)^{-2}\exp(-400/1{,}424.4)}{20} = 0.0007444.
\end{aligned}$$

A 95% confidence interval for p is $0.869 \pm 1.96\sqrt{0.0007444}$, or 0.869 ± 0.053. □

EXAMPLE 11.11

Construct a 95% confidence interval for the mean of a lognormal population using Data Set B. Compare this to the more traditional confidence interval based on the sample mean.

From Example 11.7 we have $\hat{\mu} = 6.1379$, $\hat{\sigma} = 1.3894$, and an estimated covariance matrix of

$$\frac{\hat{\Sigma}}{n} = \begin{bmatrix} 0.0965 & 0 \\ 0 & 0.0483 \end{bmatrix}.$$

The function is $g(\mu, \sigma) = \exp(\mu + \sigma^2/2)$. The partial derivatives are

$$\begin{aligned}
\frac{\partial g}{\partial \mu} &= \exp\left(\mu + \tfrac{1}{2}\sigma^2\right),\\
\frac{\partial g}{\partial \sigma} &= \sigma\exp\left(\mu + \tfrac{1}{2}\sigma^2\right),
\end{aligned}$$

and the estimates of these quantities are 1,215.75 and 1,689.16, respectively. The delta method produces the following approximation:

$$\begin{aligned}
\widehat{\mathrm{Var}}[g(\hat{\mu}, \hat{\sigma})] &= \begin{bmatrix} 1{,}215.75 & 1{,}689.16 \end{bmatrix} \begin{bmatrix} 0.0965 & 0 \\ 0 & 0.0483 \end{bmatrix} \begin{bmatrix} 1{,}215.75 \\ 1{,}689.16 \end{bmatrix}\\
&= 280{,}444.
\end{aligned}$$

The confidence interval is $1{,}215.75 \pm 1.96\sqrt{280,444}$, or $1{,}215.75 \pm 1{,}037.96$.

The customary confidence interval for a population mean is $\bar{x} \pm 1.96s/\sqrt{n}$, where s^2 is the sample variance. For Data Set B the interval is $1,424.4 \pm 1.96(3{,}435.04)/\sqrt{20}$, or $1{,}424.4 \pm 1{,}505.47$. It is not surprising that this is a wider interval because we know that (for a lognormal population) the mle is asymptotically UMVUE. □

11.6.1 Exercises

11.39 Use the delta method to construct a 95% confidence interval for the mean of a gamma distribution using Data Set B. Preliminary calculations are in Exercise 11.33.

11.40 (*) For a lognormal distribution with parameters μ and σ, you are given that the maximum likelihood estimates are $\hat{\mu} = 4.215$ and $\hat{\sigma} = 1.093$. The estimated covariance matrix of $(\hat{\mu}, \hat{\sigma})$ is

$$\begin{bmatrix} 0.1195 & 0 \\ 0 & 0.0597 \end{bmatrix}.$$

The mean of a lognormal distribution is given by $\exp(\mu + \sigma^2/2)$. Estimate the variance of the maximum likelihood estimator of the mean of this lognormal distribution using the delta method.

11.41 This is a continuation of Exercise 11.6. Let $x_1, \ldots, x_n$ be a random sample from a population with cdf $F(x) = x^p,\ 0 < x < 1$.

(a) Determine the asymptotic variance of the mle of p.

(b) Use your answer to obtain a general formula for a 95% confidence interval for p.

(c) Determine the mle of $\mathrm{E}(X)$ and obtain its asymptotic variance and a formula for a 95% confidence interval.

11.42 This is a continuation of Exercise 11.9. Let $x_1, \ldots, x_n$ be a random sample from a population with pdf $f(x) = \theta^{-1}e^{-x/\theta},\ x > 0$.

(a) Determine the asymptotic variance of the mle of θ.

(b) (*) Use your answer to obtain a general formula for a 95% confidence interval for θ.

(c) Determine the mle of $\mathrm{Var}(X)$ and obtain its asymptotic variance and a formula for a 95% confidence interval.

11.43 (*) Losses have an exponential distribution. Five observations from this distribution are 100, 200, 400, 800, 1,400, and 3,100. Use the delta method to approximate the variance of the mle of $S(1{,}500)$. Then construct a symmetric 95% confidence interval for the true value.

11.44 Estimate the covariance matrix of the mles for the data in Exercise 11.7 with both α and θ unknown. Do so by computing approximate derivatives of the loglikelihood. Then construct a 95% confidence interval for the mean.

11.45 Estimate the variance of the mle for Exercise 11.12 and use it to construct a 95% confidence interval for $\mathrm{E}(X \wedge 500)$.

11.7 Nonnormal confidence intervals

Section 11.5 created confidence intervals based on two assumptions. The first was that the normal distribution is a reasonable approximation of the true distribution of the maximum likelihood estimator. We know this assumption is asymptotically true but may not hold for

small or even moderate samples. Second, it was assumed that when there is more than one parameter, separate confidence intervals should be constructed for each parameter. Separate intervals can be used in cases such as the lognormal distribution where the parameter estimates are independent, but in most cases that is not true. When there is high correlation, it is better to postulate a confidence region, which could be done using the asymptotic covariances and a multivariate normal distribution. However, there is an easier method that does not require a normal distribution assumption (though is still based on asymptotic results).

One way to motivate a confidence region is to consider the meaning of the likelihood function. The parameter value that maximizes this function is our best choice. It is reasonable that values of the parameter which produce likelihood function values close to the maximum are good alternative choices for the true parameter value. Thus, for some choice of c, a confidence region for the parameter might be

$$\{\theta : l(\theta) \geq c\},$$

the set of all parameters for which the loglikelihood exceeds c. The discussion of the likelihood ratio test in Section **??** confirms that the loglikelihood is the correct function to use and also indicates how c should be selected to produce a $100(1-\alpha)\%$ confidence region. The value is

$$c = l(\hat{\theta}) - 0.5\chi^2_\alpha,$$

where the first term is the loglikelihood value at the maximum likelihood estimate and the second term is the $1-\alpha$ percentile from the chi-square distribution with degrees of freedom equal to the number of estimated parameters.

EXAMPLE 11.12

Use this method to construct a 95% confidence interval for the parameter of an exponential distribution. Compare the answer to the normal approximation using Data Set B.

We know that $\hat{\theta} = \bar{x}$ and for a sample of size n, $l(\bar{x}) = -n - n\ln\bar{x}$. With one degree of freedom, the 95th percentile of the chi-square distribution is 3.84. The confidence region is

$$\left\{\theta : -\frac{n\bar{x}}{\theta} - n\ln\theta \geq -n - n\ln\bar{x} - 1.92\right\},$$

which must be evaluated numerically. For Data Set B, the equation is

$$-\frac{20(1{,}424.4)}{\theta} - 20\ln\theta \geq -20 - 20\ln(1{,}424.4) - 1.92,$$
$$-\frac{28{,}488}{\theta} - 20\ln\theta \geq -167.15,$$

and the solution is $946.85 \leq \theta \leq 2{,}285.05$.

For the normal approximation, the asymptotic variance of the mle is θ^2/n, which happens to be the true variance. Inserting sample values, the normal confidence interval is

$$1{,}424.4 \pm 1.96\sqrt{1{,}424.4^2/20},$$
$$1{,}424.4 \pm 624.27,$$

which is $800.14 \le \theta \le 2{,}048.76$. Note that the widths of the two intervals are similar, but the first one is not symmetric about the sample mean. This asymmetry is reasonable in that a sample of size 20 is unlikely to be large enough to have the sample mean remove the skewness of the underlying exponential distribution. □

The extension to two parameters is similar, as illustrated in Example 11.13.

EXAMPLE 11.13

In Example 11.2, the maximum likelihood estimates for a gamma model for Data Set B were $\hat{\alpha} = 0.55616$ and $\hat{\theta} = 2{,}561.1$. Determine a 95% confidence region for the true values.

The region consists of all pairs (α, θ) that satisfy

$$(\alpha - 1)\sum_{j=1}^{20} \ln x_j - \frac{1}{\theta}\sum_{j=1}^{20} x_j - 20\ln\Gamma(\alpha) - 20\alpha\ln\theta$$
$$\ge (0.55616 - 1)\sum_{j=1}^{20} \ln x_j - \frac{1}{2{,}561.1}\sum_{j=1}^{20} x_j - 20\ln\Gamma(0.55616)$$
$$- 20(0.55616)\ln 2{,}561.1 - 2.996 = -165.289,$$

where 2.996 is one-half of the 95th percentile of a chi-square distribution with two degrees of freedom. Figure 11.1 shows the resulting confidence region. If the normal approximation were appropriate, this region would be elliptical in shape. □

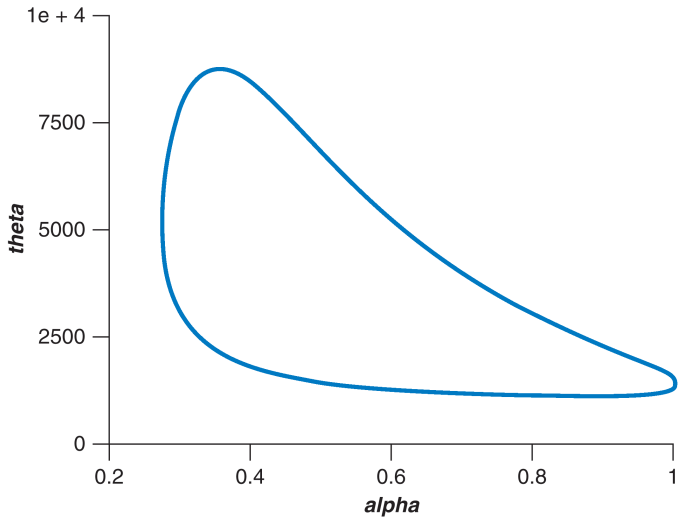

Figure 11.1 95% confidence region for gamma parameters.

For functions of parameters, the same method can be applied as illustrated in Example 11.14.

■ **EXAMPLE 11.14**

Determine a 95% confidence interval for the mean of the gamma distribution in Example 11.13.

First, reparameterize the gamma density so that the mean is a parameter, which can be done by setting $\mu = \alpha\theta$ and leaving α unchanged. The density function is now

$$f(x) = \frac{x^{\alpha-1}e^{-x\alpha/\mu}}{\Gamma(\alpha)(\mu/\alpha)^{\alpha}}.$$

Due to the invariance of mles, we have $\hat{\mu} = \hat{\alpha}\hat{\theta} = 1{,}424.4$. We then look for alternative μ values that produce loglikelihood values that are within 1.92 of the maximum (there is only one degree of freedom because there is only one parameter, the mean, being evaluated). When we try a μ-value, to give it the best chance to be accepted, the accompanying α-value should be the one that maximizes the likelihood given μ. Numerical maximizations and trial and error reveal the confidence interval $811 \le \mu \le 2{,}846$. □

11.7.1 Exercise

11.46 Use the method of this section to determine a 95% confidence interval for the probability that an observation exceeds 200 using the exponential model and Data Set B. Compare your answer to that from Example 11.10.

11.8 Solutions to Exercises

11.1 For the inverse exponential distribution,

$$l(\theta) = \sum_{j=1}^{n}(\ln\theta - \theta x_j^{-1} - 2\ln x_j) = n\ln\theta - ny\theta - 2\sum_{j=1}^{n}\ln x_j,$$

$$l'(\theta) = n\theta^{-1} - ny,\ \hat{\theta} = y^{-1},\ \text{where } y = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{x_j}.$$

For Data Set B, we have $\hat{\theta} = 197.72$ and the loglikelihood value is -159.78. Because the mean does not exist for the inverse exponential distribution, there is no traditional method-of-moments estimate available. However, it is possible to obtain a method-of-moments estimate using the negative first moment rather than the positive first moment. That is, equate the average reciprocal to the expected reciprocal:

$$\frac{1}{n}\sum_{j=1}^{n}\frac{1}{x_j} = \mathrm{E}(X^{-1}) = \theta^{-1}.$$

This special method of moments estimate is identical to the maximum likelihood estimate.

For the inverse gamma distribution with $\alpha = 2$,

$$f(x|\theta) = \frac{\theta^2 e^{-\theta/x}}{x^3\Gamma(2)},\ \ln f(x|\theta) = 2\ln\theta - \theta x^{-1} - 3\ln x,$$
$$l(\theta) = \sum_{j=1}^{n}(2\ln\theta - \theta x_j^{-1} - 3\ln x_j) = 2n\ln\theta - ny\theta - 3\sum_{j=1}^{n}\ln x_j,$$
$$l'(\theta) = 2n\theta^{-1} - ny,\ \hat{\theta} = 2/y.$$

For Data Set B, $\hat{\theta} = 395.44$ and the value of the loglikelihood function is -169.07. The method-of-moments estimate solves the equation

$$1,424.4 = \frac{\theta}{\alpha - 1} = \frac{\theta}{2-1},\ \hat{\theta} = 1,424.4,$$

which differs from the maximum likelihood estimate.

For the inverse gamma distribution with both parameters unknown,

$$f(x|\theta) = \frac{\theta^\alpha e^{-\theta/x}}{x^{\alpha+1}\Gamma(\alpha)},\ \ln f(x|\theta) = \alpha\ln\theta - \theta x^{-1} - (\alpha+1)\ln x - \ln\Gamma(\alpha).$$

The likelihood function must be maximized numerically. The answer is $\hat{\alpha} = 0.70888$ and $\hat{\theta} = 140.16$ and the loglikelihood value is -158.88. The method-of-moments estimate is the solution to the two equations

$$1{,}424.4 = \frac{\theta}{\alpha - 1},$$
$$13{,}238{,}441.9 = \frac{\theta^2}{(\alpha-1)(\alpha-2)}.$$

Squaring the first equation and dividing it into the second equation gives

$$6.52489 = \frac{\alpha - 1}{\alpha - 2},$$

which leads to $\hat{\alpha} = 2.181$ and then $\hat{\theta} = 1{,}682.2$. This result does not match the maximum likelihood estimate (which had to be the case because the mle produces a model that does not have a mean).

11.2 The density function is the derivative, $f(x) = \alpha 500^\alpha x^{-\alpha-1}$. The likelihood function is

$$L(\alpha) = \alpha^5 500^{5\alpha}(\Pi x_j)^{-\alpha-1},$$

and its logarithm is

$$l(\alpha) = 5\ln\alpha + (5\alpha)\ln 500 - (\alpha+1)\Sigma\ln x_j$$
$$= 5\ln\alpha + 31.073\alpha - 33.111(\alpha+1).$$

Setting the derivative equal to zero gives

$$0 = 5\alpha^{-1} - 2.038.$$

The estimate is $\hat{\alpha} = 5/2.038 = 2.45$.

11.3 The coefficient $(2\pi x)^{-1/2}$ is not relevant because it does not involve μ. The logarithm of the likelihood function is

$$l(\mu) = -\frac{1}{22}(11-\mu)^2 - \frac{1}{30.4}(15.2-\mu)^2 - \frac{1}{36}(18-\mu)^2 - \frac{1}{42}(21-\mu)^2 - \frac{1}{51.6}(25.8-\mu)^2.$$

The derivative is

$$l'(\mu) = \frac{11-\mu}{11} + \frac{15.2-\mu}{15.2} + \frac{18-\mu}{18} + \frac{21-\mu}{21} + \frac{25.8-\mu}{25.8} = 5 - 0.29863\mu.$$

Setting the derivative equal to zero yields $\hat{\mu} = 16.74$.

11.4 The density function is

$$f(x) = 0.5x^{-0.5}\theta^{-0.5}e^{-(x/\theta)^{0.5}}.$$

The likelihood function and subsequent calculations are

$$\begin{aligned}
L(\theta) &= \prod_{j=1}^{10} 0.5x_j^{-0.5}\theta^{-0.5}e^{-x_j^{0.5}\theta^{-0.5}} \propto \theta^{-5}\exp\left(-\theta^{-0.5}\sum_{j=1}^{10} x_j^{0.5}\right) \\
&= \theta^{-5}\exp(-488.97\theta^{-0.5}), \\
l(\theta) &= -5\ln\theta - 488.97\theta^{-0.5}, \\
l'(\theta) &= -5\theta^{-1} + 244.485\theta^{-1.5} = 0, \\
0 &= -5\theta^{0.5} + 244.485,
\end{aligned}$$

and so $\hat{\theta} = (244.485/5)^2 = 2{,}391$.

11.5 $L = 2^n\theta^n(\Pi x_j)\exp(-\theta\Sigma x_j^2)$, $l = n\ln 2 + n\ln\theta + \Sigma\ln x_j - \theta\Sigma x_j^2$, $l' = n\theta^{-1} - \Sigma x_j^2 = 0$, $\hat{\theta} = n/\Sigma x_j^2$.

11.6 $f(x) = px^{p-1}$, $L = p^n(\Pi x_j)^{p-1}$, $l = n\ln p + (p-1)\Sigma\ln x_j$, $l' = np^{-1} + \Sigma\ln x_j = 0$, $\hat{p} = -n/\Sigma\ln x_j$.

11.7 (a) $L = (\Pi x_j)^{\alpha-1}\exp(-\Sigma x_j/\theta)[\Gamma(\alpha)]^{-n}\theta^{-n\alpha}$.

$$\begin{aligned}
l &= (\alpha-1)\Sigma\ln x_j - \theta^{-1}\Sigma x_j - n\ln\Gamma(\alpha) - n\alpha\ln\theta \\
&= 81.61837(\alpha-1) - 38{,}000\theta^{-1} - 10\ln\Gamma(\alpha) - 10\alpha\ln\theta. \\
\partial l/\partial\theta &= 38{,}000\theta^{-2} - 10\alpha\theta^{-1} = 0,
\end{aligned}$$

$\hat{\theta} = 38{,}000/(10\cdot 12) = 316.67$.

(b) l is maximized at $\hat{\alpha} = 6.341$ and $\hat{\theta} = 599.3$ (with $l = -86.835$).

11.8 $\hat{\mu} = \frac{1}{5}\Sigma \ln x_i = 7.33429$. $\hat{\sigma}^2 = \frac{1}{5}\Sigma(\ln x_i)^2 - 7.33429^2 = 0.567405$, $\hat{\sigma} = 0.753263$.

$$\begin{aligned}\Pr(X > 4{,}500) &= 1 - \Phi[(\ln 4{,}500 - 7.33429)/0.753263] \\ &= 1 - \Phi(1.4305) = 0.076.\end{aligned}$$

11.9 $L = \theta^{-n}\exp(-\Sigma x_j/\theta)$. $l = -n\ln\theta - \theta^{-1}\Sigma x_j$. $l' = -n\theta^{-1} + \theta^{-2}\Sigma x_j = 0$. $\hat{\theta} = \Sigma x_j/n$.

11.10 $L = \beta^{-10}(\Pi x_j)\exp[-\Sigma x_j^2/(2\beta^2)]$. $l = -10\ln\beta + \Sigma\ln x_j - \Sigma x_j^2/(2\beta^2)$. $l' = -10\beta^{-1} + \beta^{-3}\Sigma x_j^2 = 0$. $\hat{\beta} = \sqrt{\Sigma x_j^2/10} = 3.20031$.

11.11 $f(x) = \alpha x^{-\alpha-1}$. $L = \alpha^n(\Pi x_j)^{-\alpha-1}$. $l = n\ln\alpha - (\alpha+1)\Sigma\ln x_j$. $l' = n\alpha^{-1} - \Sigma\ln x_j = 0$. $\hat{\alpha} = n/\Sigma\ln x_j$.

11.12

$$\begin{aligned}L &= \alpha^5\lambda^{5\alpha}[\Pi(\lambda + x_j)]^{-\alpha-1}, \\ l &= 5\ln\alpha + 5\alpha\ln 1{,}000 - (\alpha+1)\Sigma\ln(1{,}000 + x_j), \\ l' &= 5\alpha^{-1} + 34.5388 - 35.8331 = 0, \\ \hat{\alpha} &= 3.8629.\end{aligned}$$

11.13 (a) Three observations exceed 200. The empirical estimate is $3/20 = 0.15$.

(b) $\mathrm{E}(X) = 100\alpha/(\alpha-1) = \bar{x} = 154.5$, $\hat{\alpha} = 154.5/54.5 = 2.835$, $\Pr(X > 200) = (100/200)^{2.835} = 0.140$.

(c) $f(x) = \alpha 100^\alpha x^{-\alpha-1}$. $L = \alpha^{20}100^{20\alpha}(\Pi x_j)^{-\alpha-1}$.

$$\begin{aligned}l &= 20\ln\alpha + 20\alpha\ln 100 - (\alpha+1)\Sigma\ln x_j \\ l' &= 20\alpha^{-1} + 20\ln 100 - \Sigma\ln x_j = 0 \\ \hat{\alpha} &= 20/(\Sigma\ln x_j - 20\ln 100) = 20/(99.125 - 92.103) = 2.848.\end{aligned}$$

$\Pr(X > 200) = (100/200)^{2.848} = 0.139$.

11.14 (a)

$$\begin{aligned}\sum\frac{(x_j-\mu)^2}{x_j} &= \sum\left(x_j - 2\mu + \frac{\mu^2}{x_j}\right) \\ &= \sum\left(\frac{\mu^2}{x_j} - \frac{\mu^2}{\bar{x}}\right) + \sum\left(\frac{\mu^2}{\bar{x}} - 2\mu + x_j\right) \\ &= \mu^2\sum\left(\frac{1}{x_j} - \frac{1}{\bar{x}}\right) + \frac{n\mu^2}{\bar{x}} - 2n\mu + n\bar{x} \\ &= \mu^2\sum\left(\frac{1}{x_j} - \frac{1}{\bar{x}}\right) + \frac{n}{\bar{x}}(\bar{x}-\mu)^2.\end{aligned}$$

(b)

$$
\begin{aligned}
L &\propto \theta^{n/2} \exp\left[-\frac{\theta}{2\mu^2}\sum \frac{(x_j-\mu)^2}{x_j}\right],\\
l &= \ln L = \frac{n}{2}\ln\theta - \frac{\theta}{2\mu^2}\sum\frac{(x_j-\mu)^2}{x_j}\\
&= \frac{n}{2}\ln\theta - \frac{\theta}{2\mu^2}\left[\mu^2\sum\left(\frac{1}{x_j}-\frac{1}{\bar{x}}\right)+\frac{n}{\bar{x}}(\bar{x}-\mu)^2\right]\\
&= \frac{n}{2}\ln\theta - \frac{\theta}{2}\sum\left(\frac{1}{x_j}-\frac{1}{\bar{x}}\right) - \frac{n\theta}{2\mu^2\bar{x}}(\bar{x}-\mu)^2,\\
\frac{\partial l}{\partial \mu} &= -\frac{n\theta}{2\bar{x}}\frac{-\mu^2 2(\bar{x}-\mu)-(\bar{x}-\mu)^2 2\mu}{\mu^4} = 0,\\
\hat{\mu} &= \bar{x}.
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \theta} &= \frac{n}{2\theta} - \frac{1}{2}\sum\left(\frac{1}{x_j}-\frac{1}{\bar{x}}\right)+\frac{n}{2\mu^2\bar{x}}(\bar{x}-\mu)^2 = 0,\\
\hat{\theta} &= \frac{n}{\sum\left(\frac{1}{x_j}-\frac{1}{\bar{x}}\right)}.
\end{aligned}
$$

11.15

$$
\begin{aligned}
L(\mu,\theta) &= \prod_{j=1}^{n} f\left[x_j;\mu,(\theta m_j)^{-1}\right]\\
&= \prod_{j=1}^{n}\left(\frac{2\pi}{\theta m_j}\right)^{-\frac{1}{2}}\exp\left[-\frac{(x_j-\mu)^2 m_j\theta}{2}\right]\\
&\propto \theta^{n/2}\exp\left[-\frac{\theta}{2}\sum m_j(x_j-\mu)^2\right].\ell(\mu,\theta) = \frac{n}{2}\ln\theta - \frac{\theta}{2}\sum m_j(x_j-\mu)^2 + \text{constant},\\
\frac{\partial \ell}{\partial \mu} &= \theta\sum m_j(x_j-\mu) = 0 \Rightarrow \hat{\mu} = \frac{\sum m_j x_j}{\sum m_j} = \frac{\sum m_j x_j}{m},\\
\frac{\partial^2 \ell}{\partial \mu} &= -\theta\sum m_j < 0, \text{ hence, maximum.}\\
\frac{\partial \ell}{\partial \theta} &= \frac{n}{2}\frac{1}{\theta} - \frac{1}{2}\sum m_j(x_j-\mu)^2 = 0 \Rightarrow \hat{\theta}^{-1} = \frac{1}{n}\sum m_j(x_j-\hat{\mu})^2,\\
\hat{\theta} &= n\left[\sum m_j(x_j-\overline{x})^2\right]^{-1},\\
\frac{\partial^2 \ell}{\partial \theta^2} &= -\frac{n}{2}\frac{1}{\theta^2} < 0, \text{ hence, maximum.}
\end{aligned}
$$

11.16

$$L(\theta) = \prod_{j=1}^{n} f(x_j;\theta) = \prod_{j=1}^{n} \frac{p(x_j)e^{r(\theta)x_j}}{q(\theta)},$$
$$\ell(\theta) = \sum_{j=1}^{n} \ln p(x_j) + r(\theta)\sum_{j=1}^{n} x_j - n\ln q(\theta),$$
$$\ell'(\theta) = r'(\theta)\sum_{j=1}^{n} x_j - n\frac{q'(\theta)}{q(\theta)} = 0.$$

Therefore,

$$\frac{q'(\hat{\theta})}{r'(\hat{\theta})q(\hat{\theta})} = \overline{x}.$$

But,

$$\frac{q'(\theta)}{r'(\theta)q(\theta)} = \mathrm{E}(X) = \mu(\theta),$$

and so $\mu(\hat{\theta}) = \overline{x}$.

11.17 For the inverse exponential distribution, the cdf is $F(x) = e^{-\theta/x}$. Numerical maximization yields $\hat{\theta} = 6,662.39$ and the value of the loglikelihood function is -365.40. For the gamma distribution, the cdf requires numerical evaluation. In Excel® the function GAMMADIST(x, α, θ,true) can be used. The estimates are $\hat{\alpha} = 0.37139$ and $\hat{\theta} = 83,020$. The value of the loglikelihood function is -360.50. For the inverse gamma distribution, the cdf is available in Excel® as $1-$ GAMMADIST($1/x, \alpha, 1/\theta$,true). The estimates are $\hat{\alpha} = 0.83556$ and $\hat{\theta} = 5,113$. The value of the loglikelihood function is -363.92.

11.18

$$L(\theta) = \left(1-\frac{\theta}{10}\right)^9\left(\frac{\theta}{10}-\frac{\theta}{25}\right)^6\left(\frac{\theta}{25}\right)^5 \propto (10-\theta)^9\theta^{11},$$
$$l(\theta) = 9\ln(10-\theta) + 11\ln(\theta),$$
$$l'(\theta) = -\frac{9}{10-\theta} + \frac{11}{\theta} = 0,$$
$$9\theta = 11(10-\theta),$$
$$\theta = 110/20 = 5.5.$$

11.19 The maximum likelihood estimate is $\hat{\theta} = 93.188$.

11.20 In each case the likelihood function is $f(27)f(82)\cdots f(243)[1-F(250)]^{13}$. Table 11.6 provides the estimates for both the original and censored data sets. The censoring tended to disguise the true nature of these numbers and, in general, had a large impact on the estimates.

11.21 The calculations are done as in Example 11.4, but with θ unknown. The likelihood must be numerically maximized. For the shifted data, the estimates are $\hat{\alpha} = 1.4521$ and

Table 11.6 Estimatees for Exercise 11.20.

Model	Original	Censored
Exponential	$\hat{\theta} = 1,424.4$	$\hat{\theta} = 594.14$
Gamma	$\hat{\alpha} = 0.55616, \hat{\theta} = 2,561.1$	$\hat{\alpha} = 1.5183, \hat{\theta} = 295.69$
Inv. exponential	$\hat{\theta} = 197.72$	$\hat{\theta} = 189.78$
Inv. gamma	$\hat{\alpha} = 0.70888, \hat{\theta} = 140.16$	$\hat{\alpha} = 0.41612, \hat{\theta} = 86.290$

$\hat{\theta} = 907.98$. The two expected costs are $907.98/0.4521 = 2{,}008$ and $1{,}107.98/0.4521 = 2,451$ for the 200 and 400 deductibles, respectively. For the unshifted data, the estimates are $\hat{\alpha} = 1.4521$ and $\hat{\theta} = 707.98$. The three expected costs are $707.98/0.4521 = 1,566$, 2,008, and 2, 451 for the 0, 200, and 400 deductibles, respectively. While it is always the case that for the Pareto distribution the two approaches produce identical answers, that will not be true in general.

11.22 Table 11.6 can be used. The only difference is that observations that were surrenders are now treated as x-values and deaths are treated as y-values. Observations that ended at 5.0 continue to be treated as y-values. Once again there is no estimate for a Pareto model. The gamma parameter estimates are $\hat{\alpha} = 1.229$ and $\hat{\theta} = 6.452$.

11.23 The contribution to the likelihood for the first five values (number of drivers having zero through four accidents) is unchanged. However, for the last seven drivers, the contribution is

$$[\Pr(X \geq 5)]^7 = [1 - p(0) - p(1) - p(2) - p(3) - p(4)]^7,$$

and the maximum must be obtained numerically. The estimated values are $\hat{\lambda} = 0.16313$ and $\hat{q} = 0.02039$. These answers are similar to those for Example 11.6 because the probability of six or more accidents is so small.

11.24 We have

$$\begin{aligned}
L &= f(1{,}100)f(3{,}200)f(3{,}300)f(3{,}500)f(3{,}900)[S(4{,}000)]^{495} \\
&= \theta^{-1}e^{-1{,}100/\theta}\theta^{-1}e^{-3{,}200/\theta}\theta^{-1}e^{-3{,}300/\theta}\theta^{-1}e^{-3{,}500/\theta} \\
&\quad \times \theta^{-1}e^{-3{,}900/\theta}[e^{-4{,}000/\theta}]^{495} \\
&= \theta^{-5}e^{-1{,}995{,}000/\theta}, \\
\ln L &= -5\ln\theta - \frac{1{,}995{,}000}{\theta}, \\
\frac{d\ln L}{d\theta} &= -\frac{5}{\theta} + \frac{1{,}995{,}000}{\theta^2} = 0
\end{aligned}$$

and the solution is $\hat{\theta} = 1{,}995{,}000/5 = 399{,}000$.

11.25 The survival function is

$$S(t) = \begin{cases} e^{-t\lambda_1}, & 0 < t < 5, \\ e^{-5\lambda_1-(t-5)\lambda_2}, & 5 \le t < 10, \\ e^{-5\lambda_1-5\lambda_2-(t-10)\lambda_3}, & t \ge 10 \end{cases}$$

and the density function is

$$f(t) = -S'(t) = \begin{cases} \lambda_1 e^{-t\lambda_1}, & 0 < t < 5, \\ \lambda_2 e^{-5\lambda_1-(t-5)\lambda_2}, & 5 \le t < 10, \\ \lambda_3 e^{-5\lambda_1-5\lambda_2-(t-10)\lambda_3}, & t \ge 10. \end{cases}$$

The likelihood function and its logarithm are

$$\begin{aligned} L(\lambda_1,\lambda_2,\lambda_3) &= \lambda_1 e^{-3\lambda_1}\lambda_2^2 e^{-10\lambda_1-5\lambda_2}\lambda_3^4 e^{-20\lambda_1-20\lambda_2-11\lambda_3} \\ &\quad \times (e^{-5\lambda_1-5\lambda_2-5\lambda_3})^3 \\ &= \lambda_1 e^{-48\lambda_1}\lambda_2^2 e^{-40\lambda_2}\lambda_3^4 e^{-26\lambda_3}, \\ \ln L(\lambda_1,\lambda_2,\lambda_3) &= \ln\lambda_1 - 48\lambda_1 + 2\ln\lambda_2 - 40\lambda_2 + 4\ln\lambda_3 - 26\lambda_3. \end{aligned}$$

The partial derivative with respect to λ_1 is $\lambda_1^{-1} - 48 = 0$ for $\hat{\lambda}_1 = 1/48$. Similarly, $\hat{\lambda}_2 = 2/40$ and $\hat{\lambda}_3 = 4/26$.

11.26 The distribution and density function are

$$F(x) = 1 - e^{-(x/\theta)^2}, \; f(x) = \frac{2x}{\theta^2}e^{-(x/\theta)^2}.$$

The likelihood function is

$$\begin{aligned} L(\theta) &= f(20)f(30)f(45)[1-F(50)]^2 \\ &\propto \theta^{-2}e^{-(20/\theta)^2}\theta^{-2}e^{-(30/\theta)^2}\theta^{-2}e^{-(45/\theta)^2}\left[e^{-(50/\theta)^2}\right]^2 \\ &= \theta^{-6}e^{-8{,}325/\theta^2}. \end{aligned}$$

The logarithm and derivative are

$$\begin{aligned} l(\theta) &= -6\ln\theta - 8{,}325\theta^{-2}, \\ l'(\theta) &= -6\theta^{-1} + 16{,}650\theta^{-3}. \end{aligned}$$

Setting the derivative equal to zero yields $\hat{\theta} = (16{,}650/6)^{1/2} = 52.68$.

