when the population has a heavy right tail. For example, when estimating parameters for the normal distribution, the sample mean and variance are sufficient.[3] 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.[4] Finally, these methods allow the analyst to make arbitrary decisions regarding the moments or percentiles to use.

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. For example, we could estimate parameters by minimizing the maximum difference between the distribution function for the parametric model and the distribution function for the Nelson–Åalen estimate. Of the many possibilities, the only one used here is the maximum likelihood estimator. The general form of this estimator is presented in this introduction. This is followed with useful special cases.

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 with grouped data or when there is censoring. For example, when there is censoring at u and a censored observation is observed, the observed event is the interval from u to infinity. 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, $\boldsymbol{\theta}$. In addition, the random variables are assumed to be independent.

Definition 12.6 *The likelihood function is*

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

and the ***maximum likelihood estimate*** *of $\boldsymbol{\theta}$ is the vector that maximizes the likelihood function.*[5]

[3] 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.

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

[5] Some authors write the likelihood function as $L(\boldsymbol{\theta}|\mathbf{x})$, 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.

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 F, will usually be needed.

Because the observations are assumed to be independent, the product in the definition represents the joint probability $\Pr(X_1 \in A_1, \ldots, X_n \in A_n|\boldsymbol{\theta})$, 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 your problem.

Example 12.7 *Suppose the data in Data Set B were censored at 250. Determine the maximum likelihood estimate of θ for an exponential distribution.*

The first seven data points are uncensored. For them, the set A_j contains the single point equal to the observation x_j. When calculating the likelihood function for a single point for a continuous model, it is necessary to interpret $\Pr(X_j = x_j) = f(x_j)$. That is, the density function should be used. 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 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. □

12.2.2 Complete, individual data

When there is no truncation, and no censoring and 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.

Example 12.8 *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.

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 equation 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. □

12.2.3 Complete, grouped data

When data are complete and grouped, 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 12.9 *From Data Set C, 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. □

12.2.4 Truncated or censored data

When data are censored, there is no additional complication. As noted in Example 12.7, right censoring simply creates an interval running from the censoring point to infinity. In that example, data below the censoring point were individual data, and so the likelihood function contains both density and distribution function terms.

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 12.10 *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 5.13, 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 cost 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 observation. 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, let d be the truncation point (use zero if there is no truncation). The denominator is then $1 - F(d)$.

Example 12.11 *Determine Pareto and gamma models for the time to death for Data Set D*2.

Table 12.3 shows how the likelihood function is constructed for these values. For 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

Table 12.3 Likelihood function for Example 12.11

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	$1 - F(4.1)$				

larger as α and θ get larger.[6] For the gamma distribution the maximum is at $\hat{\alpha} = 2.617$ and $\hat{\theta} = 3.311$. □

Discrete data present no additional problems.

Example 12.12 *For Data Set A, 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

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

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

[6] 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.

Poisson distribution

$$\begin{aligned}
L(\lambda) &= \prod_{x=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^x}{x!}\right)^{n_x} = \prod_{x=0}^{\infty} \frac{e^{-n_x\lambda}\lambda^{xn_x}}{(x!)^{n_x}}, \\
l(\lambda) &= \sum_{x=0}^{\infty}(-n_x\lambda + xn_x \ln\lambda - n_x \ln x!) = -n\lambda + n\bar{x}\ln\lambda - \sum_{x=0}^{\infty} n_x \ln x!, \\
l'(\lambda) &= -n + \frac{n\bar{x}}{\lambda} = 0, \\
\hat{\lambda} &= \bar{x}.
\end{aligned}$$

For the binomial distribution

$$\begin{aligned}
L(q) &= \prod_{x=0}^{m} \left[\binom{m}{x} q^x (1-q)^{m-x}\right]^{n_x} = \prod_{x=0}^{m} \frac{m!^{n_x} q^{xn_x}(1-q)^{(m-x)n_x}}{(x!)^{n_x}[(m-x)!]^{n_x}}, \\
l(q) &= \sum_{x=0}^{m}[n_x \ln m! + xn_x \ln q + (m-x)n_x \ln(1-q)] \\
&\quad - \sum_{x=0}^{m}[n_x \ln x! + n_x \ln(m-x)!], \\
l'(q) &= \sum_{x=0}^{m} \frac{xn_x}{q} - \frac{(m-x)n_x}{1-q} = \frac{n\bar{x}}{q} - \frac{mn - n\bar{x}}{1-q} = 0, \\
\hat{q} &= \frac{\bar{x}}{m}.
\end{aligned}$$

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

12.2.5 Exercises

12.20 Repeat Example 12.8 using the inverse exponential, inverse gamma with $\alpha = 2$, and inverse gamma distributions. Compare your estimates with the method-of-moments estimates.

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

12.22 Determine maximum likelihood estimates for Data Set B using the inverse exponential, gamma, and inverse gamma distributions. Assume the

Table 12.4 Data for Exercise 12.27

Age last observed	Cause
1.7	Death
1.5	Censoring
2.6	Censoring
3.3	Death
3.5	Censoring

data have been censored at 250 and then compare your answers to those obtained in Example 12.8 and Exercise 12.20.

12.23 Repeat Example 12.10 using a Pareto distribution with both parameters unknown.

12.24 Repeat Example 12.11, this time finding the distribution of the time to surrender.

12.25 Repeat Example 12.12, 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.

12.26 (*) Lives are observed in order to estimate q_{35}. Ten lives are first observed at age 35.4 and 6 die prior to age 36 while the other 4 survive to age 36. An additional 20 lives are first observed at age 35 and 8 die prior to age 36 with the other 12 surviving to age 36. Determine the maximum likelihood estimate of q_{35} given that the time to death from age 35 has density function $f(t) = w$, $0 \le t \le 1$, with $f(t)$ unspecified for $t > 1$.

12.27 (*) The model has hazard rate function $h(t) = \lambda_1$, $0 \le t < 2$, and $h(t) = \lambda_2$, $t \ge 2$. Five items are observed from age zero, with the results in Table 12.4. Determine the maximum likelihood estimates of λ_1 and λ_2.

12.28 (*) Your goal is to estimate q_x. The time to death for a person age x has a constant density function. In a mortality study, 10 lives were first observed at age x. Of them, 1 died and 1 was removed from observation alive at age $x + 0.5$. Determine the maximum likelihood estimate of q_x.

12.29 (*) Ten lives are subject to the survival function

$$S(t) = \left(1 - \frac{t}{k}\right)^{1/2}, \quad 0 \le t \le k,$$

where t is