11.27 For the exponential distribution, the maximum likelihood estimate is the sample mean, and so $\bar{x}_P = 1000$ and $\bar{x}_S = 1500$. The likelihood with the restriction is (using i to index observations from Phil's bulbs and j to index observations from Sylvia's bulbs)

$$
\begin{aligned}
L(\theta^*) &= \prod_{i=1}^{20}(\theta^*)^{-1}\exp(-x_i/\theta^*)\prod_{j=1}^{10}(2\theta^*)^{-1}\exp(-x_j/2\theta^*)\\
&\propto (\theta^*)^{-30}\exp\left(-\sum_{i=1}^{20}\frac{x_i}{\theta^*}-\sum_{j=1}^{10}\frac{x_j}{2\theta^*}\right)\\
&= (\theta^*)^{-30}\exp(-20\bar{x}_P/\theta^* - 10\bar{x}_S/2\theta^*)\\
&= (\theta^*)^{-30}\exp(-20{,}000/\theta^* - 7{,}500/\theta^*).
\end{aligned}
$$

Taking logarithms and differentiating yields

$$
\begin{aligned}
l(\theta^*) &= -30\ln\theta^* - 27{,}500/\theta^*,\\
l'(\theta^*) &= -30(\theta^*)^{-1} + 27{,}500(\theta^*)^{-2}.
\end{aligned}
$$

Setting the derivative equal to zero gives $\hat{\theta}^* = 27{,}500/30 = 916.67$.

11.28 For the first part,

$$
\begin{aligned}
L(\theta) &= F(1{,}000)^{62}[1-F(1{,}000)]^{38}\\
&= (1-e^{-1{,}000/\theta})^{62}(e^{-1{,}000/\theta})^{38}.
\end{aligned}
$$

Let $x = e^{-1{,}000/\theta}$. Then

$$
\begin{aligned}
L(x) &= (1-x)^{62}x^{38},\\
l(x) &= 62\ln(1-x) + 38\ln x,\\
l'(x) &= -\frac{62}{1-x}+\frac{38}{x}.
\end{aligned}
$$

Setting the derivative equal to zero yields

$$
\begin{aligned}
0 &= -62x + 38(1-x)\\
&= 38 - 100x,\\
x &= 0.38,
\end{aligned}
$$

and then $\hat{\theta} = -1{,}000/\ln 0.38 = 1{,}033.50$.

With additional information,

$$
\begin{aligned}
L(\theta) &= \left[\prod_{j=1}^{62} f(x_j)\right][1-F(1{,}000)]^{38} = \left(\prod_{j=1}^{62}\theta^{-1}e^{-x_j/\theta}\right)e^{-38{,}000/\theta}\\
&= \theta^{-62}e^{-28{,}140/\theta}e^{-38{,}000/\theta} = \theta^{-62}e^{-66{,}140/\theta},\\
l(\theta) &= -62\ln\theta - 66{,}140/\theta,\\
l'(\theta) &= -62/\theta + 66{,}140/\theta^2 = 0,\\
0 &= -62\theta + 66{,}140,\\
\hat{\theta} &= 66{,}140/62 = 1{,}066.77.
\end{aligned}
$$

Table 11.7 Likelihood contributions for Exercise 11.29.

Observation	Probability	Loglikelihood
1997-1	$\frac{\Pr(N=1)}{\Pr(N=1)+\Pr(N=2)} = \frac{(1-p)p}{(1-p)p+(1-p)p^2} = \frac{1}{1+p}$	$-3\ln(1+p)$
1997-2	$\frac{\Pr(N=2)}{\Pr(N=1)+\Pr(N=2)} = \frac{(1-p)p^2}{(1-p)p+(1-p)p^2} = \frac{p}{1+p}$	$\ln p - \ln(1+p)$
1998-0	$\frac{\Pr(N=0)}{\Pr(N=0)+\Pr(N=1)} = \frac{(1-p)}{(1-p)+(1-p)p} = \frac{1}{1+p}$	$-5\ln(1+p)$
1998-1	$\frac{\Pr(N=1)}{\Pr(N=0)+\Pr(N=1)} = \frac{(1-p)p}{(1-p)+(1-p)p} = \frac{p}{1+p}$	$2\ln p - 2\ln(1+p)$
1999-0	$\frac{\Pr(N=0)}{\Pr(N=0)} = 1$	0

11.29 Each observation has a uniquely determined conditional probability. The contribution to the loglikelihood is given in Table 11.7. The total is $l(p) = 3\ln p - 11\ln(1+p)$. Setting the derivative equal to zero gives $0 = l'(p) = 3p^{-1} - 11(1+p)^{-1}$, and the solution is $\hat{p} = 3/8$.

11.30

$$
\begin{aligned}
L(w) &= \frac{\frac{1}{w}\frac{1}{w}\frac{1}{w}\left(\frac{w-4-p}{w}\right)^2}{\left(\frac{w-4}{w}\right)^5} = \frac{(w-4-p)^2}{(w-4)^5},\\
l(w) &= 2\ln(w-4-p) - 5\ln(w-4),\\
l'(w) &= \frac{2}{w-4-p} - \frac{5}{w-4} = 0,\\
0 &= l'(29) = \frac{2}{25-p} - \frac{5}{25},\\
25-p &= 10,\ p = 15.
\end{aligned}
$$

11.31 The density function is $f(x) = \theta x^{-2}e^{-\theta/x}$. The likelihood function is

$$
\begin{aligned}
L(\theta) &= \theta(66^{-2})e^{-\theta/66}\theta(91^{-2})e^{-\theta/91}\theta(186^{-2})e^{-\theta/186}(e^{-\theta/66})^7\\
&\propto \theta^3 e^{-0.148184\theta},\\
l(\theta) &= 3\ln\theta - 0.148184\theta,\\
l'(\theta) &= 3\theta^{-1} - 0.148184 = 0,\\
\theta &= 3/0.148184 = 20.25.
\end{aligned}
$$

The mode is $\theta/2 = 10.125$.

11.32

$$
\begin{aligned}
L(\alpha) &= \prod_{j=1}^{7} \frac{f(x_j|\alpha)}{1-F(100|\alpha)} = \frac{\prod_{j=1}^{7} \frac{\alpha 400^{\alpha}}{(400+x_j)^{\alpha+1}}}{\left[\left(\frac{400}{400+100}\right)^{\alpha}\right]^7}, \\
l(\alpha) &= 7\ln\alpha + 7\alpha\ln 400 - (\alpha+1)\sum_{j=1}^{7}\ln(400+x_j) - 7\alpha\ln 0.8 \\
&= 7\ln\alpha - 3.79\alpha - 47.29, \\
l'(\alpha) &= 7\alpha^{-1} - 3.79 = 0, \\
\alpha &= 7/3.79 = 1.847.
\end{aligned}
$$

11.33 In general, for the exponential distribution,

$$
\begin{aligned}
l'(\theta) &= -n\theta^{-1} + n\bar{x}\theta^{-2}, \\
l''(\theta) &= n\theta^{-2} - 2n\bar{x}\theta^{-3}, \\
\mathrm{E}[l''(\theta)] &= n\theta^{-2} - 2n\theta\theta^{-3} = -n\theta^{-2}, \\
\mathrm{Var}(\hat{\theta}) &= \theta^2/n,
\end{aligned}
$$

where the third line follows from $\mathrm{E}(\bar{X}) = \mathrm{E}(X) = \theta$. The estimated variance for Data Set B is $\widehat{\mathrm{Var}}(\hat{\theta}) = 1{,}424.4^2/20 = 101{,}445.77$ and the 95% confidence interval is $1{,}424.4 \pm 1.96(101{,}445.77)^{1/2}$, or $1{,}424.4 \pm 624.27$. Note that in this particular case, Theorem 11.4 gives the exact value of the variance, because

$$
\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n = \theta^2/n.
$$

For the gamma distribution,

$$
\begin{aligned}
l(\alpha,\theta) &= (\alpha-1)\sum_{j=1}^{n}\ln x_j - \sum_{j=1}^{n} x_j\theta^{-1} - n\ln\Gamma(\alpha) - n\alpha\ln\theta, \\
\frac{\partial l(\alpha,\theta)}{\partial\alpha} &= \sum_{j=1}^{n}\ln x_j - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} - n\ln\theta, \\
\frac{\partial l(\alpha,\theta)}{\partial\theta} &= \sum_{j=1}^{n} x_j\theta^{-2} - n\alpha\theta^{-1}, \\
\frac{\partial^2 l(\alpha,\theta)}{\partial\alpha^2} &= -n\frac{\Gamma(\alpha)\Gamma''(\alpha) - \Gamma'(\alpha)^2}{\Gamma(\alpha)^2}, \\
\frac{\partial^2 l(\alpha,\theta)}{\partial\alpha\partial\theta} &= -n\theta^{-1}, \\
\frac{\partial^2 l(\alpha,\theta)}{\partial\theta^2} &= -2\sum_{j=1}^{n} x_j\theta^{-3} + n\alpha\theta^{-2} = -2n\bar{x}\theta^{-3} + n\alpha\theta^{-2}.
\end{aligned}
$$

The first two second partial derivatives do not contain x_j, and so the expected value is equal to the indicated quantity. For the final second partial derivative, $\mathrm{E}(\bar{X}) = \mathrm{E}(X) = \alpha\theta$.

Therefore,

$$I(\alpha,\theta)=\begin{bmatrix} n\frac{\Gamma(\alpha)\Gamma''(\alpha)-\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} & n\theta^{-1} \\ n\theta^{-1} & n\alpha\theta^{-2}\end{bmatrix}.$$

The derivatives of the gamma function are available in some better computer packages, but are not available in Excel®. Using numerical derivatives of the gamma function yields

$$I(\hat{\alpha},\hat{\theta})=\begin{bmatrix} 82.467 & 0.0078091 \\ 0.0078091 & 0.0000016958\end{bmatrix},$$

and the covariance matrix is

$$\begin{bmatrix} 0.021503 & -99.0188 \\ -99.0188 & 1{,}045{,}668\end{bmatrix}.$$

Numerical second derivatives of the likelihood function (using $h_1 = 0.00005$ and $h_2 = 0.25$) yield

$$I(\hat{\alpha},\hat{\theta})=\begin{bmatrix} 82.467 & 0.0078091 \\ 0.0078091 & 0.0000016959\end{bmatrix},$$

and covariance matrix

$$\begin{bmatrix} 0.021502 & -99.0143 \\ -99.0143 & 1{,}045{,}620\end{bmatrix}.$$

The confidence interval for α is $0.55616 \pm 1.96(0.021502)^{1/2}$, or 0.55616 ± 0.28741, and for θ is $2{,}561.1 \pm 1.96(1{,}045{,}620)^{1/2}$, or $2{,}561.1 \pm 2{,}004.2$.

11.34 The density function is $f(x|\theta)=\theta^{-1}$, $0 \le x \le \theta$. The likelihood function is

$$\begin{aligned} L(\theta) &= \theta^{-n},\ 0 \le x_1,\ldots,x_n \le \theta \\ &= 0,\ \text{otherwise.}\end{aligned}$$

As a function of θ, the likelihood function is sometimes 0, and sometimes θ^{-n}. In particular, it is θ^{-n} only when θ is greater than or equal to all the xs. Equivalently, we have

$$\begin{aligned} L(\theta) &= 0,\ \theta < \max(x_1,\ldots,x_n) \\ &= \theta^{-n},\ \theta \ge \max(x_1,\ldots,x_n).\end{aligned}$$

Therefore, the likelihood function is maximized at $\hat{\theta} = \max(x_1,\ldots,x_n)$. Note that the calculus technique of setting the derivative equal to zero does not work here because the likelihood function is not continuous (and therefore not differentiable) at the maximum. From Examples 10.5 and 10.9 we know that this estimator is asymptotically unbiased and consistent, and we have its variance without recourse to Theorem 11.4. According to Theorem 11.4, we need

$$\begin{aligned} l(\theta) &= -n\ln\theta,\ \theta \ge \max(x_1,\ldots,x_n), \\ l'(\theta) &= -n\theta^{-1},\ \theta \ge \max(x_1,\ldots,x_n), \\ l''(\theta) &= n\theta^{-2},\ \theta \ge \max(x_1,\ldots,x_n).\end{aligned}$$

Then $\mathrm{E}[l''(\theta)] = n\theta^{-2}$ because with regard to the random variables, $n\theta^{-2}$ is a constant and, therefore, its expected value is itself. The information is then the negative of this number and must be negative.

With regard to assumption (ii) of Theorem 11.4,

$$\int_0^\theta \frac{\partial}{\partial\theta}\frac{1}{\theta}dx = \int_0^\theta -\theta^{-2}dx = -\frac{1}{\theta} \neq 0.$$

11.35 The first partial derivatives are

$$\frac{\partial l(\alpha,\beta)}{\partial\alpha} = -5\alpha - 3\beta + 50,$$
$$\frac{\partial l(\alpha,\beta)}{\partial\beta} = -3\alpha - 2\beta + 2.$$

The second partial derivatives are

$$\frac{\partial^2 l(\alpha,\beta)}{\partial\alpha^2} = -5,$$
$$\frac{\partial^2 l(\alpha,\beta)}{\partial\beta^2} = -2,$$
$$\frac{\partial^2 l(\alpha,\beta)}{\partial\alpha\partial\beta} = -3,$$

and so the information matrix is

$$\begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}.$$

The covariance matrix is the inverse of the information, or

$$\begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix}.$$

11.36 $\ln f(x) = -(1/2)\ln(2\pi\theta) - x^2/(2\theta)$, $\partial^2 \ln f(x)/\partial\theta^2 = (2\theta^2)^{-1} - x^2(\theta^3)^{-1}$, and $I(\theta) = n\mathrm{E}[-(2\theta^2)^{-1} + X^2(\theta^3)^{-1}] = n(2\theta^2)^{-1}$ since $X \sim N(0,\theta)$. Then $\mathrm{MSE}(\hat{\theta}) \doteq \mathrm{Var}(\hat{\theta}) \doteq 2\theta^2/n \doteq 2\hat{\theta}^2/n = 8/40 = 0.2$.

11.37 (a) $L = F(2)[1-F(2)]^3$. $F(2) = \int_0^2 2\lambda x e^{-\lambda x^2}dx = -e^{-\lambda x^2}\Big|_0^2 = 1 - e^{-4\lambda}$. $l = \ln(1-e^{-4\lambda}) - 12\lambda$. $\partial l/\partial\lambda = (1-e^{-4\lambda})^{-1}4e^{-4\lambda} - 12 = 0$. $e^{-4\lambda} = 3/4$. $\hat{\lambda} = (1/4)\ln(4/3)$.

(b) $P_1(\lambda) = 1 - e^{-4\lambda}$, $P_2(\lambda) = e^{-4\lambda}$, $P_1'(\lambda) = 4e^{-4\lambda}$, $P_2'(\lambda) = -4e^{-4\lambda}$. $I(\lambda) = 4[16e^{-8\lambda}/(1-e^{-4\lambda}) + 16e^{-8\lambda}/e^{-4\lambda}]$. $\widehat{\mathrm{Var}(\hat{\lambda})} = \{4[16(9/16)/(1/4) + 16(9/16)/(3/4)]\}^{-1} = 1/192$.

11.38 (a) Let $\theta = \mu/\Gamma(1+\tau^{-1})$. From Appendix **??**, $\mathrm{E}(X) = \theta\Gamma(1+\tau^{-1}) = \mu$.

(b) The density function is $f(x) = \exp\left\{-\left[\frac{\Gamma(1+\tau^{-1})x}{\mu}\right]^{\tau}\right\}\frac{\tau}{x}\left[\frac{\Gamma(1+\tau^{-1})x}{\mu}\right]^{\tau}$, and its logarithm is

$$\ln f(x) = -\left[\frac{\Gamma(1+\tau^{-1})x}{\mu}\right]^{\tau} + \ln\tau + \tau\ln[\Gamma(1+\tau^{-1})/\mu] + (\tau-1)\ln x.$$

The loglikelihood function is

$$l(\mu) = \sum_{j=1}^{n}\left\{-\left[\frac{\Gamma(1+\tau^{-1})x_j}{\mu}\right]^{\tau} + \ln\tau + \tau\ln[\Gamma(1+\tau^{-1})/\mu] + (\tau-1)\ln x_j\right\},$$

and its derivative is

$$l'(\mu) = \sum_{j=1}^{n}\left\{\tau\frac{\left[\Gamma(1+\tau^{-1})x_j\right]^{\tau}}{\mu^{\tau+1}} - \frac{\tau}{\mu}\right\}.$$

Setting the derivative equal to zero, moving the last term to the right-hand side, multiplying by μ, and dividing by τ produces the equation

$$\begin{aligned}
\sum_{j=1}^{n}\left[\frac{\Gamma(1+\tau^{-1})x_j}{\mu}\right]^{\tau} &= n\\
\left[\frac{\Gamma(1+\tau^{-1})}{\mu}\right]^{\tau}\sum_{j=1}^{n}x_j^{\tau} &= n\\
\left[\frac{\Gamma(1+\tau^{-1})}{n^{1/\tau}}\right]^{\tau}\sum_{j=1}^{n}x_j^{\tau} &= \mu^{\tau} \qquad (11.2)
\end{aligned}$$

and, finally,

$$\hat{\mu} = \Gamma(1+\tau^{-1})\left(\sum_{j=1}^{n}\frac{x_j^{\tau}}{n}\right)^{1/\tau}.$$

(c) The second derivative of the loglikelihood function is

$$\begin{aligned}
l''(\mu) &= \sum_{j=1}^{n}\left\{-\tau(\tau+1)\frac{\left[\Gamma(1+\tau^{-1})x_j\right]^{\tau}}{\mu^{\tau+2}} + \frac{\tau}{\mu^2}\right\}\\
&= \frac{n\tau}{\mu^2} - \tau(\tau+1)\Gamma(1+\tau^{-1})^{\tau}\mu^{-\tau-2}\sum_{j=1}^{n}x_j^{\tau}.
\end{aligned}$$

From (11.2), $\sum_{j=1}^{n}x_j^{\tau} = \left[\frac{n^{1/\tau}\hat{\mu}^{\tau}}{\Gamma(1+\tau^{-1})}\right]^{\tau}$ and therefore the observed information can be written as

$$\begin{aligned}
l''(\hat{\mu}) &= \frac{n\tau}{\hat{\mu}^2} - \tau(\tau+1)\Gamma(1+\tau^{-1})^{\tau}\hat{\mu}^{-\tau-2}\left[\frac{n^{1/\tau}\hat{\mu}^{\tau}}{\Gamma(1+\tau^{-1})}\right]^{\tau}\\
&= \frac{n\tau}{\hat{\mu}^2} - \frac{\tau(\tau+1)n}{\hat{\mu}^2} = -\frac{\tau^2 n}{\hat{\mu}^2},
\end{aligned}$$

and the negative reciprocal provides the variance estimate.

(d) The information requires the expected value of X^τ. From Appendix **??**, it is $\theta^\tau\Gamma(1+\frac{\tau}{\tau}) = \theta^\tau = [\mu/\Gamma(1+\tau^{-1})]^\tau$. Then

$$\begin{aligned}\mathrm{E}[l''(\mu)] &= \frac{n\tau}{\mu^2} - \tau(\tau+1)\Gamma(1+\tau^{-1})^\tau\mu^{-\tau-2}n[\mu/\Gamma(1+\tau^{-1})]^\tau \\ &= \frac{n\tau}{\mu^2} - \frac{\tau(\tau+1)n}{\mu^2} = -\frac{\tau^2 n}{\mu^2}.\end{aligned}$$

Changing the sign, inverting, and substituting $\hat{\mu}$ for μ produces the same estimated variance as in part (c).

(e) To obtain the distribution of $\hat{\mu}$, first obtain the distribution of $Y = X^\tau$. We have

$$\begin{aligned}S_Y(y) &= \Pr(Y > y) = \Pr(X^\tau > y) \\ &= \Pr(X > y^{1/\tau}) \\ &= \exp\left\{-\left[\frac{\Gamma(1+\tau^{-1})y^{1/\tau}}{\mu}\right]^\tau\right\} \\ &= \exp\left\{-\left[\frac{\Gamma(1+\tau^{-1})}{\mu}\right]^\tau y\right\}\end{aligned}$$

which is an exponential distribution with mean $[\mu/\Gamma(1+\tau^{-1})]^\tau$. Then $\sum_{j=1}^n X_j^\tau$ has a gamma distribution with parameters n and $[\mu/\Gamma(1+\tau^{-1})]^\tau$. Next look at

$$\frac{\Gamma(1+\tau^{-1})^\tau}{\mu^\tau}\sum_{j=1}^n X_j^\tau.$$

Multiplying by a constant changes the scale parameter, so this variable has a gamma distribution with parameters n and 1. Now raise this expression to the $1/\tau$ power. Then

$$\frac{\Gamma(1+\tau^{-1})}{\mu}\left(\sum_{j=1}^n X_j^\tau\right)^{1/\tau}$$

has a transformed gamma distribution with parameters $\alpha = n$, $\theta = 1$, and $\tau = \tau$. To create $\hat{\mu}$, this function must be multiplied by $\mu/n^{1/\tau}$, which changes the scale parameter to $\theta = \mu/n^{1/\tau}$. From Appendix **??**,

$$\mathrm{E}(\hat{\mu}) = \frac{\mu\Gamma(n+\tau^{-1})}{n^{1/\tau}\Gamma(n)}.$$

A similar argument provides the second moment and then a variance of

$$\mathrm{Var}(\hat{\mu}) = \frac{\mu^2\Gamma(n+2\tau^{-1})}{n^{2/\tau}\Gamma(n)} - \frac{\mu^2\Gamma(n+\tau^{-1})^2}{n^{2/\tau}\Gamma(n)^2}.$$

11.39 From Exercise 11.33 we have $\hat{\alpha} = 0.55616$, $\hat{\theta} = 2,561.1$, and covariance matrix

$$\begin{bmatrix} 0.021503 & -99.0188 \\ -99.0188 & 1{,}045{,}668 \end{bmatrix}.$$

The function to be estimated is $g(\alpha, \theta) = \alpha\theta$ with partial derivatives of θ and α. The approximated variance is

$$\begin{bmatrix} 2{,}561.1 & 0.55616 \end{bmatrix} \begin{bmatrix} 0.021503 & -99.0188 \\ -99.0188 & 1{,}045{,}668 \end{bmatrix} \begin{bmatrix} 2{,}561.1 \\ 0.55616 \end{bmatrix} = 182{,}402.$$

The confidence interval is $1{,}424.4 \pm 1.96\sqrt{182{,}402}$, or $1{,}424.4 \pm 837.1$.

11.40 The partial derivatives of the mean are

$$\frac{\partial e^{\mu+\sigma^2/2}}{\partial \mu} = e^{\mu+\sigma^2/2} = 123.017,$$
$$\frac{\partial e^{\mu+\sigma^2/2}}{\partial \sigma} = \sigma e^{\mu+\sigma^2/2} = 134.458.$$

The estimated variance is then

$$\begin{bmatrix} 123.017 & 134.458 \end{bmatrix} \begin{bmatrix} 0.1195 & 0 \\ 0 & 0.0597 \end{bmatrix} \begin{bmatrix} 123.017 \\ 134.458 \end{bmatrix} = 2{,}887.73.$$

11.41 (a) $f(x) = px^{p-1}$, $\ln f(x) = \ln p + (p-1)\ln x$, $\partial^2 \ln f(x)/\partial p^2 = -p^{-2}$, $I(p) = n\mathrm{E}(p^{-2}) = np^{-2}$, $\mathrm{Var}(\hat{p}) \doteq p^2/n$.

(b) From Exercise 11.6, $\hat{p} = -n/\Sigma \ln x_j$. The CI is $\hat{p} \pm 1.96\hat{p}/\sqrt{n}$.

(c) $\mu = p/(1+p)$. $\hat{p}/(1+\hat{p})$. $\partial\mu/\partial p = (1+p)^{-2}$. $\mathrm{Var}(\hat{\mu}) \doteq (1+p)^{-4}p^2/n$. The CI is $\hat{p}(1+\hat{p})^{-1} \pm 1.96\hat{p}(1+\hat{p})^{-2}/\sqrt{n}$.

11.42 (a) $\ln f(x) = -\ln\theta - x/\theta$, $\partial^2 \ln f(x)/\partial\theta^2 = \theta^{-2} - 2\theta^{-3}x$, $I(\theta) = n\mathrm{E}(-\theta^{-2} + 2\theta^{-3}X) = n\theta^{-2}$, $\mathrm{Var}(\hat{\theta}) \doteq \theta^2/n$.

(b) From Exercise 11.9, $\hat{\theta} = \bar{x}$. The CI is $\bar{x} \pm 1.96\bar{x}/\sqrt{n}$.

(c) $\mathrm{Var}(X) = \theta^2$. $\partial\,\mathrm{Var}(X)/\partial\theta = 2\theta$. $\widehat{\mathrm{Var}(X)} = \bar{x}^2$. $\mathrm{Var}[\widehat{\mathrm{Var}(X)}] \doteq (2\theta)^2\theta^2/n = 4\theta^4/n$. The CI is $\bar{x}^2 \pm 1.96(2\bar{x}^2)/\sqrt{n}$.

11.43 the maximum likelihood estimate is $\hat{\theta} = \bar{x} = 1{,}000$. $\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\bar{x}) = \theta^2/6$. The quantity to be estimated is $S(\theta) = e^{-1{,}500/\theta}$, and then

$$S'(\theta) = 1{,}500\theta^{-2}e^{-1{,}500/\theta}.$$

From the delta method

$$\begin{aligned}\text{Var}[S(\hat{\theta})] &= [S'(\hat{\theta})]^2\text{Var}(\hat{\theta})\\ &= [1{,}500(1{,}000)^{-2}e^{-1{,}500/1{,}000}]^2(1{,}000^2/6) = 0.01867.\end{aligned}$$

The standard deviation is 0.13664 and with $S(1{,}500) = 0.22313$, the confidence interval is $0.22313 \pm 1.96(0.13664)$, or 0.22313 ± 0.26781. An alternative that does not use the delta method is to start with a confidence interval for θ: $1{,}000 \pm 1.96(1{,}000)/\sqrt{6}$, which is $1{,}000 \pm 800.17$. Putting the endpoints into $S(\theta)$ produces the interval 0.00055 to 0.43463.

11.44 The loglikelihood function is $l = 81.61837(\alpha - 1) - 38{,}000\theta^{-1} - 10\ln\Gamma(\alpha) - 10\alpha\ln\theta$. Also, $\hat{\alpha} = 6.341$ and $\hat{\theta} = 599.3$. Using $v = 4$, we have

$$\begin{aligned}\frac{\partial^2 l(\alpha,\theta)}{\partial\alpha^2} &\doteq \frac{l(6.3416341, 599.3) - 2l(6.341, 599.3) + l(6.3403659, 599.3)}{(.0006341)^2}\\ &= -1.70790,\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l(\alpha,\theta)}{\partial\alpha\partial\theta} \doteq \frac{\begin{array}{r}l(6.34131705, 599.329965) - l(6.34131705, 599.270035)\\ -l(6.34068295, 599.329965) + l(6.34068295, 599.270035)\end{array}}{(.0006341)(.05993)}\\ = 0.0166861,\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l(\alpha,\theta)}{\partial\theta^2} &\doteq \frac{l(6.341, 599.35993) - 2l(6.341, 599.3) + l(6.341, 599.25007)}{(.05993)^2}\\ &= -0.000176536,\end{aligned}$$

$$I(\hat{\alpha},\hat{\theta}) = \begin{bmatrix} 1.70790 & 0.0166861 \\ 0.0166861 & 0.000176536 \end{bmatrix},$$

and its inverse is

$$\widehat{\text{Var}} = \begin{bmatrix} 7.64976 & -723.055 \\ -723.055 & 74{,}007.7 \end{bmatrix}.$$

The mean is $\alpha\theta$, and so the derivative vector is $\begin{bmatrix}599.3 & 6.341\end{bmatrix}$. The variance of $\widehat{\alpha\theta}$ is estimated as 227,763 and a 95% CI is $3{,}800 \pm 1.97\sqrt{227{,}763} = 3{,}800 \pm 935$.

11.45 $\hat{\alpha} = 3.8629$. $\ln f(x) = \ln\alpha + \alpha\ln\lambda - (\alpha+1)\ln(\lambda+x)$. $\partial^2 \ln f(x)/\partial\alpha^2 = -\alpha^{-2}$. $I(\alpha) = n\alpha^{-2}$. $\text{Var}(\hat{\alpha}) \doteq \alpha^2/n$. Inserting the estimate gives 2.9844.

$$\begin{aligned}\text{E}(X \wedge 500) &= \int_0^{500} x\alpha 1{,}000^\alpha(1{,}000+x)^{-\alpha-1}dx\\ &\quad + 500\int_{500}^{\infty}\alpha 1{,}000^\alpha(1{,}000+x)^{-\alpha-1}dx\\ &= \frac{1{,}000}{\alpha-1} - (2/3)^\alpha\frac{1{,}500}{\alpha-1}.\end{aligned}$$

Evaluated at $\hat{\alpha}$, it is 239.88. The derivative with respect to α is

$$-\frac{1{,}000}{(\alpha-1)^2}+(2/3)^{\alpha}\frac{1{,}500}{(\alpha-1)^2}-(2/3)^{\alpha}\frac{1{,}500}{\alpha-1}\ln(2/3),$$

which is -39.428 when evaluated at $\hat{\alpha}$. The variance of the LEV estimator is $(-39.4298)^2(2.9844) = 5{,}639.45$, and the CI is 239.88 ± 133.50.

11.46 The likelihood function is $L(\theta) = \theta^{-20}e^{-28{,}488/\theta}$. Then $l(\theta) = -20\ln(\theta) - 28{,}488/\theta$. With $\hat{\theta} = 1{,}424.4$, $l(\hat{\theta}) = -165.23$. A 95% confidence region solves $l(\theta) = -165.23 - 1.92 = -167.15$. The solutions are 946.87 and 2,285.07. Inserting the solutions in $S(200) = e^{-200/\theta}$ produces the interval 0.810 to 0.916. The symmetric interval was 0.816 to 0.922.

12

ESTIMATION BASED ON EMPIRICAL DATA

12.1 The Empirical Distribution

The material presented here has traditionally been presented under the heading of "survival models" with the accompanying notion that the techniques are useful only when studying lifetime distributions. Standard texts on the subject such as Klein and Moeschberger [12] and Lawless [14] contain examples that are exclusively oriented in that direction. However, the same problems that occur when modeling lifetime occur when modeling payment amounts. The examples we present are of both types. However, the latter sections focus on special considerations when constructing decrement models. Only a handful of references are presented, most of the results being well developed in the survival models literature. Readers wanting more detail and proofs should consult a text dedicated to the subject, such as the ones just mentioned.

In this chapter, it is assumed that the type of model is known but not the full description of the model. In Chapter **??**, models were divided into two types—data dependent and parametric. The definitions are repeated here.

Definition 12.1 *A **data-dependent distribution** is at least as complex as the data or knowledge that produced it, and the number of "parameters" increases as the number of data points or amount of knowledge increases.*

Loss Models: From Data to Decisions, 5th edition.
By Stuart A. Klugman, Harry H. Panjer, Gordon E. Willmot

Table 12.1 Data Set B.

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1,193	1,340	1,884	2,558	15,743

Table 12.2 Data Set C.

Payment range	Number of payments
0–7,500	99
7,500–17,500	42
17,500–32,500	29
32,500–67,500	28
67,500–125,000	17
125,000–300,000	9
Over 300,000	3

Definition 12.2 *A **parametric distribution** is a set of distribution functions, each member of which is determined by specifying one or more values called **parameters**. The number of parameters is fixed and finite.*

This chapter will focus on data-dependent distributions as models, making few, if any, assumptions about the underlying distribution. To fix the most important concepts, we begin by assuming that we have a sample of n observations that are an independent and identically distributed sample from the same (unspecified) continuous distribution. This is referred to as a **complete data** situation. In that context, we have the following definition.

Definition 12.3 *The **empirical distribution** is obtained by assigning probability $1/n$ to each data point.*

When observations are collected from a probability distribution, the ideal situation is to have the (essentially) exact[1] value of each observation. This case is referred to as "complete, individual data" and applies to Data Set B, introduced in Chapter 10 and reproduced here as Table 12.1. There are two reasons why exact data may not be available. One is grouping, in which all that is recorded is the range of values in which the observation belongs. Grouping applies to Data Set C and for Data Set A for those with five or more accidents. There data sets were introduced in Chapter 10 and are reproduced here as Tables 12.2 and 12.3 respectively.

A second reason that exact values may not be available is the presence of censoring or truncation. When data are censored from below, observations below a given value are known to be below that value but the exact value is unknown. When data are censored from above, observations above a given value are known to be above that value but the exact value is unknown. Note that censoring effectively creates grouped data. When the data are grouped in the first place, censoring has no effect. For example, the data in Data Set C may have been censored from above at 300,000, but we cannot know for sure from

[1]Some measurements are never exact. Ages may be rounded to the nearest whole number, monetary amounts to the nearest dollar, car mileage to the nearest mile, and so on. This text is not concerned with such rounding errors. Rounded values will be treated as if they are exact.

Table 12.3 Data Set A.

Number of accidents	Number of drivers
0	81,714
1	11,306
2	1,618
3	250
4	40
5 or more	7

the data set and that knowledge has no effect on how we treat the data. In contrast, were Data Set B to be censored at 1,000, we would have 15 individual observations and then 5 grouped observations in the interval from 1,000 to infinity.

In insurance settings, censoring from above is fairly common. For example, if a policy pays no more than 100,000 for an accident, any time the loss is above 100,000 the actual amount will be unknown but we will know that it happened. Note that Data Set A has been censored from above at 5. This is more common language than to say that Data Set A has some individual data and some grouped data. When studying mortality or other decrements, the study period may end with some individuals still alive. They are censored from above in that we know the death will occur sometime after their age when the study ends.

When data are truncated from below, observations below a given value are not recorded. Truncation from above implies that observations above a given value are not recorded. In insurance settings, truncation from below is fairly common. If an automobile physical damage policy has a per-claim deductible of 250, any losses below 250 will not come to the attention of the insurance company and so will not appear in any data sets. Data sets may have truncation forced on them. For example, if Data Set B were to be truncated from below at 250, the first 7 observations would disappear and the remaining 13 would be unchanged. In decrement studies it is unusual to observe individuals from birth. If someone is first observed at, say, age 20, that person is from a population where anyone who died before age 20 would not have been observed and thus is truncated from below.

As noted in Definition 12.3, the empirical distribution assigns probability $1/n$ to each data point. That definition works well when the value of each data point is recorded. An alternative definition follows.

Definition 12.4 *The* ***empirical distribution function*** *is*

$$F_n(x) = \frac{\textit{number of observations } \leq x}{n},$$

where n is the total number of observations.

■ **EXAMPLE 12.1**

Provide the empirical probability functions for the data in Data Sets A and B. For Data Set A also provide the empirical distribution function. For Data Set A assume all seven drivers who had five or more accidents had exactly five accidents.

For notation, a subscript of the sample size (or of n if the sample size is not known) is used to indicate an empirical function. Without the subscript, the function represents the true function for the underlying random variable. For Data Set A, the empirical probability function is

$$p_{94,935}(x) = \begin{cases} 81{,}714/94{,}935 = 0.860736, & x = 0, \\ 11{,}306/94{,}935 = 0.119092, & x = 1, \\ 1{,}618/94{,}935 = 0.017043, & x = 2, \\ 250/94{,}935 = 0.002633, & x = 3, \\ 40/94{,}935 = 0.000421, & x = 4, \\ 7/94{,}935 = 0.000074, & x = 5, \end{cases}$$

where the values add to 0.999999 due to rounding. The empirical distribution function is a step function with jumps at each data point.

$$F_{94,935}(x) = \begin{cases} 0/94{,}935 = 0.000000, & x < 0, \\ 81{,}714/94{,}935 = 0.860736, & 0 \le x < 1, \\ 93{,}020/94{,}935 = 0.979828, & 1 \le x < 2, \\ 94{,}638/94{,}935 = 0.996872, & 2 \le x < 3, \\ 94{,}888/94{,}935 = 0.999505, & 3 \le x < 4, \\ 94{,}928/94{,}935 = 0.999926, & 4 \le x < 5, \\ 94{,}935/94{,}935 = 1.000000, & x \ge 5. \end{cases}$$

For Data Set B,

$$p_{20}(x) = \begin{cases} 0.05, & x = 27, \\ 0.05, & x = 82, \\ 0.05, & x = 115, \\ \vdots & \vdots \\ 0.05, & x = 15{,}743. \end{cases}$$

□

In the following example, not all values are distinct.

■ **EXAMPLE 12.2**

Consider a data set containing the numbers 1.0, 1.3, 1.5, 1.5, 2.1, 2.1, 2.1, and 2.8. Determine the empirical distribution function.

The empirical distribution function is a step function with the following values:

$$F_8(x) = \begin{cases} 0, & x < 1.0, \\ 1 - \frac{7}{8} = 0.125, & 1.0 \le x < 1.3, \\ 1 - \frac{6}{8} = 0.250, & 1.3 \le x < 1.5, \\ 1 - \frac{4}{8} = 0.500, & 1.5 \le x < 2.1, \\ 1 - \frac{1}{8} = 0.875, & 2.1 \le x < 2.8, \\ 1, & x \ge 2.8. \end{cases}$$

□

To assess the quality of the estimate, we examine statistical properties, in particular the mean and variance. Working with the empirical estimate of the survival function is straightforward. To see that with complete data the empirical estimator of the survival function is unbiased and consistent, recall that the empirical estimate of $F(x)$ is $F_n(x) = Y/n$, where Y is the number of observations in the sample that are less than or equal to x. Then Y must have a binomial distribution with parameters n and $F(x)$ and

$$\mathrm{E}[F_n(x)] = \mathrm{E}\left(\frac{Y}{n}\right) = \frac{nF(x)}{n} = F(x).$$

demonstrating that the estimator is unbiased. The variance is

$$\mathrm{Var}[F_n(x)] = \mathrm{Var}\left(\frac{Y}{n}\right) = \frac{F(x)[1 - F(x)]}{n},$$

which has a limit of zero, thus verifying consistency.

To make use of the result, the best we can do for the variance is estimate it. It is unlikely we know the value of $F(x)$ because that is the quantity we are trying to estimate. The estimated variance is given by

$$\widehat{\mathrm{Var}}[F_n(x)] = \frac{F_n(x)[1 - F_n(x)]}{n}. \tag{12.1}$$

The same results hold for empirically estimated probabilities. Let $p = \Pr(a < X \le b)$. The empirical estimate of p is $\hat{p} = F_n(b) - F_n(a)$. Arguments similar to those used for $F_n(x)$ verify that $\hat{p}$ is unbiased and consistent, with $\mathrm{Var}(\hat{p}) = p(1 - p)/n$.

EXAMPLE 12.3

For the data in Example 12.2, estimate the variance of $F_8(1.4)$.

From the example we have $F_8(1.4) = 0.25$. From (12.1),

$$\widehat{\mathrm{Var}}[F_8(1.4)] = \frac{F_8(1.4)[1 - F_8(1.4)]}{8} = \frac{0.25(1 - 0.25)}{8} = 0.02344.$$

□

12.2 Empirical distributions for grouped data

For grouped data as in Data Set C, construction of the empirical distribution as defined previously is not possible. However, it is possible to approximate the empirical distribution. The strategy is to obtain values of the empirical distribution function wherever possible and then connect those values in some reasonable way. For grouped data, the distribution function is usually approximated by connecting the points with straight lines. For notation, let the group boundaries be $c_0 < c_1 < \cdots < c_k$, where often $c_0 = 0$ and $c_k = \infty$. The number of observations falling between c_{j-1} and c_j is denoted n_j with $\sum_{j=1}^{k} n_j = n$. For such data, we are able to determine the empirical distribution at each group boundary. That is, $F_n(c_j) = (1/n)\sum_{i=1}^{j} n_i$. Note that no rule is proposed for observations that fall on a group boundary. There is no correct approach, but whatever approach is used, consistency in assignment of observations to groups should be used. Note that in Data Set C it is not possible to tell how the assignments were made. If we had that knowledge, it would not affect any subsequent calculations.[2]

Definition 12.5 *For grouped data, the distribution function obtained by connecting the values of the empirical distribution function at the group boundaries with straight lines is called the* ***ogive****. The formula is*

$$F_n(x) = \frac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j), \; c_{j-1} \le x \le c_j.$$

This function is differentiable at all values except group boundaries. Therefore the density function can be obtained. To completely specify the density function, it is arbitrarily made right continuous.

Definition 12.6 *For grouped data, the empirical density function can be obtained by differentiating the ogive. The resulting function is called a* ***histogram****. The formula is*

$$f_n(x) = \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} = \frac{n_j}{n(c_j - c_{j-1})}, \; c_{j-1} \le x < c_j.$$

Many computer programs that produce histograms actually create a bar chart with bar heights proportional to n_j/n. A bar chart is acceptable if the groups have equal width, but if not, then the preceding formula is needed. The advantage of this approach is that the histogram is indeed a density function, and, among other things, areas under the histogram can be used to obtain empirical probabilities.

EXAMPLE 12.4

Construct the ogive and histogram for Data Set C.

[2]Technically, for the interval from c_{j-1} to c_j, $x = c_j$ should be included and $x = c_{j-1}$ excluded in order for $F_n(c_j)$ to be the empirical distribution function.

The distribution function is

$$F_{227}(x) = \begin{cases} 0.000058150x, & 0 \le x \le 7{,}500, \\ 0.29736 + 0.000018502x, & 7{,}500 \le x \le 17{,}500, \\ 0.47210 + 0.000008517x, & 17{,}500 \le x \le 32{,}500, \\ 0.63436 + 0.000003524x, & 32{,}500 \le x \le 67{,}500, \\ 0.78433 + 0.000001302x, & 67{,}500 \le x \le 125{,}000, \\ 0.91882 + 0.000000227x, & 125{,}000 \le x \le 300{,}000, \\ \text{undefined}, & x > 300{,}000, \end{cases}$$

where, for example, for the range $32{,}500 \le x \le 67{,}500$ the calculation is

$$F_{227}(x) = \frac{67{,}500 - x}{67{,}500 - 32{,}500}\frac{170}{227} + \frac{x - 32{,}500}{67{,}500 - 32{,}500}\frac{198}{227}.$$

The value is undefined above 300,000 because the last interval has a width of infinity. A graph of the ogive for values up to 125,000 appears in Figure 12.1. The derivative is simply a step function with the following values:

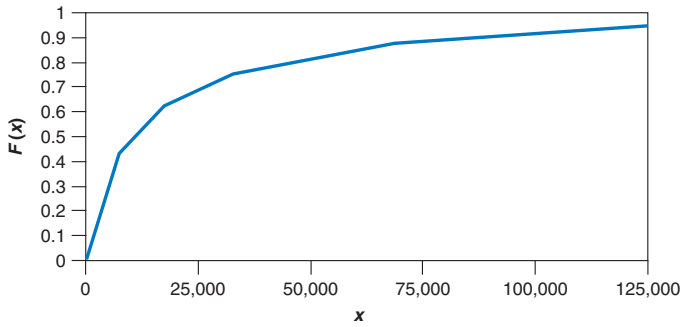

Figure 12.1 Ogive for general liability losses.

$$f_{227}(x) = \begin{cases} 0.000058150, & 0 \le x < 7{,}500, \\ 0.000018502, & 7{,}500 \le x < 17{,}500, \\ 0.000008517, & 17{,}500 \le x < 32{,}500, \\ 0.000003524, & 32{,}500 \le x < 67{,}500, \\ 0.000001302, & 67{,}500 \le x < 125{,}000, \\ 0.000000227, & 125{,}000 \le x < 300{,}000, \\ \text{undefined}, & x \ge 300{,}000. \end{cases}$$

A graph of the function up to 125,000 appears in Figure 12.2. □

12.2.1 Exercises

12.1 Construct the ogive and histogram for the data in Table 12.4.

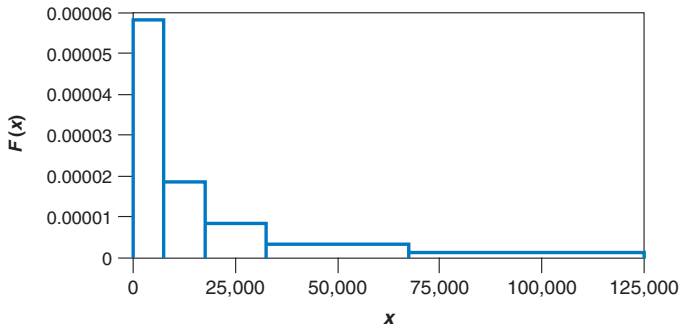

Figure 12.2 Histogram of general liability losses.

Table 12.4 Data for Exercise 12.1.

Payment range	Number of payments
0–25	6
25–50	24
50–75	30
75–100	31
100–150	57
150–250	80
250–500	85
500–1,000	54
1,000–2,000	15
2,000–4,000	10
Over 4,000	0

12.2 (*) The following 20 windstorm losses (in millions of dollars) were recorded in one year:

1	1	1	1	1	2	2	3	3	4
6	6	8	10	13	14	15	18	22	25

(a) Construct an ogive based on using class boundaries at 0.5, 2.5, 8.5, 15.5, and 29.5.

(b) Construct a histogram using the same boundaries as in part (a).

12.3 The data in Table 12.5 are from Herzog and Laverty [7]. A certain class of 15-year mortgages was followed from issue until December 31, 1993. The issues were split into those that were refinances of existing mortgages and those that were original issues. Each entry in the table provides the number of issues and the percentage of them that were still in effect after the indicated number of years. Draw as much of the two ogives (on the same graph) as is possible from the data. Does it appear from the ogives that the lifetime variable (time to mortgage termination) has a different distribution for refinanced versus original issues?

12.4 (*) The data in Table 12.6 were collected (units are millions of dollars). Construct the histogram.

Table 12.5 Data for Exercise 12.3.

	Refinances		Original	
Years	No. issued	Survived	No. issued	Survived
1.5	42,300	99.97	12,813	99.88
2.5	9,756	99.82	18,787	99.43
3.5	1,550	99.03	22,513	98.81
4.5	1,256	98.41	21,420	98.26
5.5	1,619	97.78	26,790	97.45

Table 12.6 Data for Exercise 12.4.

Loss	No. of observations
0–2	25
2–10	10
10–100	10
100–1,000	5

12.5 (*) Forty losses have been observed. Sixteen losses are between 1 and $\frac{4}{3}$ (in millions), and the sum of the 16 losses is 20. Ten losses are between $\frac{4}{3}$ and 2 with a total of 15. Ten more are between 2 and 4 with a total of 35. The remaining 4 losses are greater than 4. Using the empirical model based on these observations, determine $\mathrm{E}(X \wedge 2)$.

12.6 (*) A sample of size 2,000 contains 1,700 observations that are no greater than 6,000, 30 that are greater than 6,000 but no greater than 7,000, and 270 that are greater than 7,000. The total amount of the 30 observations that are between 6,000 and 7,000 is 200,000. The value of $\mathrm{E}(X \wedge 6{,}000)$ for the empirical distribution associated with these observations is 1,810. Determine $\mathrm{E}(X \wedge 7{,}000)$ for the empirical distribution.

12.7 (*) A random sample of unknown size produced 36 observations between 0 and 50; x between 50 and 150; y between 150 and 250; 84 between 250 and 500; 80 between 500 and 1,000; and none above 1,000. Two values of the ogive constructed from these observations are $F_n(90) = 0.21$ and $F_n(210) = 0.51$. Determine the value of x.

12.8 The data in Table 12.7 are from *Loss Distributions* [8, p. 128]. It represents the total damage done by 35 hurricanes between the years 1949 and 1980. The losses have been adjusted for inflation (using the Residential Construction Index) to be in 1981 dollars. The entries represent all hurricanes for which the trended loss was in excess of 5,000,000.

The federal government is considering funding a program that would provide 100% payment for all damages for any hurricane causing damage in excess of 5,000,000. You have been asked to make some preliminary estimates.

(a) Estimate the mean, standard deviation, coefficient of variation, and skewness for the population of hurricane losses.

(b) Estimate the first and second limited moments at 500,000,000.

12.9 (*) There have been 30 claims recorded in a random sampling of claims. There were 2 claims for 2,000, 6 for 4,000, 12 for 6,000, and 10 for 8,000. Determine the empirical skewness coefficient.

Table 12.7 Trended hurricane losses.

Year	Loss (10^3)	Year	Loss (10^3)	Year	Loss (10^3)
1964	6,766	1964	40,596	1975	192,013
1968	7,123	1949	41,409	1972	198,446
1971	10,562	1959	47,905	1964	227,338
1956	14,474	1950	49,397	1960	329,511
1961	15,351	1954	52,600	1961	361,200
1966	16,983	1973	59,917	1969	421,680
1955	18,383	1980	63,123	1954	513,586
1958	19,030	1964	77,809	1954	545,778
1974	25,304	1955	102,942	1970	750,389
1959	29,112	1967	103,217	1979	863,881
1971	30,146	1957	123,680	1965	1,638,000
1976	33,727	1979	140,136		

12.3 Empirical estimation with right censored data

In this section we generalize the empirical approach of the previous section to situations where the data are not complete. In particular, we assume that individual observations may be right censored. We have the following definition.

Definition 12.7 *An observation is* ***censored from above*** *(also called* ***right censored****) at u if when it is at or above u it is recorded as being equal to u, but when it is below u it is recorded at its observed value.*

In insurance claims data, the presence of a policy limit may give rise to right censored observations. When the amount of the loss equals or exceeds the limit u, benefits beyond that value are not paid, and so the exact value is typically not recorded. However, it is known that a loss of at least u has occurred.

When carrying out a study of the mortality of humans and the person is alive when the study ends, right censoring has occurred. The person's age at death is not known, but it is known that it is at least as large as the age when the study ended. Right censoring also affects those who leave the study prior to its end due to surrender or lapse. Note that this discussion could have been about other decrements, such as disability, policy surrender, or retirement.

For this section and the next two, we assume that the underlying random variable has a continuous distribution. While data from discrete random variables can also be right censored (Data Set A is an example), use of empirical estimators is rare and thus developing analogous formulas is unlikely to be worth the effort.

We now make specific assumptions regarding how the data are collected and recorded. It is assumed that we have a random sample for which some (but not all) of the data are right censored. For the uncensored (i.e., completely known) observations, we will denote their k unique values by $y_1 < y_2 < \cdots < y_k$. We let s_i denote the number of times that y_i appears in the sample. We also set y_0 as the minimum possible value for an observation and assume $y_0 < y_1$. Often, $y_0 = 0$. Similarly, set y_{max} as the largest observation in

the data, censored or uncensored. Hence, $y_{max} \geq y_k$. Our goal is to create an empirical (data-dependent) distribution that places probability at the values $y_1 < y_2 < \cdots < y_k$.

With regard to the censored observations, in general it is not necessary to know the exact censoring times, u. We need only the number of censored observations between the ys. That is, let b_i equal the number of right censored observations in the interval $[y_i, y_{y+1})$ for $i = 1, 2, \ldots, k-1$. We make the assumption that if an observation is censored at y_i, then the observation is censored at $y_i + 0$ (i.e., in the lifetime situation, immediately after the death). It is possible to have censored observations at values between y_0 and y_1. However, because we are placing probability only at the uncensored values these observations provide no information about those probabilities and so can be dropped. When referring to the sample size, n will denote the number of observations after these have been dropped. Observations censored at y_k or above cannot be ignored. Let b_k be the number of observations right censored at $y_k + 0$ or later. Note that if $b_k = 0$, then $y_{max} = y_k$.

The final important quantity is r_i, referred to as the number "at risk" at y_i. When thinking in terms of a mortality study, the risk set comprises the individuals who are under observation at that age. Included are all who die at that age or later and all who are censored at that age or later. Formally,

Definition 12.8 *The* ***number at risk****, r_i at observation y_i is*

$$r_i = \begin{cases} n, & i = 1 \\ r_{i-1} - s_{i-1} - b_{i-1}, & i = 2, 3, \ldots, k+1. \end{cases}$$

This formula reflects the fact that the number at risk at y_i is that at y_{i-1} less the s_{i-1} exact observations at y_{i-1} and the b_{i-1} censored observations in $[y_{i-1}, y_i)$. Note that $b_k = r_k - s_k$ and hence $r_{k+1} = 0$.

The following numerical example illustrates these ideas.

■ **EXAMPLE 12.5**

Determine the number at risk for the data in Table 12.8. Note that there is no context for this example. Once the observed data are reduced to these values, the source becomes irrelevant.

Adding the values in the s and b columns reveals that there are $n = 20$ observations. Thus, $r_1 = n = 20$. Note that $b_7 = 1$ and therefore there is one censored value larger than 12, the largest observed value. The value $y_{max} = 15$ indicates that this observation was censored at 15. The risk set values are calculated in the table. □

It should be noted that if there is no censoring so that $b_i = 0$ for all i then the data are complete and the techniques of Section 12.1 may be used. As such, the approach of this section may be viewed as a generalization.

We shall now present an heuristic derivation of a well-known generalization of the empirical distribution function. This estimator is referred to as either the Kaplan–Meier or the product limit estimator.

To proceed, we first present some basic facts regarding the distribution of a discrete random variable Y say, with support on the points $y_1 < y_2 < \cdots < y_k$. Let $p(y_j) = \Pr(Y = y_j)$, and then the survival function is (where $i|y_i > y$ means to take the sum or product over all values of i where $y_i > y$)

Table 12.8 Data for Example 12.5

i	y_i	s_i	b_i	r_i
0	0	–	–	–
1	1	1	0	20
2	2	1	1	$20-1-0=19$
3	4	2	2	$19-1-1=17$
4	5	1	1	$17-2-2=13$
5	8	3	0	$13-1-1=11$
6	9	4	1	$11-3-0=8$
7	12	2	1	$8-4-1=3$
max	15	–	–	$3-2-1=0$

$$S(y)=\Pr(Y>y)=\sum_{i|y_i>y} p(y_i).$$

Setting $y=y_j$ for $j<k$ we have

$$S(y_j)=\sum_{i=j+1}^{k} p(y_i),$$

and $S(y_k)=0$. We also have $S(y_0)=1$ from the definition of y_0.

Definition 12.9 *The* ***discrete failure rate*** *function is*

$$h(y_j)=\Pr(Y=y_j|Y\ge y_j)=p(y_j)/S(y_{j-1}),\quad j=1,2,\ldots,k.$$

Thus,

$$h(y_j)=\frac{S(y_{j-1})-S(y_j)}{S(y_{j-1})}=1-\frac{S(y_j)}{S(y_{j-1})},$$

implying that $S(y_j)/S(y_{j-1})=1-h(y_j)$. Hence,

$$S(y_j)=\frac{S(y_j)}{S(y_0)}=\prod_{i=1}^{j}\frac{S(y_i)}{S(y_{i-1})}=\prod_{i=1}^{j}[1-h(y_i)].$$

Also, $p(y_1)=h(y_1)$, and for $j=2,3,\ldots,k$,

$$p(y_j)=h(y_j)S(y_{j-1})=h(y_j)\prod_{i=1}^{j-1}[1-h(y_i)].$$

The heuristic derivation proceeds by viewing $\lambda_j=h(y_j)$ for $j=1,2,\ldots,k$ as unknown parameters, and estimating them by a nonparametric "maximum likelihood" based argument.[3] See Lawless [14] for a more detailed discussion. For the present data, the s_j

[3]Maximum likelihood estimation is covered in Chapter 11. Candidates preparing for the Society of Actuaries Long-Term Actuarial Mathematics exam will not be tested on the method itself, only the estimators presented in this chapter.

uncensored observations at y_j each contribute $p(y_j)$ to the likelihood where $p(y_1) = \lambda_1$ and

$$p(y_j) = \lambda_j \prod_{i=1}^{j-1}(1-\lambda_i), \quad j = 2, 3, \ldots, k.$$

Each of the b_j censored observations contributes

$$S(y_j) = \prod_{i=1}^{j}(1-\lambda_i), \quad j = 1, 2, \ldots, k-1,$$

to the likelihood (recall that $S(y) = S(y_j)$ for $y_j \leq y < y_{j+1}$), and the b_k censored observations at or above y_k each contribute $S(y_k) = \prod_{i=1}^{k}(1-\lambda_i)$.

The likelihood is formed by taking products over all contributions (assuming independence of all data points) namely

$$L(\lambda_1, \lambda_2, \ldots, \lambda_k) = \prod_{j=1}^{k} \left\{ [p(y_j)]^{s_j} [S(y_j)]^{b_j} \right\},$$

which, in terms of the λ_js, becomes

$$\begin{aligned} L(\lambda_1, \lambda_2, \ldots, \lambda_k) &= \lambda_1^{s_1} \left\{ \prod_{j=2}^{k} \left[\lambda_j \prod_{i=1}^{j-1} (1-\lambda_i) \right]^{s_j} \right\} \prod_{j=1}^{k} \left[\prod_{i=1}^{j} (1-\lambda_i) \right]^{b_j} \\ &= \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) \left[\prod_{j=2}^{k} \prod_{i=1}^{j-1} (1-\lambda_i)^{s_j} \right] \prod_{j=1}^{k} \prod_{i=1}^{j} (1-\lambda_i)^{b_j} \\ &= \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) \left[\prod_{i=1}^{k-1} \prod_{j=i+1}^{k} (1-\lambda_i)^{s_j} \right] \prod_{i=1}^{k} \prod_{j=i}^{k} (1-\lambda_i)^{b_j}, \end{aligned}$$

where the last line follows by interchanging the order of multiplication in each of the two double products. Thus,

$$\begin{aligned} L(\lambda_1, \lambda_2, \ldots, \lambda_k) &= \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) (1-\lambda_k)^{b_k} \prod_{i=1}^{k-1} (1-\lambda_i)^{b_i + \sum_{m=i+1}^{k} (s_m + b_m)} \\ &= \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) (1-\lambda_k)^{b_k} \prod_{i=1}^{k-1} (1-\lambda_i)^{b_i + \sum_{m=i+1}^{k} (r_m - r_{m+1})} \\ &= \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) (1-\lambda_k)^{r_k - s_k} \prod_{i=1}^{k-1} (1-\lambda_i)^{b_i + r_{i+1} - r_{k+1}}. \end{aligned}$$

Observe that $r_{k+1} = 0$ and $b_i + r_{i+1} = r_i - s_i$. Hence,

$$L(\lambda_1, \lambda_2, \ldots, \lambda_k) = \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) (1-\lambda_k)^{r_k - s_k} \prod_{i=1}^{k-1} (1-\lambda_i)^{r_i - s_i}$$
$$= \prod_{j=1}^{k} \lambda_j^{s_j} (1-\lambda_j)^{r_j - s_j}.$$

This likelihood has the appearance of a product of binomial likelihoods. That is, this is the same likelihood as if $s_1, s_2, \ldots, s_k$ were realizations of k independent binomial observations with parameters $m = r_j$ and $q = \lambda_j$. The "maximum likelihood estimate" $\hat{\lambda}_j$ of λ_j is obtained by taking logarithms, namely

$$l(\lambda_1, \lambda_2, \ldots, \lambda_k) = \ln L(\lambda_1, \lambda_2, \ldots, \lambda_k) = \sum_{j=1}^{k} \left[s_j \ln \lambda_j + (r_j - s_j) \ln(1-\lambda_j) \right],$$

implying that

$$\frac{\partial l}{\partial \lambda_j} = \frac{s_j}{\lambda_j} - \frac{r_j - s_j}{1-\lambda_j}, \quad j = 1, 2, \ldots, k.$$

Equating this latter expression to zero yields $\hat{\lambda}_j = s_j / r_j$.

For $y = y_k$, the Kaplan–Meier [10] estimate $S_n(y)$ of $S(y)$ is obtained by replacing λ_j by $\hat{\lambda}_j = s_j / r_j$ wherever it appears. Noting that $S(y) = S(y_j)$ for $y_j \le y < y_{j+1}$, it follows that

$$S_n(y) = \begin{cases} 1, & y < y_1, \\ \prod_{i=1}^{j} \left(1 - \hat{\lambda}_i\right) = \prod_{i=1}^{j} \left(1 - \frac{s_i}{r_i}\right), & y_j \le y < y_{j+1}, \quad j = 1, 2, \ldots, k-1, \\ \prod_{i=1}^{k} \left(1 - \hat{\lambda}_i\right) = \prod_{i=1}^{k} \left(1 - \frac{s_i}{r_i}\right), & y_k \le y < y_{max}. \end{cases}$$

This may be written more succinctly as $S_n(y) = \prod_{i|y_i \le y} (1 - \hat{\lambda}_i)$ for $y < y_{max}$. When $y_{max} = y_k$, interpret $y_k \le y < y_{max}$ as $y = y_k$.

EXAMPLE 12.6

Construct the Kaplan–Meier estimate of $S(y)$ using the data in Example 12.5. Indicate how the anwer would change if $s_7 = 3$ and $b_7 = 0$.

The calculations appear in Table 12.9. The estimated survival function is

$$S_{20}(y) = \begin{cases} = 1, & y < 1, \\ = 0.950, & 1 \le y < 2, \\ = 0.900, & 2 \le y < 4, \\ = 0.794, & 4 \le y < 5, \\ = 0.737, & 5 \le y < 8, \\ = 0.533, & 8 \le y < 9, \\ = 0.167, & 9 \le y < 12, \\ = 0.089, & 12 \le y < 15. \end{cases}$$

Table 12.9 Kaplan–Meier estimates for Example 12.6

i	y_i	s_i	r_i	$\hat{S}_n(y_i)$
1	1	1	20	$1 - 1/20 = 0.950$
2	2	1	19	$0.95(1 - 1/19) = 0.900$
3	4	2	17	$0.9(1 - 2/17) = 0.794$
4	5	1	13	$0.794(1 - 1/13) = 0.733$
5	8	3	11	$0.733(1 - 3/11) = 0.533$
6	9	4	8	$0.533(1 - 4/8) = 0.267$
7	12	2	3	$0.267(1 - 2/3) = 0.089$

With the change in values, we have $y_{max} = y_7 = 12$ and $S_{20}(y) = 0.267(1 - 3/3) = 0$ for $y = 12$.

□

We now discuss estimation for $y \geq y_{max}$. First, note that if $s_k = r_k$ (no censored observations at y_k), then $S_n(y_k) = 0$ and $S_n(y) = 0$ for $y \geq y_k$ is clearly the (only) obvious choice. However, if $S_n(y_k) > 0$, as in the previous example, there are no empirical data to estimate $S(y)$ for $y \geq y_{max}$, and tail estimates for $y \geq y_{max}$ (often called tail corrections) are needed. There are three popular extrapolations:

- Efron's tail correction [6] assumes that $S_n(y) = 0$ for $y \geq y_{max}$.
- Klein and Moeschberger [12, p. 118] assume that $S_n(y) = S_n(y_k)$ for $y_k \leq y < \gamma$ and $S_n(y) = 0$ for $y \geq \gamma$, where $\gamma > y_{max}$ is a plausible upper limit for the underlying random variable. For example, in a study of human mortality it might be 120 years.
- Brown, Hollander, and Korwar's exponential tail correction [2] assumes that $S_n(y_{max}) = S_n(y_k)$ and that $S_n(y) = e^{-\hat{\beta}y}$ for $y \geq y_{max}$. With $y = y_{max}$, $\hat{\beta} = -\ln S_n(y_k)/y_{max}$, and thus

$$S_n(y) = e^{y[\ln S_n(y_k)]/y_{max}} = [S_n(y_k)]^{y/y_{max}}, \quad y \geq y_{max}.$$

EXAMPLE 12.7

Apply all three tail correction methods to the data used in Example 12.6. Assume that $\gamma = 22$.

Efron's method has $S_{20}(y) = 0$, $y \geq 15$. With $\gamma = 22$, Klein and Moeschberger's method has $S_{20}(y) = 0.089$, $15 \leq y < 22$. and $S_{20}(y) = 0$, $y \geq 22$. The exponential tail correction has $S_{20}(y) = (0.089)^{y/15}$, $y \geq 15$. □

Note that if there is no censoring ($b_i = 0$ for all i), then $r_{i+1} = r_i - s_i$, and for $y_j \leq y < y_{j+1}$

$$S_n(y) = \prod_{i=1}^{j} \left(\frac{r_i - s_i}{r_i} \right) = \prod_{i=1}^{j} \frac{r_{i+1}}{r_i} = \frac{r_{j+1}}{r_1}.$$

In this case r_{j+1} is the number of observations exceeding y and $r_1 = n$. Thus with no censoring the Kaplan–Meier estimate reduces to the empirical estimate of the previous section.

An alternative to the Kaplan–Meier estimator, called the Nelson–Åalen estimator [1],[**?**], is sometimes used. To motivate the estimator, note that if $S(y)$ is the survival function of a continuous distribution with failure rate $h(y)$, then $-\ln S(y) = H(y) = \int_0^y h(t)dt$ is called the cumulative hazard rate function. The discrete analog is, in the present context, given by $\sum_{i|y_i \le y} \lambda_i$, which can intuitively be estimated by replacing λ_i by its estimate $\hat{\lambda}_i = s_i/r_i$. The Nelson–Åalen estimator of $H(y)$ is thus defined for $y < y_{max}$ to be

$$\hat{H}(y) = \begin{cases} 0, & y < y_1, \\ \sum_{i=1}^{j} \hat{\lambda}_i = \sum_{i=1}^{j} \frac{s_i}{r_i}, & y_j \le y < y_{j+1}, \quad j = 1, 2, \ldots, k-1, \\ \sum_{i=1}^{k} \hat{\lambda}_i = \sum_{i=1}^{k} \frac{s_i}{r_i}, & y_k \le y < y_{max}. \end{cases}$$

That is $\hat{H}(y) = \sum_{i|y_i \le y} \hat{\lambda}_i$ for $y < y_{max}$, and the Nelson–Åalen estimator of the survival function is $\hat{S}(y) = \exp(-\hat{H}(y))$. For $y \ge y_{max}$, the situation is similar to that involving the Kaplan–Meier estimate in the sense that a tail correction of the type discussed earlier needs to be employed. Note that, unlike the Kaplan–Meier estimate, $\hat{S}(y_k) > 0$ so that a tail correction is always needed.

EXAMPLE 12.8

Determine the Nelson–Åalen estimates for the data in Example 12.6. Continue to assume $\gamma = 22$.

The estimates of the cumulative hazard function are given in Table 12.10. The estimated survival function is

$$\hat{S}_{20}(y) = \begin{cases} = 1, & y < 1, \\ = e^{-0.050} = 0.951, & 1 \le y < 2, \\ = e^{-0.103} = 0.902, & 2 \le y < 4, \\ = e^{-0.220} = 0.803, & 4 \le y < 5, \\ = e^{-0.297} = 0.743, & 5 \le y < 8, \\ = e^{-0.570} = 0.566, & 8 \le y < 9, \\ = e^{-1.070} = 0.343, & 9 \le y < 12, \\ = e^{-1.737} = 0.176, & 12 \le y < 15. \end{cases}$$

With regard to tail correction, Efron's method has $S_{20}(y) = 0$, $y \ge 15$. Klein and Moeschberger's method has $S_{20}(y) = 0.176$, $15 \le y < 22$. and $S_{20}(y) = 0$, $y \ge 22$. The exponential tail correction has $S_{20}(y) = (0.176)^{y/15}$, $y \ge 15$. □

Table 12.10 Nelson–Åalen estimates for Example 12.8

i	y_i	s_i	r_i	$\hat{H}_n(y_i)$
1	1	1	20	$1/20 = 0.050$
2	2	1	19	$0.05 + 1/19 = 0.103$
3	4	2	17	$0.103 + 2/17 = 0.220$
4	5	1	13	$0.220 + 1/13 = 0.297$
5	8	3	11	$0.297 + 3/11 = 0.570$
6	9	4	8	$0.570 + 4/8 = 1.070$
7	12	2	3	$1.070 + 2/3 = 1.737$

To assess the quality of the two estimators we will now consider estimation of the variance. Recall that for $y < y_{max}$, the Kaplan–Meier estimator may be expressed as

$$S_n(y) = \prod_{i|y_i \le y} \left(1 - \hat{\lambda}_i\right),$$

which is a function of the $\hat{\lambda}_i$s. Thus, to estimate the variance of $S_n(y)$, we first need the covariance matrix of $(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k)$. We estimate this from the "likelihood" using standard likelihood methods. Recall that

$$L\left(\lambda_1, \lambda_2, \ldots, \lambda_k\right) = \prod_{j=1}^{k} \lambda_j^{s_j} (1 - \lambda_j)^{r_j - s_j},$$

and thus $l = \ln L$ satisfies

$$l\left(\lambda_1, \lambda_2, \ldots, \lambda_k\right) = \sum_{j=1}^{k} \left[s_j \ln \lambda_j + (r_j - s_j) \ln(1 - \lambda_j)\right].$$

Thus,

$$\frac{\partial l}{\partial \lambda_i} = \frac{s_i}{\lambda_i} - \frac{r_i - s_i}{1 - \lambda_i}, \quad i = 1, 2, \ldots, k,$$

and

$$\frac{\partial^2 l}{\partial \lambda_i^2} = -\frac{s_i}{\lambda_i^2} - \frac{r_i - s_i}{(1 - \lambda_i)^2},$$

which, with λ_i replaced by $\hat{\lambda}_i = s_i/r_i$, becomes

$$\frac{\partial^2 l}{\partial \lambda_i^2} \Big|_{\lambda_i = \hat{\lambda}_i} = -\frac{s_i}{\hat{\lambda}_i^2} - \frac{r_i - s_i}{(1 - \hat{\lambda}_i)^2} = -\frac{r_i^3}{s_i(r_i - s_i)}.$$

For $i \neq m$,

$$\frac{\partial^2 l}{\partial \lambda_i \partial \lambda_m} = 0.$$

The observed information, evaluated at the maximum likelihood estimate, is thus a diagonal matrix, which when inverted yields the estimates

$$\text{Var}(\hat{\lambda}_i) \doteq \frac{s_i(r_i - s_i)}{r_i^3} = \frac{\hat{\lambda}_i(1 - \hat{\lambda}_i)}{r_i}, \quad i = 1, 2, \ldots, k,$$

and

$$\text{Cov}(\hat{\lambda}_i, \hat{\lambda}_m) \doteq 0, \quad i \neq m.$$

These results also follow directly from the binomial form of the likelihood.

Returning to the problem at hand, the delta method[4] gives the approximate variance of $f(\hat{\theta})$ as $\text{Var}[f(\hat{\theta})] \doteq [f'(\hat{\theta})]^2 \text{Var}(\hat{\theta})$, for an estimator $\hat{\theta}$.

To proceed, note that

$$\ln S_n(y) = \sum_{i|y_i \leq y} \ln(1 - \hat{\lambda}_i),$$

and since the $\hat{\lambda}_i$s are assumed to be approximately uncorrelated,

$$\text{Var}[\ln S_n(y)] \doteq \sum_{i|y_i \leq y} \text{Var}[\ln(1 - \hat{\lambda}_i)].$$

The choice $f(x) = \ln(1 - x)$ yields

$$\text{Var}[\ln(1 - \hat{\lambda}_i)] \doteq \frac{\text{Var}(\hat{\lambda}_i)}{(1 - \hat{\lambda}_i)^2} \doteq \frac{\hat{\lambda}_i}{r_i(1 - \hat{\lambda}_i)} = \frac{s_i}{r_i(r_i - s_i)},$$

implying that

$$\text{Var}[\ln S_n(y)] \doteq \sum_{i|y_i \leq y} \frac{s_i}{r_i(r_i - s_i)}.$$

Because $S_n(y) = \exp[\ln S_n(y)]$, the delta method with $f(x) = \exp(x)$ yields

$$\text{Var}[S_n(y)] \doteq \{\exp[\ln S_n(y)]\}^2 \text{Var}[\ln S_n(y)] = [S_n(y)]^2 \text{Var}[\ln S_n(y)],$$

This yields the final version of the estimate,

$$\widehat{\text{Var}}[S_n(y)] = [S_n(y)]^2 \sum_{i|y_i \leq y} \frac{s_i}{r_i(r_i - s_i)}. \tag{12.2}$$

Equation (12.2) holds for $y < y_{max}$ in all cases. However, if $y_{max} = y_k$ (that is, there are no censored observations after the last uncensored observation), then it holds for $y \leq y_{max} = y_k$. Hence the formula always holds for $y \leq y_k$.

Formula (12.2) is known as Greenwood's approximation to the variance of $S_n(y)$, and is known to often understate the true variance.

If there is no censoring, and we take $y_j \leq y < y_{j+1}$, then Greenwood's approximation yields

$$\widehat{\text{Var}}[S_n(y)] = \widehat{\text{Var}}[S_n(y_j)] = [S_n(y_j)]^2 \sum_{i=1}^{j} \frac{s_i}{r_i(r_i - s_i)},$$

which may be expressed (using $r_{i+1} = r_i - s_i$ due to no censoring) as

$$\begin{aligned} \widehat{\text{Var}}[S_n(y)] &= [S_n(y_j)]^2 \sum_{i=1}^{j} \left(\frac{1}{r_i - s_i} - \frac{1}{r_i} \right) \\ &= [S_n(y_j)]^2 \sum_{i=1}^{j} \left(\frac{1}{r_{i+1}} - \frac{1}{r_i} \right). \end{aligned}$$

[4]The delta method is presented in Section 11.6. Candidates for the Society of Actuaries Long-Term Actuarial Mathematics exam are not expected to be able to apply this method to given problems. In this chapter it is important to know that many of the variance formulas are approximate due to being derived by the delta method.

Because $r_1 = n$, this sum telescopes to give

$$\begin{aligned}\widehat{\text{Var}}[S_n(y)] &= [S_n(y_j)]^2 \left(\frac{1}{r_{j+1}} - \frac{1}{r_1}\right) \\ &= [S_n(y_j)]^2 \left[\frac{1}{nS_n(y_j)} - \frac{1}{n}\right] \\ &= \frac{S_n(y_j)[1 - S_n(y_j)]}{n},\end{aligned}$$

which is the same estimate obtained in Section 12.1, but was derived without use of the delta method.

We remark that in the case with $r_k = s_k$ (i.e., $S_n(y_k) = 0$), Greenwood's approximation cannot be used to estimate the variance of $S_n(y_k)$. In this case, $r_k - s_k$ is often replaced by r_k in the denominator.

Turning now to the Nelson–Åalen estimator, we note that

$$\hat{H}(y) = \sum_{i|y_i \le y} \hat{\lambda}_i,$$

and the same reasoning used for Kaplan–Meier implies that $\text{Var}[\hat{H}(y)] \doteq \sum_{i|y_i \le y} Var(\hat{\lambda}_i)$, yielding the estimate

$$\widehat{\text{Var}}[\hat{H}(y)] = \sum_{i|y_i \le y} \frac{s_i(r_i - s_i)}{r_i^3}, \tag{12.3}$$

which is referred to as Klein's estimate. A commonly used alternative estimate due to Åalen is obtained by replacing $r_i - s_i$ with r_i in the numerator.

We are typically more interested in $S(t)$ then $H(t)$. Because $\hat{S}(y) = \exp[-\hat{H}(y)]$, the delta method with $f(x) = e^{-x}$ yields Klein's survival function estimate

$$\text{Var}[\hat{S}(y)] \doteq [\exp(-\hat{H}(y))]^2 \widehat{\text{Var}}[\hat{H}(y)],$$

that is, the estimated variance is

$$\widehat{\text{Var}}[\hat{S}(y)] = [\hat{S}(y)]^2 \sum_{i|y_i \le y} \frac{s_i(r_i - s_i)}{r_i^3}, \quad y < y_{max}.$$

EXAMPLE 12.9

For the data of Example 12.5, estimate the variance of the Kaplan–Meier estimators of $S(2)$ and $S(9)$, and the Nelson–Åalen estimator of $S(2)$.

For the Kaplan–Meier estimators,

$$\begin{aligned}\widehat{\text{Var}}[S_{20}(2)] &= [S_{20}(2)]^2 \left[\frac{1}{20(19)} + \frac{1}{19(18)}\right] \\ &= (0.90)^2(0.00556) \\ &= 0.0045,\end{aligned}$$

$$\widehat{\text{Var}}[S_{20}(9)] = [S_{20}(9)]^2\left[\frac{1}{20(19)}+\frac{1}{19(18)}+\frac{2}{17(15)}+\frac{1}{13(12)}+\frac{3}{11(8)}+\frac{4}{8(4)}\right]$$
$$= (0.26656)^2(0.17890)$$
$$= 0.01271,$$

and for the Nelson–Åalen estimator,

$$\widehat{\text{Var}}[\hat{S}(2) = [\hat{S}(2)]^2\left[\frac{1(19)}{(20)^3}+\frac{1(18)}{(19)^3}\right]$$
$$= (0.90246)^2(0.00500)$$
$$= 0.00407. \qquad \square$$

Variance estimates for $y \geq y_{max}$ depend on the tail correction used. Efron's method gives an estimate of 0, which is not of interest in the present context. For the exponential tail correction in the Kaplan–Meier case, we have for $y \geq y_{max}$, $S_n(y) = S_n(y_k)^{y/y_{max}}$, and the delta method with $f(x) = x^{y/y_{max}}$ yields

$$\widehat{\text{Var}}[S_n(y) = \left[\frac{y}{y_{max}} S_n\left(y_k\right)^{\frac{y}{y_{max}}-1}\right]^2 \widehat{\text{Var}}[S_n(y_k)]$$
$$= \left(\frac{y}{y_{max}}\right)^2 [S_n(y)]^2 \sum_{i=1}^{k} \frac{s_i}{r_i(r_i-s_i)}$$
$$= \left(\frac{y}{y_{max}}\right)^2 \left[\frac{S_n(y)}{S_n(y_k)}\right]^2 \widehat{\text{Var}}[S_n(y_k)].$$

Likelihood methods typically result in approximate asymptotic normality of the estimates, and this is true for Kaplan–Meier and Nelson–Åalen estimates as well. Using the results of Example 12.9, an approximate 95% confidence interval for $S(9)$ is given by $S_{20}(9) \pm 1.96\sqrt{\{\text{Var}[S_{20}(9)]\}} = 0.26656 \pm 1.96\sqrt{0.01271} = (0.04557, 0.48755)$. For $S(2)$, the Nelson–Åalen estimate gives a confidence interval of $0.90246 \pm 1.96\sqrt{0.00407} = (0.77740, 1.02753)$, whereas that based on the Kaplan–Meier estimate is $0.90 \pm 1.96\sqrt{0.0045} = (0.76852, 1.03148)$. Clearly both confidence intervals for $S(2)$ are unsatisfactory, both including values greater than one.

An alternative approach can be constructed as follows, using the Kaplan–Meier estimate as an example.

Let $Y = \ln[-\ln S_n(y)]$. Using the delta method, the variance of Y can be approximated as follows. The function of interest is $f(x) = \ln(-\ln x)$. Its derivative is

$$f'(x) = \frac{1}{-\ln x}\frac{-1}{x} = \frac{1}{x \ln x}.$$

According to the delta method,

$$\widehat{\text{Var}}(Y) = \{f'[S_n(y)]\}^2 \text{Var}[S_n(y)] = \frac{\text{Var}[S_n(y)]}{[S_n(y)\ln S_n(y)]^2}.$$

Then, an approximate 95% confidence interval for $\theta = \ln[-\ln S(y)]$ is

$$\ln[-\ln S_n(y)] \pm 1.96\frac{\sqrt{\widehat{\text{Var}}[S_n(y)]}}{S_n(y)\ln S_n(y)}.$$

Because $S(y) = \exp(-e^Y)$, evaluating each endpoint of this formula provides a confidence interval for $S(y)$. For the upper limit we have (where $\hat{v} = \widehat{\text{Var}}[S_n(y)]$)

$$\exp\left\{-e^{\ln[-\ln S_n(y)]+1.96\sqrt{\hat{v}}/[S_n(y)\ln S_n(y)]}\right\} = \exp\left\{[\ln S_n(y)]e^{1.96\sqrt{\hat{v}}/[S_n(y)\ln S_n(y)]}\right\}$$
$$= S_n(y)^U, \quad U = \exp\left[\frac{1.96\sqrt{\hat{v}}}{S_n(y)\ln S_n(y)}\right].$$

Similarly, the lower limit is $S_n(y)^{1/U}$. This interval will always be inside the range 0–1 and is referred to as a **log-transformed confidence interval**.

■ **EXAMPLE 12.10**

Construct a 95% log transformed confidence interval for $S(2)$ in Example 12.9 based on the Kaplan–Meier estimator.

In this case $S_{20}(2) = 0.9$, and $U = \exp\left[\frac{1.96(0.06708)}{0.90\ln 0.90}\right] = 0.24994$. The lower limit is $(0.90)^{1/U} = 0.65604$ and the upper limit is $(0.90)^U = 0.97401$. □

For the Nelson–Åalen estimator, a similar log-transformed confidence interval for $H(y)$ has endpoints $\hat{H}(y)U$ where $U = \exp[\pm 1.96\sqrt{\widehat{\text{Var}}[\hat{H}(y)]}/\hat{H}(y)]$. Exponentiation of the negative of these endpoints yields a corresponding interval for $S(y)$.

■ **EXAMPLE 12.11**

Construct 95% linear and log-transformed confidence interval for $H(2)$ in Example 12.9 based on the Nelson–Åalen estimate. Then construct a 95% log-transformed confidence interval for $S(2)$.

From (12.3) a 95% linear confidence interval for $H(2)$ is

$$0.10263 \pm 1.96\sqrt{\frac{1(19)}{(20)^3} + \frac{1(18)}{(19)^3}} = 0.10263 \pm 1.96\sqrt{0.00500}$$
$$= (-0.03595, 0.24121)$$

which includes negative values. For a log-transformed confidence interval, we have

$$U = \exp[\pm 1.96\sqrt{.00500}/.10263]$$
$$= 0.25916 \text{ and } 3.8586.$$

The interval is from (0.10263) (0.25916) = 0.02660 to (0.10263) (3.85865) = 0.39601. The corresponding interval for $S(2)$ is from

$$\exp(-0.39601) = 0.67300 \text{ to } \exp(-0.02660) = 0.97375.$$ □

12.3.1 Exercises

12.10 (*) You are given the following times of first claim for five randomly selected auto insurance policies: 1, 2, 3, 4, 5. You are later told that one of the five times given is actually

Table 12.11 Data for Exercise 12.11.

Time t_j	Number of deaths s_j	Number at risk r_j
5	2	15
7	1	12
10	1	10
12	2	6

the time of policy lapse but you are not told which one. The smallest product-limit estimate of $S(4)$, the probability that the first claim occurs after time 4, would result if which of the given times arose from the lapsed policy?

12.11 (*) For a mortality study with right censored data, you are given the information in Table 12.11. Calculate the estimate of the survival function at time 12 using the Nelson–Åalen estimate.

12.12 (*) Let n be the number of lives observed from birth. None were censored and no two lives died at the same age. At the time of the ninth death, the Nelson–Åalen estimate of the cumulative hazard rate is 0.511, and at the time of the tenth death it is 0.588. Estimate the value of the survival function at the time of the third death.

12.13 (*) All members of a study joined at birth; however, some may leave the study by means other than death. At the time of the third death, there was one death (i.e., $s_3 = 1$); at the time of the fourth death, there were two deaths; and at the time of the fifth death, there was one death. The following product-limit estimates were obtained: $S_n(y_3) = 0.72$, $S_n(y_4) = 0.60$, and $S_n(y_5) = 0.50$. Determine the number of censored observations between times y_4 and y_5. Assume no observations were censored at the death times.

12.14 (*) A mortality study has right censored data and no left truncated data. Uncensored observations occurred at ages 3, 5, 6, and 10. The risk sets at these ages were 50, 49, k, and 21, respectively, while the number of deaths observed at these ages were 1, 3, 5, and 7, respectively. The Nelson–Åalen estimate of the survival function at time 10 is 0.575. Determine k.

12.15 (*) Consider the observations 2,500, 2,500, 2,500, 3,617, 3,662, 4,517, 5,000, 5,000, 6,010, 6,932, 7,500, and 7,500. No truncation is possible. First, determine the Nelson–Åalen estimate of the cumulative hazard rate function at 7,000 assuming all the observations are uncensored. Second, determine the same estimate, assuming the observations at 2,500, 5,000, and 7,500 were right censored.

12.16 (*) No observations in a data set are truncated. Some are right censored. You are given $s_3 = 1$, $s_4 = 3$, and the Kaplan–Meier estimates $S_n(y_3) = 0.65$, $S_n(y_4) = 0.50$, $S_n(y_5) = 0.25$. Also, between the observations y_4 and y_5 there are six right censored observations and no observations were right censored at the same value as an uncensored observation. Determine s_5.

12.17 For Data Set A determine the empirical estimate of the probability of having two or more accidents and estimate its variance.

Table 12.12 Data for Exercise 12.18.

Age	Number of deaths
2	1
3	1
5	1
7	2
10	1
12	2
13	1
14	1

12.18 (*) Ten individuals were observed from birth. All were observed until death. Table 12.12 gives the death ages. Let V_1 denote the estimated conditional variance of ${}_3\hat{q}_7$ if calculated without any distribution assumption. Let V_2 denote the conditional variance of ${}_3\hat{q}_7$ if calculated knowing that the survival function is $S(t) = 1-t/15$. Determine $V_1 - V_2$.

12.19 (*) Observations can be censored, but there is no truncation. Let y_j and y_{j+1} be consecutive death ages. A 95% linear confidence interval for $H(y_j)$ using the Klein estimator is $(0.07125, 0.22875)$ while a similar interval for $H(y_{j+1})$ is $(0.15985, 0.38257)$. Determine s_{j+1}.

12.20 (*) A mortality study is conducted on 50 lives, all observed from age 0. At age 15 there were two deaths; at age 17 there were three censored observations; at age 25 there were four deaths; at age 30 there were c censored observations; at age 32 there were eight deaths; and at age 40 there were two deaths. Let S be the product-limit estimate of $S(35)$ and let V be the Greenwood estimate of this estimator's variance. You are given $V/S^2 = 0.011467$. Determine the value of c.

12.21 (*) Fifteen cancer patients were observed from the time of diagnosis until the earlier of death or 36 months from diagnosis. Deaths occurred as follows: At 15 months there were two deaths; at 20 months there were three deaths; at 24 months there were two deaths; at 30 months there were d deaths; at 34 months there were two deaths; and at 36 months there was one death. The Nelson–Åalen estimate of $H(35)$ is 1.5641. Determine Klein's estimate of the variance of this estimator.

12.22 (*) Ten payments were recorded as follows: 4, 4, *5*, *5*, *5*, 8, *10*, *10*, 12, and 15, with the italicized values representing payments at a policy limit. There were no deductibles. Determine the product-limit estimate of $S(11)$ and Greenwood's approximation of its variance.

12.23 (*) All observations begin on day zero. Eight observations were *4*, *8*, 8, 12, *12*, 12, 22, and 36, with the italicized values representing right censored observations. Determine the Nelson–Åalen estimate of $H(12)$ and then determine a 90% linear confidence interval for the true value using Klein's variance estimate.

12.24 You are given the data in Table 12.13 based on 40 observations. Dashes indicate missing observations that must be deduced.

Table 12.13 Data for Exercise 12.24.

i	y_i	s_i	b_i	r_i
1	4	3	-	40
2	6	-	3	31
3	9	6	4	23
4	13	4	-	-
5	15	2	4	6

Table 12.14 Data for Exercise 12.25.

i	y_i	s_i	b_i	r_i
1	3	3	6	50
2	5	7	4	41
3	7	5	2	30
4	11	5	3	23
5	16	6	4	15
6	20	2	3	5

(a) Compute the Kaplan–Meier estimate $S_{40}(y_i)$ for $i = 1, 2, \ldots, 5$.

(b) Compute the Nelson–Åalen estimate $\hat{H}(y_i)$ for $i = 1, 2, \ldots, 5$.

(c) Compute $S_{40}(24)$ using the method of Brown, Hollander and Kowar.

(d) Compute Greenwood's approximation, $\widehat{Var}[S_{40}(15)]$.

(e) Compute a 95% linear confidence interval for $S(15)$ using the Kaplan–Meier estimate.

(f) Compute a 95% log-transformed confidence interval for $S(15)$ using the Kaplan–Meier estimate.

12.25 You are given the data in Table 12.14 based on 50 observations.

(a) Compute the Kaplan–Meier estimate $S_{50}(y_i)$ for $i = 1, 2, \ldots, 6$.

(b) Compute the Nelson–Åalen estimate $\hat{S}(y_i)$ for $i = 1, 2, \ldots, 6$.

(c) Compute $S_{50}(40)$ using Efron's tail correction, and also using the exponential tail correction of Brown, Hollander and Kowar.

(d) Compute Klein's survival function estimate of the variance of $\hat{S}(20)$.

(e) Compute a 95% log-transformed confidence interval for $H(20)$ based on the Nelson–Åalen estimate.

(f) Using the tail correction method of Brown, Hollander, and Kowar, estimate the variance of $\hat{S}(40)$.

12.26 Consider the estimator

$$\tilde{S}(y) = \prod_{i|y_i \le y} \phi\left(\hat{\lambda}_i\right)$$

where ϕ is differentiable.

(a) Show that $\tilde{S}(y)$ becomes the Kaplan–Meier estimator $S_n(y)$ when $\phi(x) = 1 - x$, and $\tilde{S}(y)$ becomes the Nelson–Åalen estimator $\hat{S}(y)$ when $\phi(x) = e^{-x}$.

(b) Derive the variance estimate

$$\text{Var}\left[\tilde{S}(y)\right] \approx \left[\tilde{S}(y)\right]^2 \sum_{i|y_i \le y} \left[\frac{\phi'\left(\hat{\lambda}_i\right)}{\phi\left(\hat{\lambda}_i\right)}\right]^2 \frac{s_i(r_i - s_i)}{r_i^3}.$$

(c) Consider $\tilde{S}_m(y) = \prod_{i|y_i \le y} \phi_m\left(\hat{\lambda}_i\right)$ with $\phi_m(x) = \sum_{j=0}^{m} (-x)^j/j!$ for $m = 0, 1, 2, \ldots$. Prove that $\hat{S}(y) \le \tilde{S}_{2m}(y)$ if $m = 0, 1, 2, \ldots$, and $\hat{S}(y) \ge \tilde{S}_{2m+1}(y)$ if $m = 0, 1, 2, \ldots$, and thus that $\hat{S}(y) \ge S_n(y)$ in particular. (Hint: prove by induction on m the identity $(-1)^m \left[\phi_m(x) - e^{-x}\right] \ge 0$ for $x \ge 0$ and $m = 0, 1, 2, \ldots$).

12.4 Empirical estimation of moments

In the previous section we focused on estimation of the survival function $S(y)$, or equivalently the cumulative distribution function $F(y) = 1 - S(y)$, of a random variable Y. In many actuarial applications, other quantities such as raw moments are of interest. Of central importance in this context is the mean, particularly for premium calculation in a loss modelling context.

For estimation of the mean with complete data $Y_1, Y_2, \ldots, Y_n$, an obvious (unbiased) estimation is $\hat{\mu} = (Y_1 + Y_2 + \cdots Y_n)/n$, but for incomplete data such as that of the previous section involving right censoring, other methods are needed. In this section we continue to assume that we have the setting described in the previous section, and we will capitalize on the results obtained for $S(y)$ there. To do so, we recall that, for random variables that take on only non-negative values, the mean satisfies

$$\mu = \text{E}(Y) = \int_0^\infty S(y)dy,$$

and empirical estimation of μ may be done by replacing $S(y)$ with an estimator such as the Kaplan–Meier estimator $S_n(y)$ or the Nelson–Åalen estimator $\hat{S}(y)$. To unify the approach, we will assume that $S(y)$ is estimated for $y < y_{max}$ by the estimator given in ***Exercise 3*** of Section 12.3, namely

$$\tilde{S}(y) = \prod_{i|y_i \le y} \phi(\hat{\lambda}_i),$$

where $\phi(x) = 1 - x$ for the Kaplan–Meier estimator and $\phi(x) = e^{-x}$ for the Nelson–Åalen Estimator. The mean is obtained by replacing $S(y)$ with $\tilde{S}(y)$ in the integrand. This yields the estimator

$$\tilde{\mu} = \int_0^\infty \tilde{S}(y)dy.$$

It is convenient to write

$$\tilde{\mu} = \int_0^{y_{max}} \tilde{S}(y)dy + \tilde{\tau}(y_{max})$$

where

$$\tau(x) = \int_x^\infty S(y)dy, \quad x \geq 0$$

and

$$\tilde{\tau}(x) = \int_x^\infty \tilde{S}(y)dy, \quad x \geq 0.$$

Anticipating what follows, we wish to evaluate $\tau(y_m)$ for $m = 1, 2, \ldots, k$. For $m = k$, we have that $S(y) = S(y_k) = \prod_{i=1}^{k} \phi(\lambda_i)$ for $y_k \leq y < y_{max}$. Thus

$$\tau(y_k) = \tau(y_{max}) + (y_{max} - y_k) \prod_{i=1}^{k} \phi(\lambda_i).$$

To evaluate $\tau(y_m)$ for $m = 1, 2, \ldots, k-1$, recall that $S(y) = 1$ for $0 \leq y < y_1$ and for $y_j \leq y < y_{j+1}$, $S(y) = S(y_j) = \prod_{i=1}^{j} \phi(\lambda_i)$. Thus,

$$\begin{aligned}
\tau(y_m) &= \tau(y_k) + \int_{y_m}^{y_k} S(y)dy \\
&= \tau(y_k) + \sum_{j=m+1}^{k} \int_{y_{j-1}}^{y_j} S(y)dy \\
&= \tau(y_k) + \sum_{j=m+1}^{k} \int_{y_{j-1}}^{y_j} S(y_{j-1})dy \\
&= \tau(y_k) + \sum_{j=m+1}^{k} (y_j - y_{j-1}) \prod_{i=1}^{j-1} \phi(\lambda_i).
\end{aligned}$$

For evaluation of μ, note that

$$\mu = \int_0^{y_1} S(y)dy + \tau(y_1) = y_1 + \tau(y_1),$$

and also that for $m = 1, 2, \ldots, k-1$,

$$\tau(y_m) = \tau(y_{m+1}) + (y_{m+1} - y_m)S(y_m),$$

a recursive formula, beginning with $\tau(y_k)$.

For the estimates themselves, $\tilde{S}(y_j) = \prod_{i=1}^{j} \phi(\hat{\lambda}_i)$, and the formulas above continue to hold when $\tau(y_m)$ is replaced by $\tilde{\tau}(y_m)$, $S(y)$ by $\tilde{S}(y)$, and μ by $\tilde{\mu}$.

The estimate of the mean μ clearly depends on $\tau(y_{max})$, which in turn depends on the tail correction, i.e. on $S(y)$ for $y \geq y_{max}$. If $S(y) = 0$ for $y \geq y_{max}$, as, for example, under Efron's tail correction, then $\tau(y_{max}) = 0$. Under Klein and Moeschberger's method with $S(y) = S(y_k)$ for $y_k \leq y < \gamma$, and $S(y) = 0$ for $y \geq \gamma$ where $\gamma > y_{max}$,

$$\tau(y_{max}) = \int_{y_{max}}^{\gamma} S(y)dy + \int_{\gamma}^{\infty} S(y)dy = (\gamma - y_{max})S(y_k).$$

For the exponential tail correction of Brown, Hollander, and Korwar, $S(y) = e^{-\beta y}$ for $y \geq y_{max}$ with $\beta = -\ln S(y_k)/y_{max}$. Thus

$$\tau(y_{max}) = \int_{y_{max}}^{\infty} e^{-\beta y} dy = \frac{1}{\beta} e^{-\beta y_{max}} = \frac{y_{max} S(y_k)}{-\ln S(y_k)}.$$

The following example illustrates the calculation of $\tilde{\mu}$, where all empirical quantities are obtained by substitution of estimates.

EXAMPLE 12.12

Calculate the estimates of the mean for the Nelson–Åalen estimate of Example 12.8 under the three tail corrections. Continue to assume that $\gamma = 22$.

We have, in terms of $\tilde{\tau}(y_{max}) = \tilde{\tau}(15)$,

$$\tilde{\tau}(y_7) = \tilde{\tau}(12) = (15-12)\hat{S}(12) + \tilde{\tau}(15) = 3(0.176) + \bar{\tau}(15) = 0.528 + \tilde{\tau}(15),$$
$$\tilde{\tau}(y_6) = \tilde{\tau}(9) = (12-9)\hat{S}(9) + \tilde{\tau}(12) = 3(0.343) + \tilde{\tau}(12) = 1.557 + \tilde{\tau}(15),$$
$$\tilde{\tau}(y_5) = \tilde{\tau}(8) = (9-8)\hat{S}(8) + \tilde{\tau}(9) = 1(0.566) + \tilde{\tau}(9) = 2.123 + \tilde{\tau}(15),$$
$$\tilde{\tau}(y_4) = \tilde{\tau}(5) = (8-5)\hat{S}(5) + \tilde{\tau}(8) = 3(0.743) + \tilde{\tau}(8) = 4.352 + \tilde{\tau}(15),$$
$$\tilde{\tau}(y_3) = \tilde{\tau}(4) = (5-4)\hat{S}(4) + \tilde{\tau}(5) = 1(0.803) + \tilde{\tau}(5) = 5.155 + \tilde{\tau}(15),$$
$$\tilde{\tau}(y_2) = \tilde{\tau}(2) = (4-2)\hat{S}(2) + \tilde{\tau}(4) = 2(0.902) + \tilde{\tau}(4) = 6.959 + \tilde{\tau}(15),$$
$$\tilde{\tau}(y_1) = \tilde{\tau}(1) = (2-1)\hat{S}(1) + \tilde{\tau}(2) = 1(0.951) + \tilde{\tau}(2) = 7.910 + \tilde{\tau}(15).$$

Then $\tilde{\mu} = y_1 + \tilde{\tau}(y_1) = 1 + 7.910 + \tilde{\tau}(15) = 8.910 + \tilde{\tau}(15)$. Under Efron's method, $\tilde{\tau}(15) = 0$ and $\tilde{\mu} = 8.910$, under Klein and Moeschberger's method $\tilde{\tau}(15) = (22-15)(.176) = 1.232$ and $\tilde{\mu} = 8.910+1.232 = 10.142$. Finally, for the exponential tail correction, $\tilde{\tau}(15) = (15)(.176)/(-\ln .176) = 1.520$, and therefore $\tilde{\mu} = 8.910 + 1.520 = 10.43$. □

We next consider estimation of the variance of $\tilde{\mu}$. Clearly, $\tilde{\mu}$ is a function of $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k)$, for which we have an estimate of the variance matrix from the previous section. In particular, $\text{Var}(\hat{\lambda})$ is a $k \times k$ diagonal matrix (i.e., all off diagonal elements are 0). Thus by the multivariate delta method with the $1 \times k$ matrix A with jth entry $\partial\mu/\partial\lambda_j$, the estimated variance of $\tilde{\mu}$ is

$$\text{Var}(\tilde{\mu}) \doteq \sum_{m=1}^{k} \left(\frac{\partial \mu}{\partial \lambda_m}\right)^2 \text{Var}(\hat{\lambda}_m),$$

and it remains to identify $\partial\mu/\partial\lambda_m$ for $m = 1, 2, \ldots, k$.

To begin, first note that μ depends on $\tau(y_k)$, which in turn depends on $S(y_k) = \prod_{i=1}^{k} \phi(\lambda_i)$ but also on the tail correction employed. As such, we will express the formulas in terms of $\partial\tau(y_k)/\partial\lambda_m$ for $m = 1, 2, \ldots, k$ for the moment. We first consider $\partial\mu/\partial\lambda_m$ for $m = 1, 2, \ldots, k-1$. Then

$$\mu = y_1 + \tau(y_1) = y_1 + \tau(y_k) + \sum_{j=2}^{k} (y_j - y_{j-1}) \prod_{i=1}^{j-1} \phi(\lambda_i).$$

In the above expression, λ_m does not appear for $j \leq m$. Thus,

$$\begin{aligned}\frac{\partial \mu}{\partial \lambda_m} &= \frac{\partial \tau(y_k)}{\partial \lambda_m} + \frac{\partial}{\partial \lambda_m}\left[\sum_{j=m+1}^{k}(y_j - y_{j-1})\prod_{i=1}^{j-1}\phi(\lambda_i)\right] \\ &= \frac{\partial \tau(y_k)}{\partial \lambda_m} + \frac{\phi'(\lambda_m)}{\phi(\lambda_m)}\sum_{j=m+1}^{k}(y_j - y_{j-1})\prod_{i=1}^{j-1}\phi(\lambda_i),\end{aligned}$$

and in terms of $\tau(y_m)$, this may be expressed as

$$\frac{\partial \mu}{\partial \lambda_m} = \frac{\partial \tau(y_k)}{\partial \lambda_m} + \frac{\phi'(\lambda_m)}{\phi(\lambda_m)}\left[\tau(y_m) - \tau(y_k)\right], \quad m = 1, 2, \ldots, k-1.$$

It is also useful to note that $\int_0^{y_k} S(y)dy$ does not involve λ_k and thus $\partial\mu/\partial\lambda_k = \partial\tau(y_k)/\partial\lambda_k$. The general variance formula thus may be written as

$$\text{Var}(\tilde{\mu}) \doteq \sum_{m=1}^{k}\left\{[\tau(y_m) - \tau(y_k)]\frac{\phi'(\lambda_m)}{\phi(\lambda_m)} + \frac{\partial \tau(y_k)}{\partial \lambda_m}\right\}^2 \text{Var}(\hat{\lambda}_m).$$

But

$$\frac{\partial \tau(y_k)}{\partial \lambda_m} = (y_{max} - y_k)\frac{\phi'(\lambda_m)}{\phi(\lambda_m)}\left[\prod_{i=1}^{k}\phi(\lambda_i)\right] + \frac{\partial \tau(y_{max})}{\partial \lambda_m},$$

and thus,

$$\frac{\partial \tau(y_k)}{\partial \lambda_m} = \frac{\phi'(\lambda_m)}{\phi(\lambda_m)}[\tau(y_k) - \tau(y_{max})] + \frac{\partial \tau(y_{max})}{\partial \lambda_m},$$

in turn implying that

$$\text{Var}(\tilde{\mu}) \doteq \sum_{m=1}^{k}\left\{[\tau(y_m) - \tau(y_{max})]\frac{\phi'(\lambda_m)}{\phi(\lambda_m)} + \frac{\partial \tau(y_{max})}{\partial \lambda_m}\right\}^2 \text{Var}(\hat{\lambda}_m).$$

The variance is estimated by replacing parameters with their estimates in the above formula. This yields

$$\widehat{\text{Var}}(\tilde{\mu}) = \sum_{m=1}^{k}\left\{[\tilde{\tau}(y_m) - \tilde{\tau}(y_{max})]\frac{\phi'(\hat{\lambda}_m)}{\phi(\hat{\lambda}_m)} + \frac{\partial \tilde{\tau}(y_{max})}{\partial \hat{\lambda}_m}\right\}^2 \frac{s_m(r_m - s_m)}{r_m^3},$$

where we understand $\partial\tilde{\tau}(y_{max})/\partial\hat{\lambda}_m$ to mean $\partial\tau(y_{max})/\partial\lambda_m$ with $\tau(y_{max})$ replaced by $\tilde{\tau}(y_{max})$ and λ_m by $\hat{\lambda}_m$.

If $\tilde{\tau}(y_{max}) = 0$, then

$$\widehat{\text{Var}}(\tilde{\mu}) = \sum_{m=1}^{k}[\tilde{\tau}^2(y_m)]^2\left[\frac{\phi'(\hat{\lambda}_m)}{\phi(\hat{\lambda}_m)}\right]^2 \frac{s_m(r_m - s_m)}{r_m^3},$$

a formula that further simplifies under the Kaplan–Meier assumption $\phi(x) = 1 - x$ to (recalling that $\hat{\lambda}_m = s_m/r_m$)

$$\widehat{\text{Var}}(\tilde{\mu}) = \sum_{m=1}^{k}[\tilde{\tau}(y_m)]^2\frac{s_m}{r_m(r_m - s_m)}.$$

We note that $\tilde{\tau}(y_{max}) = 0$ if no tail correction is necessary because $S(y_k) = 0$ (in which case $\tilde{\tau}(y_k) = 0$ as well and the upper limit or the summation is $k - 1$), or under Efron's approximation.

For Klein and Moeschberger's method,

$$\tau(y_{max}) = (\gamma - y_{max})S(y_k) = (\gamma - y_{max})\prod_{i=1}^{k}\phi(\lambda_i),$$

implying that

$$\frac{\partial\tau(y_{max})}{\partial\lambda_m} = \frac{\phi'(\lambda_m)}{\phi(\lambda_m)}\tau(y_{max}),$$

resulting in the same variance formula as under Efron's method (but $\tilde{\tau}(y_m)$ is increased by $\tilde{\tau}(y_{max})$ for this latter approximation).

Turning now to the exponential tail correction with $\tau(y_{max}) = -y_{max}S(y_k)/\ln S(y_k)$, recall that $S(y_k) = \prod_{i=1}^{k}\phi(\lambda_i)$ and $\ln S(y_k) = \sum_{i=1}^{k}\ln\phi(\lambda_i)$. Thus

$$\begin{aligned}\frac{\partial\tau(y_{max})}{\partial\lambda_m} &= -\frac{y_{max}}{\ln S(y_k)}\left[\frac{\partial}{\partial\lambda_m}S(y_k)\right] + \frac{y_{max}S(y_k)}{[\ln S(y_k)]^2}\left[\frac{\partial}{\partial\lambda_m}\ln S(y_k)\right]\\ &= -\frac{y_{max}}{\ln S(y_k)}\left[\frac{\phi'(\lambda_m)}{\phi(\lambda_m)}S(y_k)\right] + \frac{y_{max}S(y_k)}{[-\ln S(y_k)]^2}\left[\frac{\phi'(\lambda_m)}{\phi(\lambda_m)}\right]\\ &= \tau(y_{max})\frac{\phi'(\lambda_m)}{\phi(\lambda_m)} + \frac{[\tau(y_{max})]^2}{y_{max}S(y_k)}\frac{\phi'(\lambda_m)}{\phi(\lambda_m)}.\end{aligned}$$

Therefore, under the exponential tail correction, the general variance estimate becomes

$$\widehat{\text{Var}}(\tilde{\mu}) = \sum_{m=1}^{k}\left\{\tilde{\tau}(y_m) + \frac{[\tilde{\tau}(y_{max})]^2}{y_{max}\tilde{S}(y_k)}\right\}^2\left[\frac{\phi'(\hat{\lambda}_m)}{\phi(\hat{\lambda}_m)}\right]^2\frac{s_m(r_m - s_m)}{r_m^3}.$$

In the Nelson–Åalen case with $\phi(x) = e^{-x}$, the term $[\phi'(\hat{\lambda}_m)/\phi\left(\hat{\lambda}_m\right)]^2$ may obviously be omitted.

EXAMPLE 12.13

Estimate the variance of the means calculated in Example 12.12.

For both Efron's and Klein and Moeschberger's approaches, the formula becomes

$$\begin{aligned}\widehat{\text{Var}}(\tilde{\mu}) &= \sum_{m=1}^{7}[\tilde{\tau}(y_m)]^2\frac{s_m(r_m - s_m)}{r_m^3}\\ &= [7.910 + \tilde{\tau}(15)]^2\frac{(1)(19)}{(20)^3} + [6.959 + \tilde{\tau}(15)]^2\frac{(1)(18)}{(19)^3}\\ &\quad + [5.155 + \tilde{\tau}(15)]^2\frac{(2)(15)}{(17)^3} + [4.352 + \tilde{\tau}(15)]^2\frac{(1)(12)}{(13)^3}\\ &\quad + [2.123 + \tilde{\tau}(15)]^2\frac{(3)(8)}{(11)^3} + [1.557 + \tilde{\tau}(15)]^2\frac{(4)(4)}{(8)^3}\\ &\quad + [0.528 + \tilde{\tau}(15)]^2\frac{(2)(1)}{(3)^3}.\end{aligned}$$

Under Efron's approach, $\tilde{\tau}(15) = 0$, yielding the variance estimate of 0.719, whereas under Klein and Moeschberger's method, $\tilde{\tau}(15) = 1.232$, yielding the variance estimate of 1.469. For the exponential tail correction, we have $\tilde{\tau}(15) = 1.520$ and

$$\frac{[\tilde{\tau}(y_{max})]^2}{y_{max}\hat{S}(y_k)} = \frac{[\tilde{\tau}(15)]^2}{15\hat{S}(12)} = \frac{(1.520)^2}{15(0.176)} = 0.875,$$

and the variance estimate is thus

$$\widehat{\text{Var}}(\tilde{\mu}) = \sum_{m=1}^{7} [\tilde{\tau}(y_m) + 0.875]^2 \frac{s_m(r_m - s_m)}{r_m^3}.$$

That is,

$$\begin{aligned}
\widehat{\text{Var}}(\tilde{\mu}) =& (7.910 + 1.520 + 0.875)^2 \frac{(1)(19)}{(20)^3} + (6.959 + 1.520 + 0.875)^2 \frac{(1)(18)}{(19)^3} \\
&+ (5.155 + 1.520 + 0.875)^2 \frac{(2)(15)}{(17)^3} + (4.352 + 1.520 + 0.875)^2 \frac{(1)(12)}{(13)^3} \\
&+ (2.123 + 1.520 + 0.875)^2 \frac{(3)(8)}{(11)^3} + (1.557 + 1.520 + 0.875)^2 \frac{(4)(4)}{(8)^3} \\
&+ (0.528 + 1.520 + 0.875)^2 \frac{(2)(1)}{(3)^3} \\
=& 2.568.
\end{aligned}$$

□

For higher moments, a similar approach may be used. We have, for the α-th moment,

$$\text{E}(Y^\alpha) = \alpha \int_0^\infty y^{\alpha-1} S(y) dy,$$

which may be estimated (using $y_0 = 0$ without loss of generality) by

$$\begin{aligned}
\tilde{\mu}_\alpha &= \alpha \int_0^\infty y^{\alpha-1} \tilde{S}(y) dy \\
&= \alpha \left[\sum_{j=1}^{k} \int_{y_{j-1}}^{y_j} y^{\alpha-1} \tilde{S}(y) dy \right] + \alpha \int_{y_k}^{y_{max}} y^{\alpha-1} \tilde{S}(y) dy + \alpha \int_{y_{max}}^{\infty} y^{\alpha-1} \tilde{S}(y) dy \\
&= \alpha \left[\sum_{j=1}^{k} \int_{y_{j-1}}^{y_j} y^{\alpha-1} \tilde{S}(y_{j-1}) dy \right] + \alpha \int_{y_k}^{y_{max}} y^{\alpha-1} \tilde{S}(y_k) dy + \alpha \int_{y_{max}}^{\infty} y^{\alpha-1} \tilde{S}(y) dy \\
&= \alpha \left[\sum_{j=1}^{k} \tilde{S}(y_{j-1}) \int_{y_{j-1}}^{y_j} y^{\alpha-1} dy \right] + \alpha \tilde{S}(y_k) \int_{y_k}^{y_{max}} y^{\alpha-1} dy + \alpha \int_{y_{max}}^{\infty} y^{\alpha-1} \tilde{S}(y) dy \\
&= \left[\sum_{j=1}^{k} \left(y_j^\alpha - y_{j-1}^\alpha \right) \tilde{S}(y_{j-1}) \right] + (y_{max}^\alpha - y_k^\alpha) \tilde{S}(y_k) + \alpha \int_{y_{max}}^{\infty} y^{\alpha-1} \tilde{S}(y) dy \\
&= y_1^\alpha + \left[\sum_{j=2}^{k} (y_j^\alpha - y_{j-1}^\alpha) \prod_{i=1}^{j-1} \phi(\hat{\lambda}_i) \right] + \left[(y_{max}^\alpha - y_k^\alpha) \prod_{i=1}^{k} \phi(\hat{\lambda}_i) \right] + \alpha \int_{y_{max}}^{\infty} y^{\alpha-1} \tilde{S}(y) dy.
\end{aligned}$$

Again, the final integral on the right side depends on the tail correction, and is 0 if $\tilde{S}(y_k) = 0$ or under Efron's tail correction. It is useful to note that under the exponential tail correction, $\tilde{S}(y) = e^{-\tilde{\beta}y}$ for $y \geq y_{max}$ with $\tilde{\beta} = -\ln \tilde{S}(y_k)/y_{max}$, and if $\alpha = 1, 2, \ldots$,

$$\begin{aligned}\alpha \int_{y_{max}}^{\infty} y^{\alpha-1} e^{-\tilde{\beta}y} dy &= \frac{\alpha!}{\tilde{\beta}^{\alpha}} \int_{y_{max}}^{\infty} \frac{\tilde{\beta}^{\alpha} y^{\alpha-1} e^{-\tilde{\beta}y}}{(\alpha-1)!} dy \\ &= \frac{\alpha!}{\tilde{\beta}^{\alpha}} \sum_{j=0}^{\alpha-1} \frac{(\tilde{\beta} y_{max})^j e^{-\tilde{\beta} y_{max}}}{j!},\end{aligned}$$

using the tail function representation of the gamma distribution. That is, under the exponential tail correction,

$$\alpha \int_{y_{max}}^{\infty} y^{\alpha-1} \tilde{S}(y) dy = \alpha (y_{max})^{\alpha} \tilde{S}(y_k) \sum_{j=0}^{\alpha-1} \frac{1}{j!} \left[-\ln \tilde{S}(y_k)\right]^{j-\alpha}, \quad \alpha = 1, 2, 3, \ldots.$$

In particular, for the second moment ($\alpha = 2$),

$$2 \int_{y_{max}}^{\infty} y \tilde{S}(y) dy = 2 (y_{max})^2 \tilde{S}(y_k) \left\{ [-\ln \tilde{S}(y_k)]^{-1} + [-\ln \tilde{S}(y_k)]^{-2} \right\}.$$

Variance estimation for $\tilde{\mu}_{\alpha}$ may be done in a similar manner as for the mean, if desired.

12.4.1 Exercises

12.27 For the data of Exercise 12.24 and using the Kaplan–Meier estimate:

(a) Compute the mean survival time estimate assuming Efron's tail correction.

(b) Compute the mean survival time estimate using the exponential tail correction of Brown, Hollander, and Korwar.

(c) Estimate the variance of the estimate in (a).

12.28 For the data of Exercise 12.25, using the Nelson–Åalen estimate and the exponential tail correction of Brown, Hollander, and Korwar:

(a) Estimate the mean $\tilde{\mu}$.

(b) Estimate the variance of $\tilde{\mu}$ in (b).

12.29 For the data in Example 12.5 and subsequent examples, using the Nelson–Åalen estimate with the exponential tail correction of Brown, Hollander, and Korwar, estimate the variance of Y.

12.5 Empirical estimation with left truncated data

The results of Section 12.3 apply in situations when the data are (right) censored. In this section we discuss the situation where the data may also be (left) truncated. We have the following definitions.

Definition 12.10 *An observation is* ***truncated from below*** (*also called* ***left truncated***) *at* d *if when it is at or below* d *it is not recorded, but when it is above* d *it is recorded at its observed value.*

An observation is ***truncated from above*** (*also called* ***right truncated***) *at* u *if when it is at or above* u *it is not recorded, but when it is below* u *it is recorded at its observed value.*

In insurance survival data and claim data, the most common occurrences are left truncation and right censoring. Left truncation occurs when an ordinary deductible of d is applied. When a policyholder has a loss below d, he or she realizes no benefits will be paid and so does not inform the insurer. When the loss is above d, the amount of the loss is assumed to be reported.[5] A policy limit leads to an example of right censoring. When the amount of the loss equals or exceeds u, benefits beyond that value are not paid, and so the exact value is not recorded. However, it is known that a loss of at least u has occurred.

For decrement studies, such as of human mortality, it is impractical to follow people from birth to death. It is more common to follow a group of people of varying ages for a few years during the study period. When a person joins a study, he or she is alive at that time. This person's age at death must be at least as great as the age at entry to the study and thus has been left truncated. If the person is alive when the study ends, right censoring has occurred. The person's age at death is not known, but it is known that it is at least as large as the age when the study ended. Right censoring also affects those who leave the study prior to its end due to surrender. Note that this discussion could have been about other decrements, such as disability, policy surrender, or retirement.

Because left truncation and right censoring are the most common occurrences in actuarial work, they are the only cases that are covered in this section. To save words, *truncated* always means truncated from below and *censored* always means censored from above.

When trying to construct an empirical distribution from truncated or censored data, the first task is to create notation to represent the data. For individual (as opposed to grouped) data, the following facts are needed. First is the truncation point for that observation. Let that value be d_j for the jth observation. If there was no truncation, $d_j = 0$.[6] Next record the observation itself. The notation used depends on whether or not that observation was censored. If it was not censored, let its value be x_j. If it was censored, let its value be u_j. When this subject is presented more formally, a distinction is made between the case where the censoring point is known in advance and where it is not. For example, a liability insurance policy with a policy limit usually has the censoring point known prior to the receipt of any claims. By comparison, in a mortality study of insured lives, those that surrender their policy do so at an age that was not known when the policy was sold. In this chapter no distinction is made between the two cases.

To construct the estimate, the raw data must be summarized in a useful manner. The most interesting values are the uncensored observations. As in Section 12.3, let $y_1 < y_2 < \cdots < y_k$ be the k unique values of the x_js that appear in the sample, where k must be less than or equal to the number of uncensored observations. We also continue to let s_j be the number of times the uncensored observation y_j appears in the sample. Again, an important quantity is r_j, the number "at risk" at y_j. In a decrement study, r_j represents the number under observation and subject to the decrement at that time. To be under observation at y_j,

[5] In some cases an insured may elect to not report a loss that is slightly above the deductible if it is likely the next premium will be increased.

[6] Throughout, we assume that negative values are not possible.

an individual must (1) either be censored or have an observation that is on or after y_j and (2) not have a truncation value that is on or after y_j. That is,

$$r_j = (\text{number of } x_i\text{s } \geq y_j) + (\text{number of } u_i\text{s } \geq y_j) - (\text{number of } d_i\text{s} \geq y_j).$$

Alternatively, because the total number of d_is is equal to the total number of x_is and u_is, we also have

$$r_j = (\text{number of } d_i\text{s} < y_j) - (\text{number of } x_i\text{s } < y_j) - (\text{number of } u_i\text{s } < y_j). \quad (12.4)$$

This latter version is a bit easier to conceptualize because it includes all who have entered the study prior to the given age less those who have already left. The key point is that the number at risk is the number of people observed alive at age y_j. If the data are loss amounts, the risk set is the number of policies with observed loss amounts (either the actual amount or the maximum amount due to a policy limit) greater than or equal to y_j less those with deductibles greater than or equal to y_j. These relationships lead to a recursive version of the formula,

$$\begin{aligned} r_j = r_{j-1} &+ (\text{number of } d_i\text{s between } y_{j-1} \text{ and } y_j) \\ &- (\text{number of } x_i\text{s equal to } y_{j-1}) \\ &- (\text{number of } u_i\text{s between } y_{j-1} \text{ and } y_j), \end{aligned} \quad (12.5)$$

where *between* is interpreted to mean greater than or equal to y_{j-1} and less than y_j, and r_0 is set equal to zero.

A consequence of the above definitions is that if a censoring or truncation time equals that of a death, the death is assumed to have happened first. That is, the censored observation is considered at risk while the truncated observation is not.

The definition of r_j presented here is consistent with that in Section 12.3. That is, if $d_j = 0$ for all observations, the formulas presented here reduce match those presented earlier. The following example illustrates calculating the number at risk when there is trunction.

EXAMPLE 12.14

Using Data Set D, introduced in Chapter 10 and reproduced here as Table 12.15, calculate the r_j values using both (12.4) and (12.5). To provide some context and explain the entries in the table, think of this as a study of mortality by duration for a five-year term insurance policy. The study period is a fixed time period. Thus, policies 31–40 were already insured when first observed. There are two ways in which censoring occurs. For some (1, 2, 3, 5, etc.) the study either ended while they were still alive or they terminated their policies prior to the end of the five-year term. For others (19–32, 36, etc.) the five-year term ended with them still alive. Note that the eight observed deaths are at six distinct times and that $y_6 = 4.8$. Becuase the highest censoring time is 5.0, $y_{max} = 5.0$.

The calculations appear in Table 12.16. □

The approach to developing an empirical estimator of the survival function is to use the formulas developed in Section 12.3, but with the this more general definition of r_j. A

Table 12.15 Values for Example 12.14

i	d_i	x_i	u_i	i	d_i	x_i	u_i
1	0	–	0.1	16	0	4.8	–
2	0	–	0.5	17	0	–	4.8
3	0	–	0.8	18	0	–	4.8
4	0	0.8	–	19–30	0	–	5.0
5	0	–	1.8	31	0.3	–	5.0
6	0	–	1.8	32	0.7	–	5.0
7	0	–	2.1	33	1.0	4.1	–
8	0	–	2.5	34	1.8	3.1	–
9	0	–	2.8	35	2.1	–	3.9
10	0	2.9	–	36	2.9	–	5.0
11	0	2.9	–	37	2.9	–	4.8
12	0	–	3.9	38	3.2	4.0	–
13	0	4.0	–	39	3.4	–	5.0
14	0	–	4.0	40	3.9	–	5.0
15	0	–	4.1				

Table 12.16 Risk set calculations for Example 12.14

j	y_j	s_j	r_j
1	0.8	1	$32-0-2=30$ or $0+32-0-2=30$
2	2.9	2	$35-1-8=26$ or $30+3-1-6=26$
3	3.1	1	$37-3-8=26$ or $26+2-2-0=26$
4	4.0	2	$40-4-10=26$ or $26+3-1-2=26$
5	4.1	1	$40-6-11=23$ or $26+0-2-1=23$
6	4.8	1	$40-7-12=21$ or $23+0-1-1=21$

theoretical treatment that incorporates left truncation is considerably more complex. The reader is referred to [14] for details.

The formula for the Kaplan–Meier estimate is the same presented earlier, namely

$$S_n(y)=\begin{cases}1, & y<y_1,\\ \prod_{i=1}^{j}\left(1-\hat{\lambda}_i\right)=\prod_{i=1}^{j}\left(1-\frac{s_i}{r_i}\right), & y_j\le y<y_{j+1}, \quad j=1,2,\ldots,k-1,\\ \prod_{i=1}^{k}\left(1-\hat{\lambda}_i\right)=\prod_{i=1}^{k}\left(1-\frac{s_i}{r_i}\right), & y_k\le y<y_{max}.\end{cases}$$

The same tail corrections developed in Section 12.3 can be used for $y \ge y_{max}$ in cases where $S_n(y_k) > 0$.

EXAMPLE 12.15

Determine the Kaplan–Meier estimate for the data in Example 12.14

Based on the previous example, we have

$$S_{40}(y) = \begin{cases} 1, & 0 \le y < 0.8, \\ \frac{30-1}{30} = 0.9667, & 0.8 \le y < 2.9, \\ 0.9667\frac{26-2}{26} = 0.8923, & 2.9 \le y < 3.1, \\ 0.8923\frac{26-1}{26} = 0.8580, & 3.1 \le y < 4.0, \\ 0.8580\frac{26-2}{26} = 0.7920, & 4.0 \le y < 4.1, \\ 0.7920\frac{23-1}{23} = 0.7576, & 4.1 \le y < 4.8, \\ 0.7576\frac{21-1}{21} = 0.7215, & 4.8 \le y < 5.0. \end{cases}$$

□

In this example a tail correction is not needed because an estimate of survival beyond the five-year term is of no value when analyzing these policyholders.

The same analogy holds for the Nelson–Åalen estimator where the fomula for the cumulative hazard function remains

$$\hat{H}(y) = \begin{cases} 0, & y < y_1, \\ \sum_{i=1}^{j} \frac{s_i}{r_i}, & y_j \le y < y_{j+1}, \quad j = 1, 2, \ldots, k-1, \\ \sum_{i=1}^{k} \frac{s_i}{r_i}, & y_k \le y < y_{max}. \end{cases}$$

As before, $\hat{S}(y) = \exp[-\hat{H}(y)]$ for $y < y_{max}$ and for $y \ge y_{max}$ the same tail corrections can be used.

■ **EXAMPLE 12.16**

Determine the Nelson–Åalen estimate of the survival function for the data in Example 12.14.

The estimated functions are

$$\hat{H}(y) = \begin{cases} 0, & 0 \le y < 0.8, \\ \frac{1}{30} = 0.0333, & 0.8 \le y < 2.9, \\ 0.0333 + \frac{2}{26} = 0.1103, & 2.9 \le y < 3.1, \\ 0.1103 + \frac{1}{26} = 0.1487, & 3.1 \le y < 4.0, \\ 0.1487 + \frac{2}{26} = 0.2256, & 4.0 \le y < 4.1, \\ 0.2256 + \frac{1}{23} = 0.2691, & 4.1 \le y < 4.8, \\ 0.2691 + \frac{1}{21} = 0.3167, & 4.8 \le y < 5.0. \end{cases}$$

$$\hat{S}(y) = \begin{cases} 1, & 0 \le y < 0.8, \\ e^{-0.0333} = 0.9672, & 0.8 \le y < 2.9, \\ e^{-0.1103} = 0.8956, & 2.9 \le y < 3.1, \\ e^{-0.1487} = 0.8618, & 3.1 \le y < 4.0, \\ e^{-0.2256} = 0.7980, & 4.0 \le y < 4.1, \\ e^{-0.2691} = 0.7641, & 4.1 \le y < 4.8, \\ e^{-0.3167} = 0.7285, & 4.8 \le y < 5.0. \end{cases}$$

□

In this section the results were not formally developed, as was done for the the case with only right censored data. However, all the results, including formulas for moment estimates and estimates of the variance of the estimators hold when left truncation is added. However, it is important to note that when the data are truncated, the resulting distribution function is the distribution function of observations given that they are above the smallest truncation point (i.e., the smallest d value). Empirically, there is no information about observations below that value, and thus there can be no information for that range. Finally, if it turns out that there was no censoring or truncation, using the formulas in this section will lead to the same results as when using the empirical formulas in Section 12.1.

12.5.1 Exercises

12.30 Repeat Example 12.14, treating "surrender" as "death." The easiest way to do this is to reverse the x and u labels. In this case death produces censoring because those who die are lost to observation and thus their surrender time is never observed. Treat those who lasted the entire five years as surrenders at that time.

12.31 Determine the Kaplan–Meier estimate for the time to surrender for Data Set D. Treat those who lasted the entire five years as surrenders at that time.

Table 12.17 Data for Exercise 12.37.

y_j	r_j	s_j
1	100	15
8	65	20
17	40	13
25	31	31

12.32 Determine the Nelson–Åalen estimate of $H(t)$ and $S(t)$ for Data Set D where the variable is time to surrender.

12.33 Determine the Kaplan–Meier and Nelson–Åalen estimates of the distribution function of the amount of a workers compensation loss. First use the raw data from Data Set B. Then repeat the exercise, modifying the data by left truncation at 100 and right censoring at 1,000.

12.34 (*) Three hundred mice were observed at birth. An additional 20 mice were first observed at age 2 (days) and 30 more were first observed at age 4. There were 6 deaths at age 1, 10 at age 3, 10 at age 4, a at age 5, b at age 9, and 6 at age 12. In addition, 45 mice were lost to observation at age 7, 35 at age 10, and 15 at age 13. The following product-limit estimates were obtained: $S_{350}(7) = 0.892$ and $S_{350}(13) = 0.856$. Determine the values of a and b.

12.35 Construct 95% confidence intervals for $H(3)$ by both the linear and log-transformed formulas using all 40 observations in Data Set D with surrender being the variable of interest.

12.36 (*) For the interval from zero to one year, the exposure (r) is 15 and the number of deaths (s) is 3. For the interval from one to two years, the exposure is 80 and the number of deaths is 24. For two to three years, the values are 25 and 5; for three to four years, they are 60 and 6; and for four to five years, they are 10 and 3. Determine Greenwood's approximation to the variance of $\hat{S}(4)$.

12.37 (*) You are given the values in Table 12.17. Determine the standard deviation of the Nelson–Åalen estimator of the cumulative hazard function at time 20.

12.6 Kernel density models

One problem with empirical distributions is that they are always discrete. If it is known that the true distribution is continuous, the empirical distribution may be viewed as a poor approximation. In this section, a method of obtaining a smooth, empirical-like distribution, called a kernel density distribution, is introduced. We have the following definition.

Definition 12.11 *A **kernel smoothed distribution** is obtained by replacing each data point with a continuous random variable and then assigning probability $1/n$ to each such random variable. The random variables used must be identical except for a location or scale change that is related to its associated data point.*

Note that the empirical distirubtion is a special type of kernel smoothed distribution in which the random variable assignes probability 1 to the data point. With regard to kernel

smoothing, there are several distributions that could be used, three of which are introduced here.

While not necessary, it is customary that the continuous variable have a mean equal to the value of the point it replaces, ensuring that overall the kernel estimate has the same mean as the empirical estimate. One way to think about such a model is that it produces the final observed value in two steps. The first step is to draw a value at random from the empirical distribution. The second step is to draw a value at random from a continuous distribution whose mean is equal to the value drawn at the first step. The selected continuous distribution is called the *kernel*.

For notation, let $p(y_j)$ be the probability assigned to the value y_j $(j = 1, \ldots, k)$ by the empirical distribution. Let $K_y(x)$ be a distribution function for a continuous distribution such that its mean is y. Let $k_y(x)$ be the corresponding density function.

Definition 12.12 *A **kernel density estimator** of a distribution function is*

$$\hat{F}(x) = \sum_{j=1}^{k} p(y_j) K_{y_j}(x),$$

and the estimator of the density function is

$$\hat{f}(x) = \sum_{j=1}^{k} p(y_j) k_{y_j}(x).$$

The function $k_y(x)$ is called the kernel. Three kernels are now introduced: uniform, triangular, and gamma.

Definition 12.13 *The **uniform kernel** is given by*

$$k_y(x) = \begin{cases} 0, & x < y - b, \\ \dfrac{1}{2b}, & y - b \le x \le y + b, \\ 0, & x > y + b, \end{cases}$$

$$K_y(x) = \begin{cases} 0, & x < y - b, \\ \dfrac{x - y + b}{2b}, & y - b \le x \le y + b, \\ 1, & x > y + b. \end{cases}$$

*The **triangular kernel** is given by*

$$k_y(x) = \begin{cases} 0, & x < y-b, \\ \dfrac{x-y+b}{b^2}, & y-b \le x \le y, \\ \dfrac{y+b-x}{b^2}, & y \le x \le y+b, \\ 0, & x > y+b, \end{cases}$$

$$K_y(x) = \begin{cases} 0, & x < y-b, \\ \dfrac{(x-y+b)^2}{2b^2}, & y-b \le x \le y, \\ 1-\dfrac{(y+b-x)^2}{2b^2}, & y \le x \le y+b, \\ 1, & x > y+b. \end{cases}$$

*The **gamma** kernel is given by letting the kernel have a gamma distribution with shape parameter α and scale parameter y/α. That is,*

$$k_y(x) = \frac{x^{\alpha-1}e^{-x\alpha/y}}{(y/\alpha)^\alpha \Gamma(\alpha)} \text{ and } K_y(x) = \Gamma(\alpha; \alpha x/y).$$

Note that the gamma distribution has a mean of $\alpha(y/\alpha) = y$ and a variance of $\alpha(y/\alpha)^2 = y^2/\alpha$.

In each case there is a parameter that relates to the spread of the kernel. In the first two cases it is the value of $b > 0$, which is called the *bandwidth*. In the gamma case, the value of α controls the spread, with a larger value indicating a smaller spread. There are other kernels that cover the range from zero to infinity.

EXAMPLE 12.17

Determine the kernel density estimate for Example 12.2 using each of the three kernels.

The empirical distribution places probability $\frac{1}{8}$ at 1.0, $\frac{1}{8}$ at 1.3, $\frac{2}{8}$ at 1.5, $\frac{3}{8}$ at 2.1, and $\frac{1}{8}$ at 2.8. For a uniform kernel with a bandwidth of 0.1 we do not get much separation. The data point at 1.0 is replaced by a horizontal density function running from 0.9 to 1.1 with a height of $\frac{1}{8}\frac{1}{2(0.1)} = 0.625$. In comparison, with a bandwidth of 1.0, that same data point is replaced by a horizontal density function running from 0.0 to 2.0 with a height of $\frac{1}{8}\frac{1}{2(1)} = 0.0625$. Figures 12.3 and 12.4 provide plots of the density functions.

It should be clear that the larger bandwidth provides more smoothing. In the limit, as the bandwidth approaches zero, the kernel density estimate matches the empirical estimate. Note that, if the bandwidth is too large, probability will be assigned to negative values, which may be an undesirable result. Methods exist for dealing with that issue, but they are not presented here.

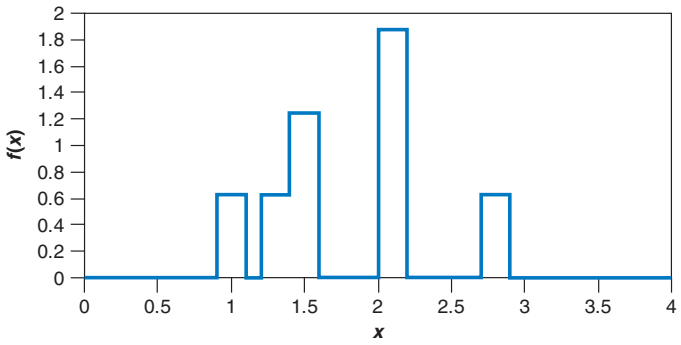

Figure 12.3 Uniform kernel density with bandwidth 0.1.

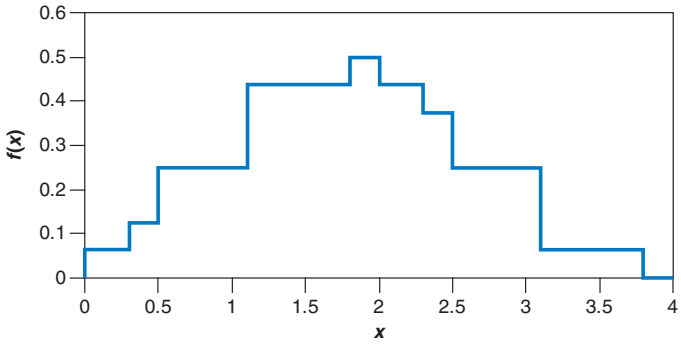

Figure 12.4 Uniform kernel density with bandwidth 1.0.

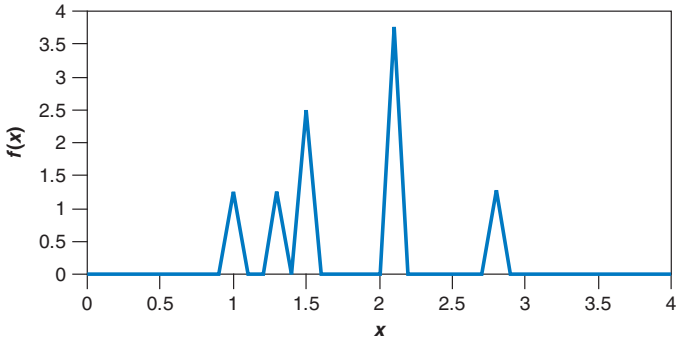

Figure 12.5 Triangular kernel density with bandwidth 0.1.

For the triangular kernel, each point is replaced by a triangle. Pictures for the same two bandwidths used previously appear in Figures 12.5 and 12.6.

Once again, the larger bandwidth provides more smoothing. The gamma kernel simply provides a mixture of gamma distributions where each data point provides the

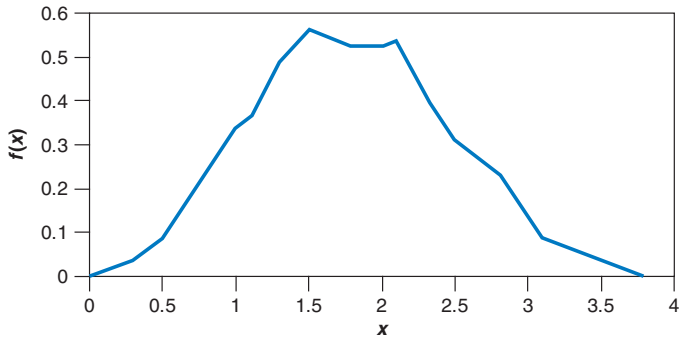

Figure 12.6 Triangular kernel density with bandwidth 1.0.

mean and the empirical probabilities provide the weights. The density function is

$$f_\alpha(x) = \sum_{j=1}^{5} p(y_j) \frac{x^{\alpha-1} e^{-x\alpha/y_j}}{(y_j/\alpha)^\alpha \Gamma(\alpha)}$$

and is graphed in Figures 12.7 and 12.8 for two α values.[7] For this kernel, decreasing the value of α increases the amount of smoothing. Further discussion of the gamma kernel can be found in [3], where the author recommends $\alpha = \sqrt{n}/(\hat{\mu}_4'/\hat{\mu}_2' - 1)^{1/2}$.□

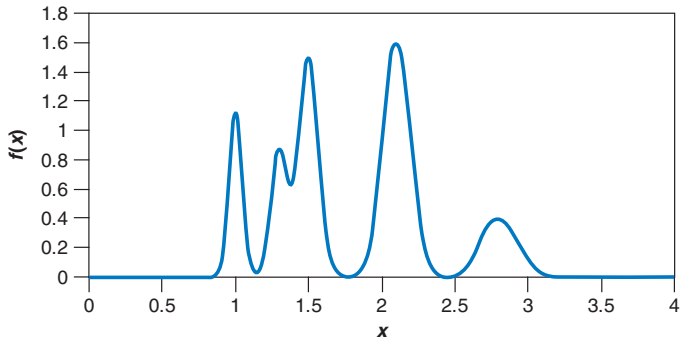

Figure 12.7 Gamma kernel density with $\alpha = 500$.

12.6.1 Exercises

12.38 Provide the formula for the Pareto kernel.

12.39 Construct a kernel density estimate for the time to surrender for Data Set D. Be aware of the fact that this is a mixed distribution (probability is continuous from 0 to 5 but is discrete at 5).

[7] When computing values of the density function, overflow and underflow problems can be reduced by computing the logarithm of the elements of the ratio, that is, $(\alpha - 1)\ln x - x\alpha/y_j - \alpha\ln(y_j/\alpha) - \ln\Gamma(\alpha)$, and then exponentiating the result.

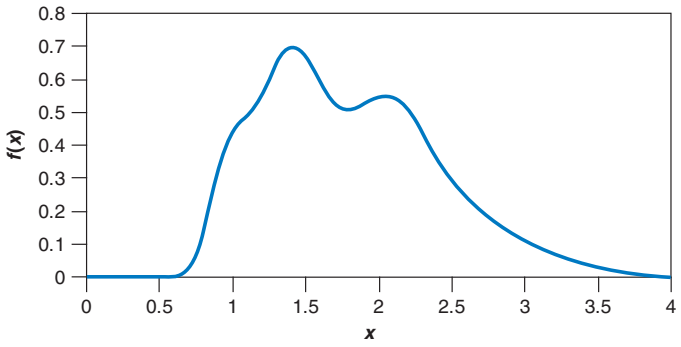

Figure 12.8 Gamma kernel density with $\alpha = 50$.

Table 12.18 Data for Exercise 12.40.

t_j	s_j	r_j
10	1	20
34	1	19
47	1	18
75	1	17
156	1	16
171	1	15

12.40 (*) You are given the data in Table 12.18 on time to death. Using the uniform kernel with a bandwidth of 60, determine $\hat{f}(100)$.

12.41 (*) You are given the following ages at time of death for 10 individuals: 25, 30, 35, 35, 37, 39, 45, 47, 49, and 55. Using a uniform kernel with a bandwidth of $b = 10$, determine the kernel density estimate of the probability of survival to age 40.

12.42 (*) Given the five observations 82, 126, 161, 294, and 384, determine each of the following estimates of $F(150)$:

(a) The empirical estimate.

(b) The kernel density estimate based on a uniform kernel with bandwidth $b = 50$.

(c) The kernel density estimate based on a triangular kernel with bandwidth $b = 50$.

12.7 Approximations for large data sets

12.7.1 Introduction

The discussion in this section is motivated by the circumstances that accompany the determination of a model for the time to death (or other decrement) for use in pricing, reserving, or funding insurance programs. The particular circumstances are:

- Values of the survival function are required only at discrete values, normally integral ages measured in years.

- A large volume of data has been collected over a fixed time period, with most observations truncated, censored, or both.
- No parametric distribution is available that provides an adequate model given the volume of available data.

These circumstances typically apply when an insurance company (or a group of insurance companies) conduct a mortality study based on the historical experience of a very large portfolio of life insurance policies. (For the remainder of this section we shall refer only to mortality. The results apply equally to the study of other decrements such as disablement or surrender.)

The typical mortality table is essentially a distribution function or a survival function with values presented only at integral ages. While there are parametric models that do well over parts of the age range (such as the Makeham model at ages over about 30), there are too many changes in the pattern from age 0 to ages well over 100 to allow for a simple functional description of the survival function.

The typical mortality study is conducted over a short period of time such as three to five years. For example, all persons who are covered by an insurance company's policies at some time from January 1, 2014, through December 31, 2016, might be included. Some of these persons may have purchased their policies prior to 2014 and were still covered when the study period started. During the study period some persons will die, some will cancel (surrender) their policy, some will have their policy expire due to policy provisions (such as with term insurance policies that expire during the study period), and some will still be insured when the study ends. It is assumed that if a policy is cancelled or expires the eventual age at death will not be known to the insurance company. Some persons will purchase their life insurance policy during the study period and be covered for some of the remaining part of the study period. These policies will be subject to the same decrements (death, surrender, expiration) as other policies. With regard to the age at death, almost every policy in the study will be left truncated.[8] If the policy was issued prior to 2014, the truncation point will be the age on January 1, 2014. For those who buy insurance during the study period, the truncation point is the age at which the contract begins. For any person who exits the study due to a cause other than death, their observation is right censored at the age of exit, because all that is known about them is that death will be at some unknown later age.

When no simple parametric distribution is appropriate and when large amounts of data are available, it is reasonable to use a nonparametric model because the large amount of data will ensure that key features of the survival function will be captured. Because there are both left truncation (due to the age at entry into the study) and right censoring (due to termination of the study at a fixed time), when there are large amounts of data, constructing the Kaplan–Meier estimate may require a very large amount of sorting and counting. Over the years a variety of methods have been introduced and entire texts have been written about the problem of constructing mortality tables from this kind of data (e.g., [**?**] and [15]). While the context for the examples presented here is the construction of mortality tables, the methods can apply any time the circumstances described previously apply.

We begin by examining the two ways in which data are usually collected. Estimators will be presented for both situations. The formulas will be presented in this section and their derivation and properties will be provided in Section 12.8. In all cases, a set of values

[8]The only exception would be a policy issued during the study period to someone just born.

(ages), $c_0 < c_1 < \cdots < c_k$ has been established in advance and the goal is to estimate the survival function at these values and no others (with some sort of interpolation to be used to provide intermediate values as needed). All of the methods are designed to estimate the conditional one-period probability of death, $q_j = [S(c_j) - S(c_{j+1})]/S(c_j)$, where j may refer to the interval and not to a particular age. From those values, $S(c_j)/S(c_0)$ can be evaluated as

$$\frac{S(c_j)}{S(c_0)} = \prod_{i=0}^{j-1}(1 - q_i).$$

12.7.2 Using individual data points

In this setting, data are recorded for each person observed. This approach is sometimes referred to as a *seriatim* method because the data points are analyzed as a series of individual observations. The estimator takes the form $\hat{q}_j = d_j/e_j$ where d_j is the number of observed deaths in the interval and e_j is a measure of exposure, representing the number of individuals who had a chance to be an observed death in that interval. Should a death occur at one of the boundary values between successive intervals, the death is counted in the preceding interval. When there are no entrants after age c_j into the interval and no exitants except for death during the interval (referred to as complete data), e_j represents the number of persons alive at age c_j and the number of deaths has a binomial distribution. With incomplete data it is necessary to determine a suitable convenient approximation, in particular, preferably one that requires only a single pass through the data set. To illustrate this challenge, consider the following example.

EXAMPLE 12.18

A mortality study is based on observations during the period January 1, 2014, through December 31, 2016. Five policies were observed, with the following information recorded. For simplicity, a date of 3-2000 is interpreted as March 1, 2000, and all events are treated as occurring on the first day of the month of occurrence. Furthermore, all months are treated as being one-twelfth of a year in length. Summarize the information in a manner that is sufficient for estimating mortality probabilities.

1. Born 4-1981, purchased insurance policy on 8-2013, was an active policyholder on 1-2017.
2. Born 6-1981, purchased insurance policy on 7-2013, died 9-2015.
3. Born 8-1981, purchased insurance policy on 2-2015, surrendered policy on 2-2016.
4. Born 5-1981, purchased insurance policy on 6-2014, died 3-2015.
5. Born 7-1981, purchased insurance policy on 3-2014, surrendered policy on 5-2016.

The key information is the age of the individual when first observed, the age when last observed, and the reason observation ended. For the five policies, using "x" for death and "s" for any other reason, the values are (where an age of 33-6 means 33 years and 6 months) $(32\text{-}9, 35\text{-}9, s)$, $(32\text{-}7, 34\text{-}3, x)$, $(33\text{-}6, 34\text{-}6, s)$, $(33\text{-}1, 33\text{-}10, x)$, $(32\text{-}8, 34\text{-}10, s)$. Note that policies 1 and 2 were purchased prior to the start of the study, so they are first observed when the study begins. No distinction needs to be

made between those whose observation ends by surrender as opposed to the ending of the study. □

The next step is to tally information for each age interval, building up totals for d_j and e_j. Counting deaths is straightforward. For exposures, there are two approaches that are commonly used.

Exact exposure method

Following this method, we set the exposure equal to the exact total time under observation within the age interval. When a death occurs, that person's exposure ends at the exact age of death. It will be shown in Section 12.8 that d_j/e_j is the maximum likelihood estimator of the hazard rate, under the assumption that the hazard rate is constant over the interval $(c_j, c_{j+1}]$. Further properties of this estimator will also be discussed in that section. The estimated hazard rate can then be converted into a conditional probability of death using the formula $q_j = 1 - \exp(-d_j/e_j)$.

Actuarial exposure method

Under this method, the exposure period for deaths extends to the end of the age interval, rather than the exact age at death. This has the advantage of reproducing the empirical estimator for complete data but has been shown to be an inconsistent estimator in other cases. In this case, the estimate of the conditional probability of death is obtained as $q_j = d_j/e_j$.

When the conditional probability of death is small, with a large number of observations, the choice of method is unlikely to materially affect the results.

EXAMPLE 12.19

Estimate all possible mortality rates at integral ages for Example 12.18 using both methods.

First observe that data are available for ages 32, 33, 34, and 35. The deaths are in the intervals 33-34 and 34-35. With a seriatim approach, each policy is analyzed and its contribution to the exposure for each interval added to the running total. For each age interval, the contributions for each interval are totalled, using the exact exposure method of recording time under observation. The numbers represent months.

32-33: $e_{32} = 3 + 5 + 0 + 0 + 4 = 12$.
33-34: $e_{33} = 12 + 12 + 6 + 9 + 12 = 51$.
34-35: $e_{34} = 12 + 3 + 6 + 0 + 10 = 31$.
35-36: $e_{35} = 9 + 0 + 0 + 0 + 0 = 9$.

As a check, the total contributions for the five policies are 36, 20, 12, 9, and 26, respectively, which matches the times on observation. The only two nonzero mortality probabilities are estimated as $\hat{q}_{33} = 1 - \exp[-1/(51/12)] = 0.20966$ and $\hat{q}_{34} = 1 - \exp[-1/(31/12)] = 0.32097$.

Under the actuarial exposure method the exposure for the interval 33-34 for the fourth person increases from 9 to 11 months for a total of 53 (note that it is not a full year of exposure as observation did not begin until age 33-1). For the interval 34-35 the exposure for the second person increases from 3 to 12 months for a total of 40. The mortality probability estimates are $\hat{q}_{33} = 1/(53/12) = 0.2264$ and $\hat{q}_{34} = 1/(40/12) = 0.3$. □

■ **EXAMPLE 12.20**

Use the actuarial exposure method to estimate all mortality probabilities for Data Set D.

Noting that the deaths at time 4.0 are assigned to the interval 3-4, the estimated values are $\hat{q}_0 = 1/29.4 = 0.0340$, $\hat{q}_1 = 0/28.8 = 0$, $\hat{q}_2 = 2/27.5 = 0.0.0727$, $\hat{q}_3 = 3/27.3 = 0.1099$, $\hat{q}_4 = 2/29.4 = 0.0930$. The corresponding survival probabilities are (with the Kaplan–Meier estimates in parentheses) $\hat{S}(1) = 0.9660$ (0.9667), $\hat{S}(2) = 0.9660$ (0.9667), $\hat{S}(3) = 0.8957$ (0.8923), $\hat{S}(4) = 0.7973$ (0.7920), $\hat{S}(5) = 0.7231$ (0.7215). □

12.7.2.1 Insuring ages While the examples have been in a life insurance context, the methodology applies to any situation with left truncation and right censoring. However, there is a situation that is specific to life insurance studies. Consider a one-year term insurance policy. Suppose an applicant was born on February 15, 1981, and applies for this insurance on October 15, 2016. Premiums are charged by whole-number ages. Some companies will use the age at the last birthday (35 in this case) and some will use the age at the nearest birthday (36 in this case). One company will base the premium on q_{35} and one on q_{36} when both should be using $q_{35.67}$, the applicant's true age. Suppose a company uses age last birthday. When estimating q_{35} it is not interested in the probability that a person exactly age 35 dies in the next year (the usual interpretation) but rather the probability that a random person who is assigned age 35 at issue (who can be anywhere between 35 and 36 years old) dies in the next year. One solution is to obtain a table based on exact ages, assume that the average applicant is 35.5, and use an interpolated value when determining premiums. A second solution is to perform the mortality study using the ages assigned by the company rather than the policyholder's true age. In the example, the applicant is considered to be exactly age 35 on October 15, 2016, and is thus assigned a new birthday of October 15, 1981. When this is done, the study is said to use *insuring ages* and the resulting values can be used directly for insurance calculations.

■ **EXAMPLE 12.21**

Suppose the company in Example 12.18 assigned insuring ages by age last birthday. Use the actuarial exposure method to estimate all possible mortality values.

1. Born 4-1981, purchased insurance policy on 8-2013, was an active policyholder on 1-2017. New birthday is 8-1981, enters at 32-5, exits at 35-5.

2. Born 6-1981, purchased insurance policy on 7-2013, died 9-2015. New birthday is 7-1981, enters at 32-6, dies at 34-2.

3. Born 8-1981, purchased insurance policy on 2-2015, surrendered policy on 2-2016. New birthday is 2-1982, enters at 33-0, exits at 34-0.

4. Born 5-1981, purchased insurance policy on 6-2014, died 3-2015. New birthday is 6-1981, enters at 33-0, dies at 33-9.

5. Born 7-1981, purchased insurance policy on 3-2014, surrendered policy on 5-2016. New birthday is 3-1982, enters at 32-0, exits at 34-2.

The exposures are now:

32-33: $e_{32} = 7 + 6 + 0 + 0 + 12 = 25$.
33-34: $e_{33} = 12 + 12 + 12 + 12 + 12 = 60$.
34-35: $e_{34} = 12 + 12 + 0 + 0 + 2 = 26$.
35-36: $e_{35} = 5 + 0 + 0 + 0 + 0 = 5$.

As expected, the exposures are assigned to younger age intervals, reflecting the fact that each applicant is assigned an age that is less than their true age. The estimates are $\hat{q}_{33} = 1/(60/12) = 0.2$ and $\hat{q}_{34} = 1/(26/12) = 0.4615$. □

Note that with insuring ages those who enter observation after the study begins are first observed on their newly assigned birthday. Thus there are no approximation issues with regard to those numbers.

12.7.2.2 Anniversary-based mortality studies Mortality studies described so far in this section are often called calendar-based or *date-to-date* studies because the period of study runs from one calendar date to another calendar date. It is also common for mortality studies of insured persons to use a different setup.

Instead of having the observations run from one calendar date to another calendar date, observation for a particular policyholder begins on the first policy anniversary following the fixed start date of the study and ends on the last anniversary prior to the study's fixed end date. Such studies are often called *anniversary-to-anniversary* studies. We can illustrate this through a previous example.

Consider Example 12.18 with the study now running from anniversaries in 2014 to anniversaries in 2016. The first policy comes under observation on 8-2014 at insuring age 33-0 and exits the study on 8-2016 at insuring age 35-0. Policyholder 2 begins observation on 7-2014 at insuring age 33-0. Policyholder 5 surrendered after the 2016 anniversary, so observation ends on 3-2016 at age 34-0. All other ages remain the same. In this setting, all subjects begin observations at an integral age and all who are active policyholders at the end of the study do so at an integral age. Only the ages of death and surrender may be other than integers (and note that with the actuarial exposure method, in calculating the exposure, deaths are placed at the next integral age). There is a price to be paid for this convenience. In a three-year study like the one in the example, no single policyholder can be observed for more than two years. In the date-to-date version, some policies will contribute three years of exposure.

All of the examples used one-year time periods. If the length of an interval $(c_{j+1} - c_j)$ is not equal to 1, an adjustment is necessary. Exposures should be the fraction of the period under observation and not the length of time.

12.7.3 Interval-based methods

Instead of recording the exact age at which an event happens, all that is recorded is the age interval in which it took place and the nature of the event. As with the individual method, for a portfolio of insurance policies, only running totals need to be recorded, and the end result is just four to six[9] numbers for each age interval:

1. The number of persons at the beginning of the interval carried over from the previous interval.
2. The number of additional persons entering at the beginning of the interval.
3. The number of persons entering during the interval.
4. The number of persons leaving by death during the interval.
5. The number of persons leaving during the interval for reasons other than death.
6. The number of persons leaving at the end of the interval by other than death.

■ **EXAMPLE 12.22**

Consider two versions of Example 12.18. In the first one use exact ages and a date-to-date study. Then use age last birthday insuring ages and an anniversary-to-anniversary study. For each, construct a table of the required values.

For the exact age study, the entry age, exit age, and cause of exit values for the five lives were $(32\text{-}9, 35\text{-}9, s)$, $(32\text{-}7, 34\text{-}3, x)$, $(33\text{-}6, 34\text{-}6, s)$, $(33\text{-}1, 33\text{-}10, x)$, and $(32\text{-}8, 34\text{-}10, s)$. There are no lives at age 32. Between ages 32 and 33, three lives enter and none leave. Between ages 33 and 34, two lives enter and one leaves by death. No lives enter or exit at age boundaries. The full set of values is in Table 12.19. For the insuring age study the values are $(33\text{-}0, 35\text{-}0, s)$, $(33\text{-}0, 34\text{-}2, x)$, $(33\text{-}0, 34\text{-}0, s)$, $(33\text{-}0, 33\text{-}9, x)$, $(32\text{-}0, 34\text{-}0, s)$, In this case there are several entries and exits at integral ages. The values are in Table 12.20. It is important to note that in the first table those who enter and leave do so during the age interval while in the second table those who enter do so at the beginning of the interval and those who leave by reason other than death do so at the end of the age interval. □

The analysis of this situation is relatively simple. For the interval from age c_j to age c_{j+1}, let P_j be the number of lives under observation at age c_j. This number includes those carried over from the prior interval as well as those entering at age c_j. Let n_j, d_j, and w_j be the number entering, dying, and leaving during the interval. Note that in general $P_{j+1} \neq P_j + n_j - d_j - w_j$ as the right-hand side must be adjusted by those who leave or enter at exact age c_{j+1}. Estimating the mortality probability depends on the method selected and an assumption about when the events that occur during the age interval take place.

One approach is to assume a uniform distribution of the events during the interval. For the exact exposure method, the P_j who start the interval have the potential to contribute

[9] As will be seen in the examples, not every category need be present in a given situation.

Table 12.19 Exact age values for Example 12.22.

Age	Number at age	Number entering	Number dying	Number leaving	Number at next age
32	0	3	0	0	3
33	3	2	1	0	4
34	4	0	1	2	1
35	1	0	0	1	0

Table 12.20 Insuring age values for Example 12.22.

Age	Number at age	Number entering	Number dying	Number leaving	Number at next age
32	0	1	0	0	1
33	1	4	1	2	2
34	2	0	1	1	0

a full unit of exposure and the n_j entrants during the year add another half-year each (on average). Similarly, those who die or leave subtract one-half year on average. Thus the net exposure is $P_j + (n_j - d_j - w_j)/2$. For the actuarial exposure method, those who die do not reduce the exposure, and it becomes $P_j + (n_j - w_j)/2$.

Another approach is to adapt the Kaplan–Meier estimator to this situation. Suppose the deaths all occur at midyear and all other decrements occur uniformly through the year. Then the risk set at midyear is $P_j + (n_j - w_j)/2$ and the estimator is the same as the actuarial estimator.

EXAMPLE 12.23

Apply both methods to the two data sets in Example 12.22.

For the exact age data the exact exposures for ages 32-35 are $0+(3-0-0)/2 = 1.5$, $3 + (2 - 1 - 0)/2 = 3.5$, $4 + (0 - 1 - 2)/2 = 2.5$, and $1 + (0 - 0 - 1)/2 = 0.5$, respectively. Using actuarial exposures changes the values to $0 + (3 - 0)/2 = 1.5$, $3 + (2 - 0)/2 = 4.0$, $4 + (0 - 2)/2 = 3.0$, and $1 + (0 - 1)/2 = 0.5$.

For the insuring age data, the exact exposures for ages 32-34 are $1+(0-0-0)/2 = 1.0$, $5 + (0 - 1 - 0)/2 = 4.5$, and $2 + (0 - 1 - 0)/2 = 1.5$. The actuarial exposures are $1 + (0 - 0)/2 = 1.0$, $5 + (0 - 0)/2 = 5.0$, and $2 + (0 - 0)/2 = 2.0$. Note that the entrants all occur at the beginning of the interval and so are included in P_j while those who leave do so at the end of the interval and are not included in w_j. □

The goal of all the estimation procedures in this book is to deduce the probability distribution for the random variable in the absence of truncation and censoring. For loss data, that would be the probabilities if there were no deductible or limit, that is, ground-up losses. For lifetime data it would be the probability distribution of the age at death if we could follow the person from birth to death. These are often referred to as *single-decrement probabilities* and are typically denoted q'_j in life insurance mathematics. In the life insurance context,

Table 12.21 Single-decrement mortality probabilities for Example 12.24.

j	P_j	n_j^b	n_j^m	d_j	w_j^m	w_j^e	$q_j'^{(d)}$
0	0	30	3	1	3	0	$1/30.0 = 0.0333$
1	29	0	1	0	2	0	$0/28.5 = 0.0000$
2	28	0	3	2	3	0	$2/28.0 = 0.0714$
3	26	0	3	3	3	0	$3/26.0 = 0.1154$
4	23	0	0	2	4	17	$2/21.0 = 0.0952$

the censoring rates are often as important as the mortality rates. For example, in the context of Data Set D, both time to death and time to withdrawal may be of interest. In the former case, withdrawals cause observations to be censored. In the latter case, censoring is caused by death. A superscript identifies the decrement of interest. For example, suppose the decrements were death (d) and withdrawal (w). Then $q_j'^{(w)}$ is the actuarial notation for probability that a person alive and insured at age c_j withdraws prior to age c_{j+1} in an environment where withdrawal is the only decrement, that is, that death is not possible. When the causes of censoring are other important decrements, an often-used assumption is that all the decrements are stochastically independent. That is, that they do not influence each other. For example, a person who withdraws at a particular age has the same probability of dying in the following month as a person who does not.

EXAMPLE 12.24

Estimate single-decrement probabilities using Data Set D and the actuarial method. Make reasonable assumptions.

First consider the decrement death. In the notation of this section, the relevant quantities are in Table 12.21. The notation n_j^b refers to entrants who do so at the beginning of the interval while n_j^m refers to those entering during the interval. These categories are based not on the time of the event but rather on its nature. The 30 policies that were issued during the study are at time zero by definition. A policy that was at duration 1.0 when the study began could have entered at any duration. Such policies are included in n_j^m for the preceding interval. For those who leave other than by death, the notation is w_j^m for those leaving during the interval and w_j^e for those doing so at the end. Again, those who do so by chance are assigned to the previous interval. Deaths are also assigned to the previous interval.

For withdrawals, the values of $q_j'^{(w)}$ are given in Table 12.22. To keep the notation consistent, d_j now refers to withdrawals and w_j refers to other decrements. □

EXAMPLE 12.25

Loss data for policies with deductibles of 0, 250, and 500 and policy limits of 5,000, 7,500, and 10,000 were collected. The data are in Table 12.23. Use the actuarial method to estimate the distribution function for losses.

The calculations appear in Table 12.24. Because the deductibles and limits are at the endpoints of intervals, the only reasonable assumption is the first one presented. □

Table 12.22 Single-decrement withdrawal probabilities for Example 12.24.

j	P_j	n_j^b	n_j^m	d_j	w_j^m	w_j^e	$q_j'^{(d)}$
0	0	30	3	3	1	0	$3/31.0 = 0.0968$
1	29	0	1	2	0	0	$2/29.5 = 0.0678$
2	28	0	3	3	2	0	$3/28.5 = 0.1053$
3	26	0	3	3	3	0	$3/26.0 = 0.1154$
4	23	0	0	4	2	17	$4/22.0 = 0.1818$

Table 12.23 Data for Example 12.25.

	Deductible			
Range	0	250	500	Total
0–100	15			15
100–250	16			16
250–500	34	96		130
500–1,000	73	175	251	499
1,000–2,500	131	339	478	948
2,500–5,000	83	213	311	607
5,000–7,500	12	48	88	148
7,500–10,000	1	4	11	16
At 5,000	7	17	18	42
At 7,500	5	10	15	30
At 10,000	2	1	4	7
Total	379	903	1,176	2,458

12.7.4 Exercises

12.43 Verify the calculations in Table 12.22.

12.44 For an anniversary-to-anniversary study, the values in Table 12.25 were obtained. Estimate $q_{45}'^{(d)}$ and $q_{46}'^{(d)}$ using the exact Kaplan–Meier estimate, exact exposure, and actuarial exposure.

12.45 Twenty-two insurance payments are recorded in Table 12.26. Use the fewest reasonable number of intervals and an interval-based method with actuarial exposure to estimate the probability that a policy with a deductible of 500 will have a payment in excess of 5,000.

12.46 (*) Nineteen losses were observed. Six had a deductible of 250, six had a deductible of 500, and seven had a deductible of 1,000. Three losses were paid at a policy limit, those values being 1,000, 2,750, and 5,500. For the 16 losses not paid at the limit, one was in the interval $(250, 500)$, two in $(500, 1{,}000)$, four in $(1{,}000, 2{,}750)$, seven in $(2{,}750, 5{,}500)$, one in $(5{,}500, 6{,}000)$, and one in $(6{,}000, 10{,}000)$. Estimate the probability that a policy with a deductible of 500 will have a claim payment in excess of 5,500.

Table 12.24 Calculations for Example 12.25.

c_j	P_j	n_j^b	d_j	w_j^e	$q_j'^{(d)}$	$\hat{F}(c_j)$
0	0	379	15	0	$15/379 = 0.0396$	0.0000
100	364	0	16	0	$16/364 = 0.0440$	0.0396
250	348	903	130	0	$130/1{,}251 = 0.1039$	0.0818
500	1,121	1,176	499	0	$499/2{,}297 = 0.2172$	0.1772
1,000	1,798	0	948	0	$948/1{,}798 = 0.5273$	0.3560
2,500	850	0	607	42	$607/850 = 0.7141$	0.6955
5,000	201	0	148	30	$148/201 = 0.7363$	0.9130
7,500	23	0	16	7	$16/23 = 0.6957$	0.9770
10,000	0					0.9930

Table 12.25 Data for Exercise 12.44.

d	u	x	d	u	x
45	46.0		45	45.8	
45	46.0		46	47.0	
45		45.3	46	47.0	
45		46.7	46	46.3	
45		45.4	46		46.2
45	47.0		46		46.4
45	45.4		46	46.9	

12.8 Maximum likelihood estimation of decrement probabilities

In Section 12.7 methods were introduced for estimating mortality probabilities with large data sets. One of the methods was a seriatim method using exact exposure. In this section that estimator will be shown to be maximum likelihood under a particular assumption. To do this, we need to develop some notation. Suppose we are interested in estimating the probability an individual alive at age a dies prior to age b where $a > b$. This is denoted $q = [S(a) - S(b)]/S(a)$. Let X be the random variable with survival function $S(x)$, the probability of surviving from birth to age x. Now let Y be the random variable X conditioned on $X > a$. Its survival function is $S_Y(y) = \Pr(X > y|X > a) = S(y)/S(a)$.

We now introduce a critical assumption about the shape of the survival function within the interval under consideration. Assume that $S_Y(y) = \exp[-(y-a)\lambda]$ for $a < y \le b$. This means that the survival function decreases exponentially within the interval. Equivalently, the hazard rate (called the force of mortality in life insurance mathematics) is assumed to be constant within the interval. Beyond b a different hazard rate will be used. Our objective is to estimate the conditional probability q. Thus we can perform the estimation using only data from and a functional form for this interval. Values of the survival function beyond b will not be needed.

Now consider data collected on n individuals, all of whom were observed during the age interval $(a, b]$. For individual j, let g_j be the age that the person was first observed within the interval and let h_j be the age the person was last observed within the interval (thus

Table 12.26 Data for Exercise 12.45.

Deductible	Payment[a]	Deductible	Payment
250	2,221	500	3,660
250	*2,500*	500	215
250	207	500	1,302
250	3,735	500	*10,000*
250	*5,000*	1,000	1,643
250	517	1,000	3,395
250	5,743	1,000	3,981
500	*2,500*	1,000	3,836
500	525	1,000	*5,000*
500	4,393	1,000	1,850
500	*5,000*	1,000	6,722

[a]Numbers in italics indicate that the amount paid was at the policy limit.

$a \le g_j < h_j \le b$). Let $\delta_j = 0$ if the individual was alive when last observed and $\delta_j = 1$ if the individual was last observed due to death. For this analysis, we assume that each individual's censoring age (everyone who does not die in the interval will be censored, either by reaching age b or through some event that removes them from observation) is known in advance. Thus the only random quantities are δ_j, and for individuals with $\delta_j = 1$, the age at death. The likelihood function is

$$\begin{aligned} L(\lambda) &= \prod_{j=1}^{n} \left[\frac{S(h_j)}{S(g_j)}\right]^{\delta_j} \left[\frac{f(h_j)}{S(g_j)}\right]^{1-\delta_j} = \prod_{j=1}^{n} \left[e^{-(h_j-g_j)\lambda}\right]^{\delta_j} \left[\lambda^{-1} e^{-(h_j-g_j)\lambda}\right]^{1-\delta_j} \\ &= \lambda^d \prod_{j=1}^{n} e^{-(h_j-g_j)\lambda} = \lambda^d \exp\left[-\sum_{j=1}^{n}(h_j-g_j)\lambda\right] = \lambda^d \exp(-e\lambda), \end{aligned}$$

where $d = \sum_{j=1}^{n} \delta_j$ is the number of observed deaths and $e = \sum_{j=1}^{n}(h_j - g_j)$ is the total time the individuals were observed in the interval (which was called exact exposure in Section 12.7). Taking logarithms, differentiating, and solving produces

$$\begin{aligned} l(\lambda) &= d \ln \lambda - e\lambda, \\ l'(\lambda) &= \frac{d}{\lambda} - e = 0, \\ \hat{\lambda} &= \frac{d}{e}. \end{aligned}$$

Finally, the maximum likelihood estimate of the probability of death is $\hat{q} = 1 - \exp[-(b-a)\hat{\lambda}] = 1 - \exp[-(b-a)d/e]$.

Studies often involve random censoring where individuals may leave for reasons other than death at times that were not known in advance. If all decrements (e.g., death, disability, and retirement) are stochastically independent (that is, the timing of one event does not influence any of the others), then the maximum likelihood estimator turns out to be identical to the one derived in this section. Although we do not derive the result, note that it

follows from the fact that the likelihood function can be decomposed into separate factors for each decrement.

The variance of this estimator can be approximated using the observed information approach. The second derivative of the loglikelihood function is

$$l''(\lambda) = -\frac{d}{\lambda^2}.$$

Substituting the estimator produces

$$l''(\hat{\lambda}) = -\frac{d}{(d/e)^2} = -\frac{e^2}{d}$$

and so $\widehat{\text{Var}}(\hat{\lambda}) = d/e^2$. Using the delta method,

$$\begin{aligned} \frac{dq}{d\lambda} &= -(b-a)\exp[-(b-a)\lambda], \\ \widehat{\text{Var}}(\hat{q}) &= (1-\hat{q})^2(b-a)^2\frac{d}{e^2}. \end{aligned}$$

Recall from Section 12.7 that there is an alternative called actuarial exposure with $\hat{q} = (b-a)d/e$ with e calculated in a different manner. When analyzing results from this approach, it is common to assume that d is the result of a binomial experiment with sample size $e/(b-a)$. Then,

$$\widehat{\text{Var}}(\hat{q}) = \frac{\hat{q}(1-\hat{q})}{e/(b-a)}.$$

If the $1-\hat{q}$ terms are dropped (and they are often close to 1) the two variance formulas are identical (noting that the values of e will be slightly different).

EXAMPLE 12.26

A pension plan assigns every employee an integral age on their date of hire (thus retirement age is always reached on an anniversary of being hired and not on a birthday). Because of the nature of the employee contracts, employees can only quit their job on annual anniversaries of being hired or six months after an anniversary. They can die at any age. Using assigned ages, a mortality study observed employees from age 35 to age 36. There were 10,000 who began at age 35. Of them, 100 died between ages 35 and 35.5 at an average age of 35.27, 400 quit at age 35.5, 110 died between ages 35.5 and 36 at an average age of 35.78, and the remaining 9,390 were alive and employed at age 36. Using both exact and actuarial exposure, estimate the single-decrement value of q_{35} and the standard deviation of the estimator.

The exact exposure is $100(0.27) + 400(0.5) + 110(0.78) + 9{,}390(1) = 9{,}702.8$. Then $\hat{q}_{35} = 1 - \exp(-210/9{,}702.8) = 0.02141$. The estimated standard deviation is $0.97859(210)^{1/2}/9{,}702.8 = 0.00146$. Recall that actuarial exposure assumes deaths occur at the end of the interval. The exposure is now $100(1) + 400(0.5) + 110(1) + 9{,}390(1) = 9{,}800$. Then $\hat{q}_{35} = 210/9{,}800 = 0.02143$. The estimated standard deviation is $\sqrt{0.02143(0.97857)/9,800} = 0.00146$. □

12.8.1 Exercise

12.47 In Exercise 12.44 mortality estimates for q_{45} and q_{46} were obtained by Kaplan–Meier, exact exposure, and actuarial exposure. Approximate the variances of these estimates (using Greenwood's formula in the Kaplan–Meier case).

12.9 Estimation of transition intensities

The discussion to this point has concerned estimating the probability of a decrement in the absence of other decrements. An unstated assumption was that the environment in which the observations are made is one where once any decrement occurs, the individual is no longer observed.

A common, and more complex, situation is one where after a decrement occurs, the individual remains under observation with the possibility of further decrements. A simple example is a disability income policy. A healthy individual can die, become disabled, or surrender their policy. Those who become disabled continue to be observed with possible decrements being recovery or death. Scenarios such as this are referred to as **multi-state models**. Such models are discussed in detail in [4]. In this section we cover estimation of the transition intensities associated with such models. The results presented are based on [20].

For notation, let the possible states be $\{0, 1, \ldots, k\}$ and let μ_x^{ij} be the force of transition to state j for an individual who is currently between ages x and $x+1$ and is in state i. This notation is based on an assumption that the force of transition is constant over an integral age. This is similar to the earlier assumption that the force of decrement is constant over a given age.

■ **EXAMPLE 12.27**

For the disability income policy previously described, identify the states and indicate which transition intensities likely have positive values. Of those with positive values, which can be estimated from available data?

The assigning of numbers to states is arbitrary. For this situation, consider 0 = healthy and an active policyholder, 1 = disabled and receiving benefits, 2 = surrendered, and 3 = dead. Possible transitions are $0 \to 1$, $0 \to 2$, $0 \to 3$, $1 \to 0$, $1 \to 3$, $2 \to 3$. All but the last one can be observed (we are unlikely to know when a surrendered policyholder dies). States 0 and 1 were carefully defined to exclude cases where a surrendered policyhodler becomes disabled or healthy. □

While not shown here, maximum likelihood estimates turn out to be based on exact exposure for the time spent in each state. For those between ages x and $x+1$ (which can be generalized for periods of other than one year), let T_i be the total time policyholders are observed in state i and d_{ij} be the number of observed transitions from state i to state j. Then $\hat{\mu}_x^{ij} = d_{ij}/T_i$. Similarly $\widehat{\text{Var}}(\hat{\mu}_x^{ij}) = d_{ij}/{T_i}^2$.

■ **EXAMPLE 12.28**

Table 12.27 Calculations for Exercise 12.1.

Payment range	Number of payments	Ogive value	Histogram value
0–25	6	$\frac{6}{392} = 0.0153$	$\frac{6}{392(25)} = 0.000612$
25–50	24	$\frac{30}{392} = 0.0765$	$\frac{24}{392(25)} = 0.002449$
50–75	30	$\frac{60}{392} = 0.1531$	$\frac{30}{392(25)} = 0.003061$
75–100	31	$\frac{91}{392} = 0.2321$	$\frac{31}{392(25)} = 0.003163$
100–150	57	$\frac{148}{392} = 0.3776$	$\frac{57}{392(50)} = 0.002908$
150–250	80	$\frac{228}{392} = 0.5816$	$\frac{80}{392(100)} = 0.002041$
250–500	85	$\frac{313}{392} = 0.7985$	$\frac{85}{392(250)} = 0.000867$
500–1,000	54	$\frac{367}{392} = 0.9362$	$\frac{54}{392(500)} = 0.000276$
1,000–2,000	15	$\frac{382}{392} = 0.9745$	$\frac{15}{392(1000)} = 0.000038$
2,000–4,000	10	$\frac{392}{392} = 1.0000$	$\frac{10}{392(2000)} = 0.000013$

Consider five policyholders who, between ages 50 and 51, are observed to do the following (decimals are fractions of a year):

- Disabled at age 50, dies at age 50.27.
- Healthy at age 50, disabled at age 50.34, dies at age 50.78.
- Healthy at age 50, surrendered at age 50.80.
- Purchases policy (healthy) at age 50.31, healthy at age 51.
- Healthy at age 50, disabled at 50.12, healthy at 50.45, dies at age 50.91.

Calculate the maximum likelihood estimates of the transition intensities.

Time spent healthy is $50.34-50+50.80-50+51-50.31+50.12-50+50.91-50.45 = 2.41$. While healthy, two became disabled, one surrendered, and one died. Hence the estimated transition intensities are $\hat{\mu}_{50}^{01} = 2/2.41$, $\hat{\mu}_{50}^{02} = 1/2.41$, $\hat{\mu}_{50}^{03} = 1/2.41$. Time spent disabled is $50.27 - 50 + 50.78 - 50.34 + 50.45 - 50.12 = 1.04$. While disabled, one became healthy and two died. The estimates are $\hat{\mu}_{50}^{10} = 2/1.04$, $\hat{\mu}_{50}^{13} = 1/1.04$. □

Constructing interval-based methods is more difficult because it is unclear when to place the transitions. Those who make one transition in the year may be reasonably placed at mid-age. However, those who make two transitions would more reasonably be placed at the one-third and two-thirds points. This would require careful data-keeping and the counting of many different cases.

12.10 Solutions to Exercises

12.1 There are 392 observations and the calculations are in Table 12.27. For each interval, the ogive value is for the right-hand endpoint of the interval, while the histogram value is for the entire interval. Graphs of the ogive and histogram appear in Figures 12.9 and 12.10.

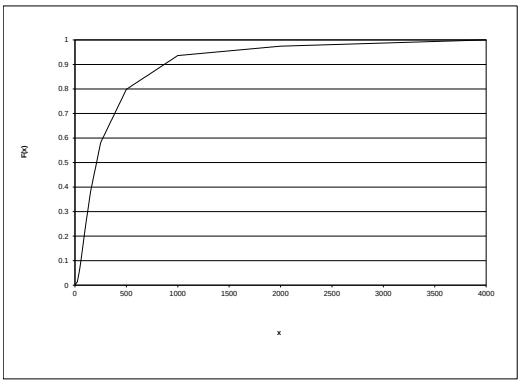

Figure 12.9 Ogive for Exercise 12.1.

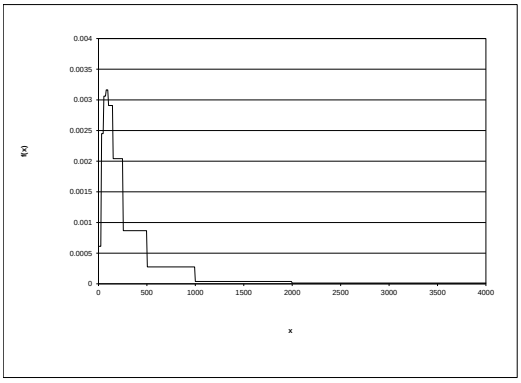

Figure 12.10 Histogram for Exercise 12.1.

12.2 (a) The ogive connects the points $(0.5, 0)$, $(2.5, 0.35)$, $(8.5, 0.65)$, $(15.5, 0.85)$, and $(29.5, 1)$.

(b) The histogram has height $0.35/2 = 0.175$ on the interval $(0.5, 2.5)$, height $0.3/6 = 0.05$ on the interval $(2.5, 8.5)$, height $0.2/7 = 0.028571$ on the interval $(8.5, 15.1)$, and height $0.15/14 = 0.010714$ on the interval $(15.5, 29.5)$.

12.3 The plot appears in Figure 12.11. The points are the complements of the survival probabilities at the indicated times.

Because one curve lies completely above the other it appears possible that original issues have a shorter lifetime.

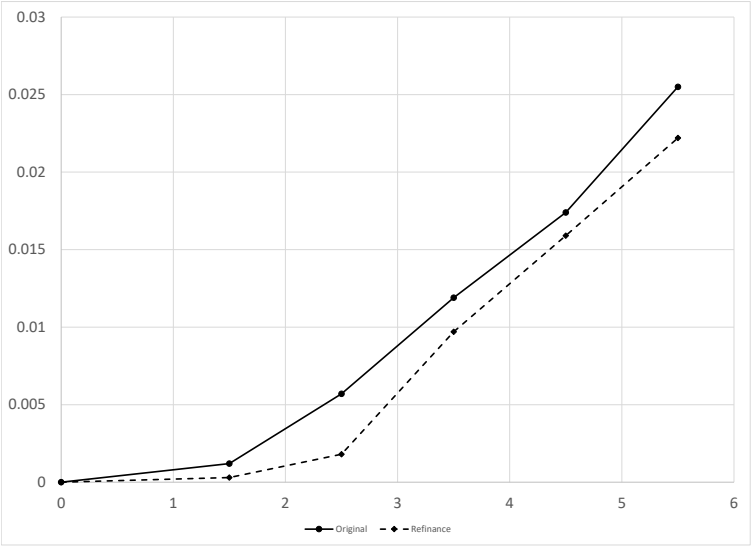

Figure 12.11 Ogive for mortgage lifetime for Exercise 12.3.

12.4 The heights of the histogram bars are, respectively, $0.5/2 = 0.25$, $0.2/8 = 0.025$, $0.2/90 = 0.00222$, $0.1/900 = 0.000111$. The histogram appears in Figure 12.12.

12.5 The empirical model places probability $1/n$ at each data point. Then

$$\begin{aligned} \mathrm{E}(X \wedge 2) &= \sum_{x_j<2} x_j(1/40) + \sum_{x_j \geq 2} 2(1/40) \\ &= (20+15)(1/40) + (14)(2)(1/40) \\ &= 1.575. \end{aligned}$$

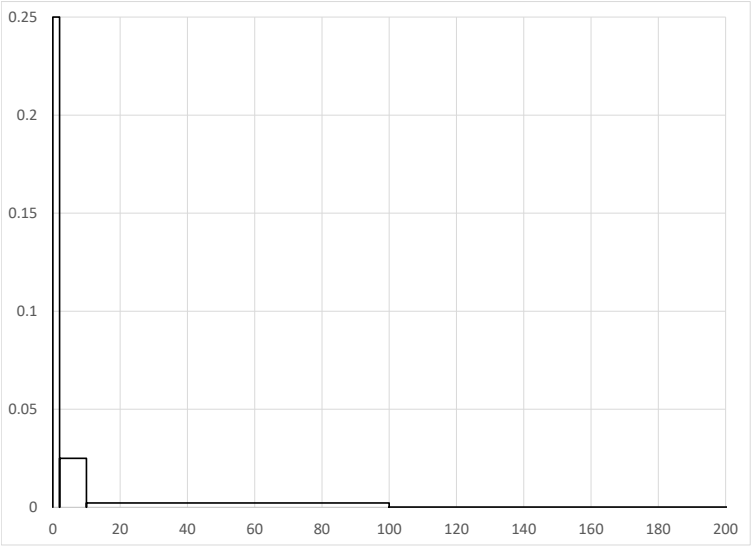

Figure 12.12 Histogram for Exercise 12.4.

12.6 We have

$$\begin{aligned}
\mathrm{E}(X \wedge 7{,}000) &= \frac{1}{2{,}000}\left(\sum_{x_j \le 7000} x_j + \sum_{x_j > 7000} 7{,}000\right) \\
&= \frac{1}{2{,}000}\left(\sum_{x_j \le 6{,}000} x_j + \sum_{x_j > 6{,}000} 6{,}000\right. \\
&\qquad \left. + \sum_{6000 < x_j \le 7000} (x_j - 6{,}000) + \sum_{x_j > 7000} 1{,}000\right) \\
&= \mathrm{E}(X \wedge 6{,}000) + [200{,}000 - 30(6{,}000) + 270(1{,}000)]/2{,}000 \\
&= 1m, 955.
\end{aligned}$$

12.7 Let n be the sample size. The equations are

$$0.21 = \frac{36}{n} + \frac{0.4x}{n} \text{ and } 0.51 = \frac{36}{n} + \frac{x}{n} + \frac{0.6y}{n}.$$

Also, $n = 200 + x + y$. The equations can be rewritten as

$$0.21(200 + x + y) = 36 + 0.4x \text{ and } 0.51(200 + x + y) = 36 + x + 0.6y.$$

The linear equations can be solved for $x = 120$.

12.8 (a) $\hat{\mu} = \sum x_i/35 = 204{,}900$. $\hat{\mu}'_2 = \sum x_i^2/35 = 1.4134 \times 10^{11}$.
$\hat{\sigma} = 325{,}807$. $\hat{\mu}'_3 = 1.70087 \times 10^{17}$, $\hat{\mu}_3 = 9.62339 \times 10^{16}$,
$\hat{c} = 325{,}807/204{,}900 = 1.590078$, $\hat{\gamma}_1 = 2.78257$.

(b) $E_n(500{,}000) = [\sum_{j=1}^{30} y_j + 5(500{,}000)]/35 = 153{,}139$.
$E_n^{(2)}(500{,}000) = [\sum_{j=1}^{30} y_j^2 + 5(500{,}000)^2]/35 = 53{,}732{,}687{,}032$.

12.9 $\hat{\mu}'_1 = [2(2{,}000) + 6(4{,}000) + 12(6{,}000) + 10(8{,}000)]/30 = 6{,}000$,
$\hat{\mu}_2 = [2(-4{,}000)^2 + 6(-2{,}000)^2 + 12(0)^2 + 10(2{,}000)^2]/30 = 3{,}200{,}000$,
$\hat{\mu}_3 = [2(-4{,}000)^3 + 6(-2{,}000)^3 + 12(0)^3 + 10(2{,}000)^3]/30 = -3{,}200{,}000{,}000$.
$\hat{\gamma}_1 = -3{,}200{,}000{,}000/(3{,}200{,}000)^{1.5} = -0.55902$.

12.10 Suppose the lapse was at time 1. The estimate of $S(4)$ is $(3/4)(2/3)(1/2) = 0.25$. If it is at time 2, the estimate is (4/5)(2/3)(1/2)=0.27. If it is at time 3, the estimate is (4/5)(3/4)(1/2) = 0.3. If it is at time 4, the estimate is (4/5)(3/4)(2/3) = 0.4. If it is at time 5, the estimate is (4/5)(3/4)(2/3)(1/2) = 0.20. Therefore, the answer is time 5.

12.11 $\hat{H}(12) = \frac{2}{15} + \frac{1}{12} + \frac{1}{10} + \frac{2}{6} = 0.65$. The estimate of the survival function is $\hat{S}(12) = e^{-0.65} = 0.522$.

12.12 $\hat{H}(t_{10}) - \hat{H}(t_9) = \frac{1}{n-9} = 0.077$, $n = 22$. $\hat{H}(t_3) = \frac{1}{22} + \frac{1}{21} + \frac{1}{20} = 0.14307$, $\hat{S}(t_3) = e^{-0.14307} = 0.8667$.

12.13 $0.60 = 0.72\frac{r_4-2}{r_4}$, $r_4 = 12$. $0.50 = 0.60\frac{r_5-1}{r_5}$, $r_5 = 6$. With two deaths at the fourth death time and the risk set decreasing by 6, there must have been four censored observations.

12.14 $0.575 = \hat{S}(10) = e^{-\hat{H}(10)}$. $\hat{H}(10) = -\ln(0.575) = 0.5534 = \frac{1}{50} + \frac{3}{49} + \frac{5}{k} + \frac{7}{12}$. The solution is $k = 36$.

12.15 With no censoring, the r values are 12, 9, 8, 7, 6, 4, and 3, and the s values are 3, 1, 1, 1, 2, 1, and 1. Then

$$\hat{H}(7,000) = \frac{3}{12} + \frac{1}{9} + \frac{1}{8} + \frac{1}{7} + \frac{2}{6} + \frac{1}{4} + \frac{1}{3} = 1.5456.$$

With censoring, there are only five uncensored values with r values 9, 8, 7, 4, and 3, and all five s values are 1. Then

$$\hat{H}(7{,}000) = \frac{1}{9} + \frac{1}{8} + \frac{1}{7} + \frac{1}{4} + \frac{1}{3} = 0.9623.$$

12.16 $\frac{0.5}{0.65} = \frac{r_4-s_4}{r_4} = \frac{r_4-3}{r_4} \Longrightarrow r_4 = 13$. The risk set a the fifth death time is the 13 from the previous death time less the 3 who died and less the 6 who were censored. That leaves $r_5 = 13 - 3 - 6 = 4$. Then, $\frac{0.25}{0.5} = \frac{r_5-s_5}{r_5} = \frac{4-s_5}{4} \Longrightarrow s_5 = 2$.

12.17 From the data set, there were 1,915 out of 94,935 that had two or more accidents. The estimated probability is $1,915/94,935 = 0.02017$, and the estimated variance is

$$0.02017(0.97983)/94,935 = 2.08176 \times 10^{-7}.$$

12.18 Without any distributional assumptions, the variance is estimated as $(1/5)(4/5)/5 = 0.032$. From the distributional assumption, the true value of ${}_3q_7$ is $[(8/15)-(5/15)]/(8/15) = 3/8$, and the variance is $(3/8)(5/8)/5 = 0.046875$. The difference is -0.014875.

12.19 The Nelson–Åalen estimates are the centers of the confidence intervals, which are 0.15 and 0.27121. Therefore, $s_{j+1}/r_{j+1} = 0.12121$. From the first confidence interval, the estimated variance is $(0.07875/1.96)^2 = 0.0016143$, while for the second interval it is $(0.11136/1.96)^2 = 0.0032282$ and, therefore, $s_{j+1}(r_{j+1} - s_{j+1})/r_{i+1}^3 = 0.0016139$. Substituting $s_{j+1} = 0.12121r_{j+1}$ produces

$$\frac{0.12121r_{j+1}(0.87879r_{j+1})}{r_{j+1}^3} = 0.0016139.$$

This leads to $0.10652 = 0.0016139r_{j+1}$ for $r_{j+1} = 66$ and then $s_{j+1} = 0.12121(66) = 8$.

12.20 Greenwood's estimate is

$$V = S^2\left(\frac{2}{50(48)} + \frac{4}{45(41)} + \frac{8}{(41-c)(33-c)}\right).$$

Then,

$$\begin{aligned} 0.011467 = \frac{V}{S^2} &= 0.003001 + \frac{8}{(41-c)(33-c)}, \\ (41-c)(33-c) &= 8/(0.008466) = 945. \end{aligned}$$

Solving the quadratic equation yields $c = 6$.

12.21 For the Nelson–Åalen estimate,

$$\begin{aligned} 1.5641 = \hat{H}(35) &= \frac{2}{15} + \frac{3}{13} + \frac{2}{10} + \frac{d}{8} + \frac{2}{8-d}, \\ (1.5641 - 0.5641)8(8-d) &= d(8-d) + 16, \\ 0 &= d^2 - 16d + 48, \\ d &= 4. \end{aligned}$$

Note that the other solution to the quadratic equation ($d = 12$) is not plausible as only 8 individuals were available to become deaths.

The variance is

$$\frac{2(13)}{15^3} + \frac{3(10)}{13^3} + \frac{2(8)}{10^3} + \frac{4(4)}{8^3} + \frac{2(2)}{4^3} = 0.13111.$$

Table 12.28 Solution to Exercise 12.24.

i	y_i	s_i	b_i	r_i
1	4	3	**6**	40
2	6	**5**	3	31
3	9	6	4	23
4	13	4	**3**	**13**
5	15	2	4	6

12.22 The uncensored observations are 4 and 8, the two r values are 10 and 5, and the two s values are 2 and 1. Then $\hat{S}(11) = \frac{8}{10}\frac{4}{5} = 0.64$. Greenwood's estimate is

$$(0.64)^2\left[\frac{2}{10(8)} + \frac{1}{5(4)}\right] = 0.03072.$$

12.23 At day 8 (the first uncensored time), $r = 7\ (8-1)$ and $s = 1$. At day 12, $r = 5$ and $s = 2$. Then $\hat{H}(12) = \frac{1}{7} + \frac{2}{5} = 0.5429$. Also, $\widehat{\text{Var}}[\hat{H}(12)] = \frac{1(6)}{7^3} + \frac{2(3)}{5^3} = 0.06549$. The interval is $0.5429 \pm 1.645\sqrt{0.06549}$, which is $(0.1219, 0.9639)$

12.24 (a) The missing values are added in Table 12.28.

$$S_{40}(4) = 1 - \frac{3}{40} = \frac{37}{40} = 0.925$$
$$S_{40}(6) = S_{40}(4)\left(1 - \frac{5}{31}\right) = \frac{37}{40}\cdot\frac{26}{31} = \frac{481}{620} = 0.77581$$
$$S_{40}(9) = S_{40}(6)\left(1 - \frac{6}{23}\right) = \frac{481}{620}\cdot\frac{17}{23} = \frac{8,177}{14,260} = 0.57342$$
$$S_{40}(13) = S_{40}(9)\left(1 - \frac{4}{13}\right) = \frac{8,177}{14,260}\cdot\frac{9}{13} = \frac{5,661}{14,260} = 0.39698$$
$$S_{40}(15) = S_{40}(13)\left(1 - \frac{2}{6}\right) = \frac{5,661}{14,260}\cdot\frac{2}{3} = \frac{1,887}{7,130} = 0.26466$$

(b)

$$\hat{H}(4) = \frac{3}{40} = 0.075$$
$$\hat{H}(6) = \hat{H}(4) + \frac{5}{31} = 0.23629$$
$$\hat{H}(9) = \hat{H}(6) + \frac{6}{23} = 0.49716$$
$$\hat{H}(13) = \hat{H}(9) + \frac{4}{13} = 0.80485$$
$$\hat{H}(15) = \hat{H}(13) + \frac{2}{6} = 1.13819$$

(c)

$$S_{40}(24) = \{S_{40}(15)\}^{24/15} = \left(\frac{1,887}{7,130}\right)^{1.6} = (0.26466)^{1.6} = 0.11920$$

(d)

$$\sum_{i=1}^{5} \frac{s_i}{r_i\,(r_i - s_i)} = \frac{3}{40(37)} + \frac{5}{31(26)} + \frac{6}{23(1)7} + \frac{4}{13(9)} + \frac{2}{6(4)} = 0.14110$$

$$\widehat{Var}\,\{S_{40}(15)\} = (0.26466)^2(0.14110) = 0.00988$$

(e)

$$0.26466 \pm 1.96\sqrt{0.00988} = (0.06984, 0.45948)$$

(f)

$$U = \exp\left[\frac{1.96\sqrt{0.00988}}{0.26466\ln(0.26466)}\right] = 0.57479$$

$$\left(S_{40}(15)^{1/U}, S_{40}(15)^{U}\right) = (0.09899, 0.46577)$$

12.25 (a)

$$\begin{aligned}
&S_{50}(3) = 1 - \tfrac{3}{50} = 0.94\\
&S_{50}(5) = S_{50}(3)\left(1 - \tfrac{7}{41}\right) = 0.77951\\
&S_{50}(7) = S_{50}(5)\left(1 - \tfrac{5}{30}\right) = 0.64959\\
&S_{50}(11) = S_{50}(7)\left(1 - \tfrac{5}{23}\right) = 0.50838\\
&S_{50}(16) = S_{50}(11)\left(1 - \tfrac{6}{15}\right) = 0.30503\\
&S_{50}(20) = S_{50}(16)\left(1 - \tfrac{2}{5}\right) = 0.18302
\end{aligned}$$

(b)

$$\begin{aligned}
&\hat{H}(3) = \tfrac{3}{50} = 0.06, && \hat{S}(3) = e^{-0.06} = 0.94176\\
&\hat{H}(5) = \hat{H}(3) + \tfrac{7}{41} = 0.23073, && \hat{S}(5) = e^{-0.23073} = 0.79395\\
&\hat{H}(7) = \hat{H}(5) + \tfrac{5}{30} = 0.39740, && \hat{S}(7) = e^{-0.39740} = 0.67207\\
&\hat{H}(11) = \hat{H}(7) + \tfrac{5}{23} = 0.61479, && \hat{S}(11) = e^{-0.61479} = 054075\\
&\hat{H}(16) = \hat{H}(11) + \tfrac{6}{15} = 1.01479, && \hat{S}(16) = e^{-1.01479} = 0.36248\\
&\hat{H}(20) = \hat{H}(16) + \tfrac{2}{5} = 1.41479, && \hat{S}(20) = e^{-1.41479} = 0.24298
\end{aligned}$$

(c)

$$\begin{aligned}
S_{50}(40) &= 0 \text{ EFRON}\\
S_{50}(40) &= [S_{50}(20)]^{40/20} = (0.18302)^2 = 0.03350 \text{ BHK}
\end{aligned}$$

(d)

$$\begin{aligned}\widehat{Var}\left[\hat{H}(20)\right] &= \sum_{i=1}^{6}\frac{s_i(r_i-s_i)}{r_i^3}\\ &= \frac{3(47)}{50^3}+\frac{7(34)}{41^3}+\frac{5(25)}{30^3}+\frac{5(18)}{23^3}+\frac{6(9)}{15^3}+\frac{2(3)}{5^3}\\ &= 0.08061\\ \text{Var}\left[\hat{S}(20)\right] &\approx (0.24298)^2(0.08061) = 0.00476 = (0.06898)^2\end{aligned}$$

(e) $U = \exp\left[\pm 1.96\sqrt{0.08061}/1.41479\right] = (0.67480, 1.48190) \Rightarrow$ 95% confidence interval is from (1.41479)(0.67480) = 0.95471 to (1.41479)(1.48190) = 2.0966.

(f) $\text{Var}\left\{\hat{S}(40)\right\} = \text{Var}\left\{\left[\hat{S}(20)\right]^2\right\}$ under B, H, & K.

Let $g(x) = x^2 \Rightarrow g'(x) = 2x \Rightarrow$

$$\begin{aligned}\text{Var}\left\{\hat{S}(40)\right\} &= \text{Var}\left\{\left[\hat{S}(20)\right]^2\right\} \approx \left\{2\hat{S}(20)\right\}^2 \text{Var}\left\{\hat{S}(20)\right\}\\ &\approx 4\left\{\hat{S}(20)\right\}^2 \text{Var}\left\{\hat{S}(20)\right\}\end{aligned}$$

which is

$$\begin{aligned}&= 4(0.24298)^2(0.00476)\\ &= (0.03352)^2\\ &= 0.00112.\end{aligned}$$

12.26 (a) For the Kaplan–Meier estimator, $S_n(y) = \prod_{i:y_i\le y}\left(1-\hat{\lambda}_i\right)$ and for the Nelson–Åalen estimator, $\hat{S}(y) = \prod_{i:y_i\le y} e^{-\hat{\lambda}_i}$.

(b) Note that $\ln\tilde{S}(y) = \sum_{i|y_i\le y}\ln\phi\left(\hat{\lambda}_i\right)$, and by the delta method with $f(x) = \ln\phi(x)$,

$$\text{Var}\left[\ln\phi\left(\hat{\lambda}_i\right)\right] \approx \left[\frac{\phi'\left(\hat{\lambda}_i\right)}{\phi\left(\hat{\lambda}_i\right)}\right]^2 \text{Var}\left(\hat{\lambda}_i\right) \approx \left[\frac{\phi'\left(\hat{\lambda}_i\right)}{\phi\left(\hat{\lambda}_i\right)}\right]^2 \frac{s_i(r_i-s_i)}{r_i^3}.$$

Because the $\hat{\lambda}_i$ are approximately uncorrelated,

$$\text{Var}\left[\ln\tilde{S}(y)\right] \approx \sum_{i|y_i\le y}\text{Var}\left[\ln\phi\left(\hat{\lambda}_i\right)\right] \approx \sum_{i|y_i\le y}\left[\frac{\phi'\left(\hat{\lambda}_i\right)}{\phi\left(\hat{\lambda}_i\right)}\right]^2 \frac{s_i(r_i-s_i)}{r_i^3}.$$

By the delta method again with $f(x) = e^x$,

$$\text{Var}\left[\tilde{S}(y)\right] \approx \left[\tilde{S}(y)\right]^2 \text{Var}\left[\ln\tilde{S}(y)\right],$$

from which the result follows.
(c) For $m = 0$, $(-1)^0 [\phi_0(x) - e^{-x}] = 1 - e^{-x} \geq 0$. Next assume that $(-1)^m [\phi_m(x) - e^{-x}] \geq 0$ for m. Then let $f_{m+1}(x) = (-1)^{m+1} [\phi_{m+1}(x) - e^{-x}]$ and so $f_{m+1}(0) = 0$. Also,

$$\begin{aligned} f'_{m+1}(x) &= (-1)^{m+1} \left[\sum_{j=1}^{m+1} \frac{(-1)^j x^{j-1}}{(j-1)!} + e^{-x} \right] \\ &= (-1)^{m+1} \left[\sum_{j=0}^{m} \frac{(-1)^{j+1} x^j}{j!} + e^{-x} \right] \\ &= (-1)^m \left[\sum_{j=0}^{m} \frac{(-x)^j}{j!} - e^{-x} \right] \\ &= (-1)^m \left[\phi_m(x) - e^{-x} \right], \end{aligned}$$

and so $f'_{m+1}(x) \geq 0$ by the inductive hypothesis. Because $f_{m+1}(0) = 0$ and $f'_{m+1}(x) \geq 0$, it follows that $f_{m+1}(x) \geq 0$, implying by induction on m that $(-1)^m [\phi_m(x) - e^{-x}] \geq 0$ for $m = 0, 1, 2, \ldots$. This in turn implies that

$$e^{-x} \leq \phi_{2m}(x) = \sum_{j=0}^{2m} \frac{(-x)^j}{j!}, \quad m = 0, 1, 2, \ldots,$$

and

$$e^{-x} \geq \phi_{2m+1}(x) = \sum_{j=0}^{2m+1} \frac{(-x)^j}{j!}, \quad m = 0, 1, 2, \ldots,$$

and the result follows by also noting that $\tilde{S}_1(y) = S_n(y)$, the Kaplan–Meier estimate.

12.27 We have $k = 5$ and $y_k = y_{max} = 15$ in this case. The estimates of the survival function are given in the solutions to Exercise 12.24.
(a)

$$\begin{aligned} \tilde{\tau}(y_4) &= (15 - 13) S_{40}(13) = 2(0.39698) = 0.79397 \\ \tilde{\tau}(y_3) &= \tilde{\tau}(y_4) + (13 - 9) S_{40}(9) = 0.79397 + 4(0.57342) = 3.08766 \\ \tilde{\tau}(y_2) &= \tilde{\tau}(y_3) + (9 - 6) S_{40}(6) = 3.08766 + 3(0.77581) = 5.41508 \\ \tilde{\tau}(y_1) &= \tilde{\tau}(y_2) + (6 - 4) S_{40}(4) = 5.41508 + 2(0.925) = 7.26508 \\ \tilde{\mu}_1 &= 4(1) + \tilde{\tau}(y_1) = 4 + 7.26508 = 11.26508 \end{aligned}$$

(b)

$$\begin{aligned} e^{-\hat{\beta}(15)} &= S_{40}(15) = 0.26466 \Rightarrow \tilde{\beta} = -\ln 0.26466/15 = 0.08862 \\ \tilde{\tau}(y_5) &= (15)(0.26466)/(-\ln 0.26466) = 2.98637 \\ \tilde{\mu}_1 &= 11.26508 + 2.98637 = 14.25144 \end{aligned}$$

(c)

$$\begin{aligned}\sum_{m=1}^{4}\left[\tilde{\tau}(y_m)\right]^2\frac{s_m}{r_m\left(r_m-s_m\right)} &= (7.26508)^2\frac{3}{(40)(37)}+(5.41508)^2\frac{5}{(31)(26)}\\ &\quad+(3.08766)^2\frac{6}{(23)(17)}+(.79397)^2\frac{4}{(13)(9)}\\ &= 0.45674\end{aligned}$$

12.28 We have $k=6$ and $y_k=y_{max}=20$ in this case. The estimates of $\hat{S}(y)$ are given in the solution to Exercise 12.25.
(a)

$$\begin{aligned}e^{-20\tilde{\beta}} &= \tilde{S}(20)=0.24298\Rightarrow\tilde{\beta}=-\ln 0.24298/20=0.07074\\ \tilde{\tau}(y_6) &= (20)(0.24298)/(-\ln 0.24298)=3.43481\\ \tilde{\tau}(y_5) &= (20-16)\tilde{S}(16)+\tilde{\tau}(y_6)=4(0.36248)+3.43481=4.88473\\ \tilde{\tau}(y_4) &= (16-11)\tilde{S}(11)+\tilde{\tau}(y_5)=5(0.54075)+4.88473=7.58848\\ \tilde{\tau}(y_3) &= (11-7)\tilde{S}(7)+\tilde{\tau}(y_4)=4(0.67207)+7.58848=10.27676\\ \tilde{\tau}(y_2) &= (7-5)\tilde{S}(5)+\tilde{\tau}(y_3)=2(0.79395)+10.27676=11.86466\\ \tilde{\tau}(y_1) &= (5-3)\tilde{S}(3)+\tilde{\tau}(y_2)=2(0.94176)+11.86466=13.74818\\ \tilde{\mu} &= y_1+\tilde{\tau}(y_1)=3+13.74818=16.74818\end{aligned}$$

(b)

$$\begin{aligned}\sum_{m=1}^{6}\left\{\tilde{\tau}(y_m)+\frac{\left[\tilde{\tau}(y_6)\right]^2}{y_6\tilde{S}(y_6)}\right\}^2\frac{s_m(r_m-s_m)}{r_m^3} &= \sum_{m=1}^{6}\left[\tilde{\tau}(y_m)+2.42779\right]^2\frac{s_m(r_m-s_m)}{r_m^3}\\ &= (13.74818+2.42779)^2\frac{(3)(47)}{(50)^3}\\ &\quad+(11.86466+2.42779)^2\frac{(7)(34)}{(41)^3}\\ &\quad+(10.27676+2.42779)^2\frac{(5)(25)}{(30)^3}\\ &\quad+(7.58848+2.42779)^2\frac{(5)(18)}{(23)^3}\\ &\quad+(4.88473+2.42779)^2\frac{(6)(9)}{(15)^3}\\ &\quad+(3.43481+2.42779)^2\frac{(2)(3)}{(5)^3}\\ &= 4.99525\end{aligned}$$

12.29 The estimate of the mean is 10.43. The second moment is estimated by

$$\begin{aligned}
\tilde{\mu}_2 &= y_1^2 + \left[\sum_{j=2}^{7} \left(y_j^2 - y_{j-1}^2\right) \tilde{S}(y_{j-1})\right] + \left(y_{max}^2 - y_7^2\right) \tilde{S}(y_7) \\
&\quad + 2(y_{max})^2 \tilde{S}(y_k) \left\{\left[-\ln \tilde{S}(y_k)\right]^2 + \left[-\ln \tilde{S}(y_k)\right]^{-1}\right\} \\
&= 1^2 + (2^2 - 1^2)(0.95123) + (4^2 - 2^2)(0.90246) + (5^2 - 4^2)(0.80230) \\
&\quad + (8^2 - 5^2)(0.74289) + (9^2 - 8^2)(0.56557) + (12^2 - 9^2)(0.34303) \\
&\quad + (15^2 - 12^2)(0.176) + 2(15)^2(0.12619)\left[(-\ln 0.12619)^{-2} + (-\ln 0.12619)^{-1}\right] \\
&= 1 + 3(0.95123) + 12(0.90246) + 9(0.80230) + 39(0.74289) \\
&\quad + 17(0.56557) + 63(0.34303) + 81(0.176) + 40.68595 \\
&= 137.04415
\end{aligned}$$

The variance is thus estimated by $137.04415 - (10.43)^2 = 28.25925$.

12.30 The calculations are in Tables 12.29 and 12.30.

Table 12.29 Calculations for Exercise 12.30.

i	d_i	u_i	x_i	i	d_i	u_i	x_i
1	0	–	0.1	16	0	4.8	–
2	0	–	0.5	17	0	–	4.8
3	0	–	0.8	18	0	–	4.8
4	0	0.8	–	19–30	0	–	5.0
5	0	–	1.8	31	0.3	–	5.0
6	0	–	1.8	32	0.7	–	5.0
7	0	–	2.1	33	1.0	4.1	–
8	0	–	2.5	34	1.8	3.1	–
9	0	–	2.8	35	2.1	–	3.9
10	0	2.9	–	36	2.9	–	5.0
11	0	2.9	–	37	2.9	–	4.8
12	0	–	3.9	38	3.2	4.0	–
13	0	4.0	–	39	3.4	–	5.0
14	0	–	4.0	40	3.9	–	5.0
15	0	–	4.1				

12.31

Table 12.30 Further calculations for Exercise 12.30.

j	y_j	s_j	r_j
1	0.1	1	$30-0-0=30$ or $0+30-0-0=30$
2	0.5	1	$31-1-0=30$ or $30+1-1-0=30$
3	0.8	1	$32-2-0=30$ or $30+1-1-0=30$
4	1.8	2	$33-3-1=29$ or $30+1-1-1=29$
5	2.1	1	$34-5-1=28$ or $29+1-2-0=28$
6	2.5	1	$35-6-1=28$ or $28+1-1-0=28$
7	2.8	1	$35-7-1=27$ or $28+0-1-0=27$
8	3.9	2	$39-8-4=27$ or $27+4-1-3=27$
9	4.0	1	$40-10-4=26$ or $27+1-2-0=26$
10	4.1	1	$40-11-6=23$ or $26+0-1-2=23$
11	4.8	3	$40-12-7=21$ or $23+0-1-1=21$
12	5.0	17	$40-15-8=17$ or $21+0-3-1=17$

$$S_{40}(t)=\begin{cases}1, & 0\le t<0.1,\\ \frac{30-1}{30}=0.9667, & 0.1\le t<0.5,\\ 0.9667\frac{30-1}{30}=0.9344, & 0.5\le t<0.8,\\ 0.9344\frac{30-1}{30}=0.9033, & 0.8\le t<1.8,\\ 0.9033\frac{29-2}{29}=0.8410, & 1.8\le t<2.1,\\ 0.8410\frac{28-1}{28}=0.8110, & 2.1\le t<2.5,\\ 0.8110\frac{28-1}{28}=0.7820, & 2.5\le t<2.8,\\ 0.7820\frac{27-1}{27}=0.7530, & 2.8\le t<3.9,\\ 0.7530\frac{27-2}{27}=0.6973, & 3.9\le t<4.0,\\ 0.6973\frac{26-1}{26}=0.6704, & 4.0\le t<4.1,\\ 0.6704\frac{23-1}{23}=0.6413, & 4.1\le t<4.8,\\ 0.6413\frac{21-3}{21}=0.5497, & 4.8\le t<5.0,\\ 0.5497\frac{17-17}{17}=0, & t\ge 5.0.\end{cases}$$

$$\hat{H}(t) = \begin{cases} 0, & 0 \le t < 0.1, \\ \frac{1}{30} = 0.0333, & 0.1 \le t < 0.5, \\ 0.0333 + \frac{1}{30} = 0.0667, & 0.5 \le t < 0.8, \\ 0.0667 + \frac{1}{30} = 0.1000, & 0.8 \le t < 1.8, \\ 0.1000 + \frac{2}{29} = 0.1690, & 1.8 \le t < 2.1, \\ 0.1690 + \frac{1}{28} = 0.2047, & 2.1 \le t < 2.5, \\ 0.2047 + \frac{1}{28} = 0.2404, & 2.5 \le t < 2.8, \\ 0.2404 + \frac{1}{27} = 0.2774, & 2.8 \le t < 3.9, \\ 0.2774 + \frac{2}{27} = 0.3515, & 3.9 \le t < 4.0, \\ 0.3515 + \frac{1}{26} = 0.3900, & 4.0 \le t < 4.1, \\ 0.3900 + \frac{1}{23} = 0.4334, & 4.1 \le t < 4.8, \\ 0.4334 + \frac{3}{21} = 0.5763, & 4.8 \le t < 5.0, \\ 0.5763 + \frac{17}{17} = 1.5763, & t \ge 5.0. \end{cases}$$

$$\hat{S}(t) = \begin{cases} e^{-0} = 1, & 0 \le t < 0.1, \\ e^{-0.0333} = 0.9672, & 0.1 \le t < 0.5, \\ e^{-0.0667} = 0.9355, & 0.5 \le t < 0.8, \\ e^{-0.1000} = 0.9048, & 0.8 \le t < 1.8, \\ e^{-0.1690} = 0.8445, & 1.8 \le t < 2.1, \\ e^{-0.2047} = 0.8149, & 2.1 \le t < 2.5, \\ e^{-0.2404} = 0.7863, & 2.5 \le t < 2.8, \\ e^{-0.2774} = 0.7578, & 2.8 \le t < 3.9, \\ e^{-0.3515} = 0.7036, & 3.9 \le t < 4.0, \\ e^{-0.3900} = 0.6771, & 4.0 \le t < 4.1, \\ e^{-0.4334} = 0.6483, & 4.1 \le t < 4.8, \\ e^{-0.5763} = 0.5620, & 4.8 \le t < 5.0, \\ e^{-1.5763} = 0.2076, & t \ge 5.0. \end{cases}$$

12.33 Using the raw data, the results are in Table 12.31. When the deductible and limit are imposed, the results are as in Table 12.32. Because 1,000 is a censoring point and not an observed loss value, there is no change in the survival function at 1,000.

Table 12.31 Calculations for Exercise 12.33.

value(x)	r	s	$S_{KM}(x)$	$H_{NA}(x)$	$S_{NA}(x)$
27	20	1	0.95	0.0500	0.9512
82	19	1	0.90	0.1026	0.9025
115	18	1	0.85	0.1582	0.8537
126	17	1	0.80	0.2170	0.8049
155	16	1	0.75	0.2795	0.7562
161	15	1	0.70	0.3462	0.7074
243	14	1	0.65	0.4176	0.6586
294	13	1	0.60	0.4945	0.6099
340	12	1	0.55	0.5779	0.5611
384	11	1	0.50	0.6688	0.5123
457	10	1	0.45	0.7688	0.4636
680	9	1	0.40	0.8799	0.4148
855	8	1	0.35	1.0049	0.3661
877	7	1	0.30	1.1477	0.3174
974	6	1	0.25	1.3144	0.2686
1,193	5	1	0.20	1.5144	0.2199
1,340	4	1	0.15	1.7644	0.1713
1,884	3	1	0.10	2.0977	0.1227
2,558	2	1	0.05	2.5977	0.0744
15,743	1	1	0.00	3.5977	0.0274

12.34 The information may be organized as in Table 12.33

12.35 From Exercise 12.32, $\hat{H}(3) = 0.2774$. The variance is estimated as

$$\widehat{\text{Var}}[\hat{H}(3)] = \frac{1(29)}{30^3} + \frac{1(29)}{30^3} + \frac{1(29)}{30^3} + \frac{2(27)}{29^3} + \frac{1(27)}{28^3} + \frac{1(27)}{28^3} + \frac{1(26)}{27^3} = 0.0092172.$$

The linear confidence interval is

$$0.2774 \pm 1.96\sqrt{0.0092172} \text{ or } 0.0892 \text{ to } 0.4656.$$

The log-transformed interval requires

$$U = \exp\left[\pm\frac{1.96\sqrt{0.0092172}}{0.2774}\right] = 1.97060, \text{ or } 0.50746.$$

The lower limit is $0.2774(0.50746) = 0.14077$
and the upper limit is $0.2774(1.97060) = 0.54664$.

12.36 First, obtain the estimated survival probability as

$$\hat{S}(4) = \frac{12}{15}\frac{56}{80}\frac{20}{25}\frac{54}{60} = 0.4032.$$

Table 12.32 Further calculations for Exercise 12.33.

value(x)	r	s	$S_{KM}(x)$	$H_{NA}(x)$	$S_{NA}(x)$
115	18	1	0.9444	0.0556	0.9459
126	17	1	0.8889	0.1144	0.8919
155	16	1	0.8333	0.1769	0.8379
161	15	1	0.7778	0.2435	0.7839
243	14	1	0.7222	0.3150	0.7298
294	13	1	0.6667	0.3919	0.6758
340	12	1	0.6111	0.4752	0.6218
384	11	1	0.5556	0.5661	0.5677
457	10	1	0.5000	0.6661	0.5137
680	9	1	0.4444	0.7773	0.4596
855	8	1	0.3889	0.9023	0.4056
877	7	1	0.3333	1.0451	0.3517
974	6	1	0.2778	1.2118	0.2977

Table 12.33 Calculations for Exercise 12.34.

age(t)	#ds	#xs	#us	r	$\hat{S}(t)$
0	300				
1		6		300	$\frac{294}{300} = 0.98$
2	20				
3		10		314	$0.98\frac{304}{314} = 0.94879$
4	30	10		304	$0.94879\frac{294}{304} = 0.91758$
5		a		324	$0.91758\frac{324-a}{324} = 0.892$
7			45		$\Longrightarrow a = 9$
9		b		$279-a = 270$	$0.892\frac{270-b}{270}$
10			35		
12		6		$244-a-b = 235-b$	$0.892\frac{270-b}{270}\frac{229-b}{235-b} = 0.856$
13			15		$\Longrightarrow b = 4$

Greenwood's formula gives

$$(0.4032)^2\left(\frac{3}{15(12)}+\frac{24}{80(56)}+\frac{5}{25(20)}+\frac{6}{60(54)}\right)=0.00551.$$

12.37 The standard deviation is

$$\left(\frac{15(85)}{100^3}+\frac{20(45)}{65^3}+\frac{13(27)}{40^3}\right)^{1/2}=0.10018.$$

Table 12.34 Data for Exercise 12.39.

y_j	$p(y_j)$
0.1	0.0333
0.5	0.0323
0.8	0.0311
1.8	0.0623
2.1	0.0300
2.5	0.0290
2.8	0.0290
3.9	0.0557
4.0	0.0269
4.1	0.0291
4.8	0.0916

12.38 In order for the mean to be equal to y, we must have $\theta/(\alpha-1) = y$. Letting α be arbitrary (and greater than 1), use a Pareto distribution with $\theta = y(\alpha-1)$. This makes the kernel function

$$k_y(x) = \frac{\alpha[(\alpha-1)y]^\alpha}{[(\alpha-1)y+x]^{\alpha+1}}.$$

12.39 The data points and probabilities can be taken from Exercise 12.31. They are given in Table 12.34

The probability at 5.0 is discrete and so should not be spread out by the kernel density estimator. Because of the value at 0.1, the largest available bandwidth is 0.1. Using this bandwidth and the triangular kernel produces the graph in Figure 12.13.

This graph is clearly not satisfactory. The gamma kernel is not available because there would be positive probability at values greater than 5. Your author tried to solve this by using the beta distribution. With θ known to be 5, the mean (to be set equal to y) is $5a/(a+b)$. To have some smoothing control and to determine parameter values, the sum $a+b$ was fixed. Using a value of 50 for the sum, the kernel is

$$k_y(x) = \frac{\Gamma(50)}{\Gamma(10y)\Gamma(50-10y)}\left(\frac{x}{5}\right)^{10y}\left(1-\frac{x}{5}\right)^{50-10y-1}\frac{1}{x},\ 0<x<5$$

and the resulting smoothed estimate appears in Figure 12.14.

12.40 With a bandwidth of 60, the height of the kernel is 1/120. At a value of 100, the following data points contribute probability 1/20: 47, 75, and 156. Therefore, the height is 3(1/20)(1/120) = 1/800.

12.41 The uniform kernel spreads the probability of 0.1 to 10 units either side of an observation. The observation at 25 contributes a density of 0.005 from 15 to 35, thus contributing nothing to survival past age 40. The same applies to the point at 30. The points at 35 each contribute probability from 25 to 45 and 0.25 of that probability is above 40. Together they

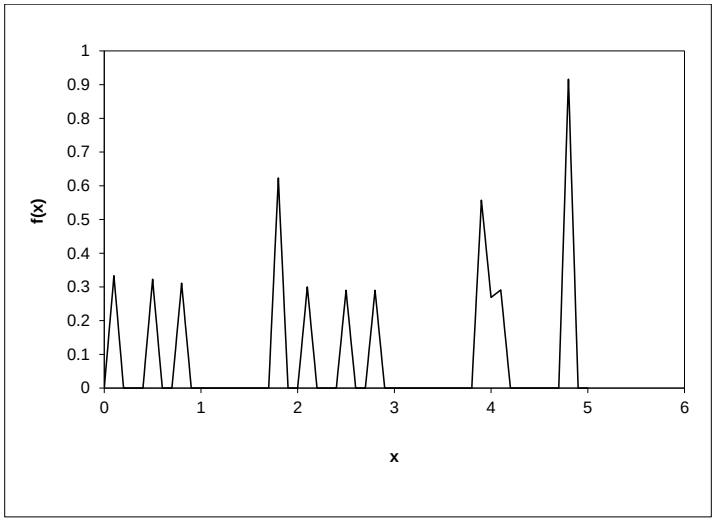

Figure 12.13 Triangular kernel for Exercise 12.39.

contribute $2(0.25)(0.1) = 0.05$. The point at 37 contributes $(7/20)(0.1) = 0.035$. The next four points contribute a total of $(9/20 + 15/20 + 17/20 + 19/20)(0.1) = 0.3$. The final point (at 55) contributes all its probability at points above 40 and so contributes 0.1 to the total, which is $0.05 + 0.035 + 0.3 + 0.1 = 0.485$.

12.42 (a) With two of the five points below 150, the empirical estimate is $2/5 = 0.4$.

(b) The point at 82 has probability 0.2 spread from 32 to 132; all is below 150, so the contribution is 0.2. The point at 126 has probability from 76 to 176; the contribution below 150 is $(74/100)(0.2) = 0.148$. The point at 161 contributes $(39/100)(0.2) = 0.078$. The last two points contribute nothing, so the estimate is $0.2 + 0.148 + 0.078 = 0.426$.

(c) As in part (b), the first point contributes 0.2 and the last two points contribute nothing. The triangle has base 100 and area 0.2, so the height must be 0.004. For the point at 126, the probability excluded is the triangle from 150 to 176, which has base 26 and height at 150 of $0.004(26/50) = 0.00208$. The area is $0.5(26)(0.00208) = 0.02704$ and the contribution is 0.17296. For the point at 161, the area included is the triangle from 111 to 150, which has base 39 and height at 150 of $0.004(39/50) = 0.00312$. The area is $0.5(39)(0.00312) = 0.06084$. The estimate is $0.2 + 0.17296 + 0.06084 = 0.4338$.

12.43 The only change is the entries in the d_j and w_j^m columns are swapped. The exposures then change. For example, at $j = 2$, the exposure is $28 + 0 + 3(1/2) - 2(1/2) = 28.5$.

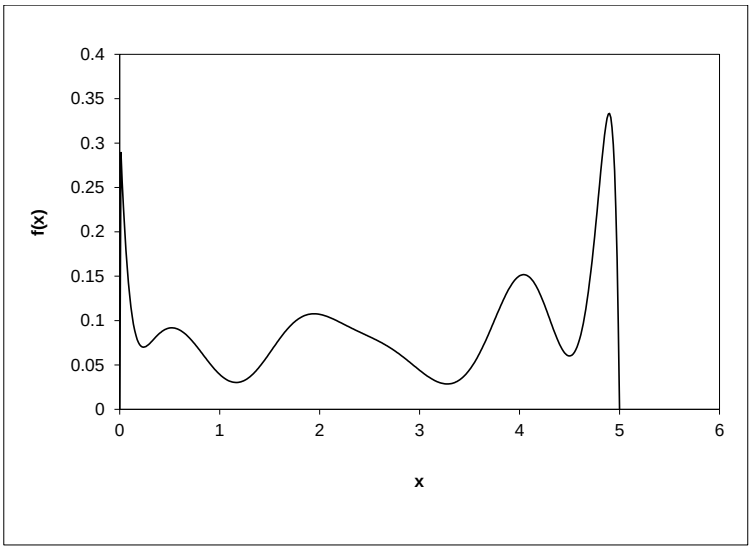

Figure 12.14 Beta kernel for Exercise 12.39.

Table 12.35 Kaplan–Meier calculations for Exercise 12.44.

x	r	s	$\hat{S}(x)$
45.3	8	1	$7/8 = 0.875$
45.4	7	1	$0.875(6/7) = 0.750$
46.2	8	1	$0.750(7/8) = 0.656$
46.4	6	1	$0.656(5/6) = 0.549$
46.7	5	1	$0.549(4/5) = 0.438$

12.44 The Kaplan–Meier calculations appear in Table 12.35. Then $\hat{q}_{45} = 1 - 0.750 = 0.250$ and $\hat{q}_{46} = 1 - 0.438/0.750 = 0.416$.

For exact exposure at age 45, the first eight individuals contribute $1 + 1 + 0.3 + 1 + 0.4 + 1 + 0.4 + 0.8 = 5.9$ of exposure for $\hat{q}_{45} = 1 - e^{-2/5.9} = 0.28751$. For age 46, eight individuals contribute for an exposure of $0.7 + 1 + 1 + 1 + 0.3 + 0.2 + 0.4 + 0.9 = 5.5$ for $\hat{q}_{46} = 1 - e^{-3/5.5} = 0.42042$.

For actuarial exposure at age 45 the two deaths add 1.3 to the exposure for $\hat{q}_{45} = 2/7.2 = 0.27778$. For age 46, the three deaths add 1.7 for $\hat{q}_{46} = 3/7.2 = 0.41667$.

12.45 The calculations are in Table 12.36. The intervals were selected so that all deductibles and limits are at the boundaries and so no assumption is needed about the intermediate values. This setup means that no assumptions need to be made about the timing of entrants and withdrawals. When working with the observations, the deductible must be

Table 12.36 Calculations for Exercise 12.45.

c_j	P_j	n_j^b	w_j^e	d_j	$q_j'^{(d)}$	$\hat{S}(c_j)$
250	0	7	0	1	1/7	1.000
500	6	8	0	2	2/14	$1.000(6/7) = 0.857$
1,000	12	7	1	4	4/19	$0.857(12/14) = 0.735$
2,750	14	0	1	1	1/14	$0.735(15/19) = 0.580$
3,000	12	0	0	0	0/12	$0.580(13/14) = 0.539$
3,500	12	0	1	6	6/12	$0.539(12/12) = 0.539$
5,250	5	0	1	0	0/5	$0.539(6/12) = 0.269$
5,500	4	0	1	1	1/4	$0.269(5/5) = 0.269$
6,000	2	0	0	1	1/2	$0.269(3/4) = 0.202$
10,250	1	0	1	0	0/1	$0.202(1/2) = 0.101$
10,500	0	0	0	0	–	$0.101(1/1) = 0.101$

added to the payment in order to produce the loss amount. The requested probability is $\hat{S}(5{,}500)/\hat{S}(500) = 0.269/0.857 = 0.314$.

12.46 For the intervals, the n-values are 6, 6, 7, 0, 0, 0, and 0, the w-values are 0, 0, 1, 1, 1, 0, and 0, and the d-values are 1, 2, 4, 7, 1, 1, and 0. the P values are then 6, $6-0-1+6=11$, $11-0-2+7=16$, $16-1-4+0=11$, $11-1-7+0=3$, $3-1-1+0=1$, and $1-0-1+0=0$. The estimates are $\hat{S}(500) = 5/6$ and $\hat{S}(6{,}000) = (5/6)(9/11)(12/16)(4/11)(2/3) = 15/121$. The answer is the ratio $(15/121)/(5/6) = 18/121 = 0.14876$.

12.47 For the Kaplan-Meier estimates, the variances are (by Greenwood's formula)

$$\hat{Var}(\hat{q}_{45}) = (0.750)^2\left[\frac{1}{8(7)} + \frac{1}{7(6)}\right] = 0.02344$$
$$\hat{Var}(\hat{q}_{46}) = (0.584)^2\left[\frac{1}{8(7)} + \frac{1}{6(5)} + \frac{1}{5(4)}\right] = 0.03451.$$

For exact exposure, the deaths were 2 and 3 and the exposures 5.9 and 5.5 for the two ages. The estimated variances are

$$\hat{Var}(\hat{q}_{45}) = (1-0.28751)^2\frac{2}{5.9^2} = 0.02917$$
$$\hat{Var}(\hat{q}_{46}) = (1-0.42042)^2\frac{3}{5.5^2} = 0.03331.$$

The actuarial exposures are 7.2 for both ages. The estimated variances are

$$\hat{Var}(\hat{q}_{45}) = \frac{0.28751(0.71249)}{7.2} = 0.02845$$
$$\hat{Var}(\hat{q}_{46}) = \frac{0.42042(0.57958)}{7.2} = 0.03384.$$

REFERENCES

1. Åalen, O. (1978), "Nonparametric Inference for a Family of Counting Processes," *Annals of Statistics*, **6**, 701–726.
2. Brown, J., Hollander, M. and Korwar, R. (1974), "Nonparametric Tests of Independence for Censored Data, with Applications to Heart Transplant Studies," in *Reliability and Biometry: Statistical Analysis of Lifelength*, (Proschen, F. and Serfling, R., eds.), SIAM, 327–354.
3. Carriere, J. (1993), "Nonparametric Estimators of a Distribution Function Based on Mixtures of Gamma Distributions," *Actuarial Research Clearing House*, **1993.3**, 1–11.
4. Dickson, D., Hardy, M., and Waters, H. (2013), *Actuarial Mathematics for Life Contingent Risks*, 2nd ed., Cambridge: Cambridge University Press.
5. Dropkin, L. (1959), "Some Considerations on Automobile Rating Systems Utilizing Individual Driving Records," *Proceedings of the Casualty Actuarial Society*, **XLVI**, 165–176.
6. Efron, B. (1967), "The Two Sample Problem with Censored Data," in *Proceedings of the Fifth Berkely Symposium on Mathematical Statistics and Probability*, **4**, 831–853.
7. Herzog, T. and Laverty, J. (1995), "Experience of Refinanced FHA Section 203(b) Single Family Mortages," *Actuarial Research Clearing House*, **1995.1**, 97–129.
8. Hogg, R. and Klugman, S. (1984), *Loss Distributions*, New York: Wiley.
9. Hogg, R., McKean, J., and Craig, A. (2005), *Introduction to Mathematical Statistics*, 6th ed., Upper Saddle River, NJ: Prentice-Hall.
10. Kaplan, E. and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, **53**, 457–481.
11. Klein, J. (1991), "Small-sample Moments of Some Estimators of the Variance of the Kaplan–Meier and Nelson–Åalen Estimators," *Scandinavian Journal of Statistics*, **18**, 333–340.
12. Klein, J. and Moeschberger, M. (2003), *Survival Analysis, Techniques for Censored and Truncated Data*, 2nd ed., New York: Springer-Verlag.

13. Klugman, S., Panjer, H., and Willmot, G. (2013), *Loss Models: Further Topics*, New York: Wiley.

14. Lawless, J. (2003), *Statistical Models and Methods for Lifetime Data*, 2nd ed., New York: Wiley.

15. London, D. (1988), *Survival Models and Their Estimation*, 3rd ed., Winsted, CT: ACTEX.

16. Rao, C. (1965), *Linear Statistical Inference and Its Applications*, New York: Wiley.

17. Rohatgi, V. (1976), *An Introduction to Probability Theory and Mathematical Statistics*, New York: Wiley.

18. Society of Actuaries Committee on Actuarial Principles (1992), "Principles of Actuarial Science," *Transactions of the Society of Actuaries*, **XLIV**, 565–628.

19. Society of Actuaries Committee on Actuarial Principles (1995), "Principles Regarding Provisions for Life Risks," *Transactions of the Society of Actuaries*, **XLVII**, 775–793.

20. Waters. H. (1984). "An Approach to the Study of Multiple State Models," *Journal of the Institute of Actuaries*, **111(2)**, 363–374.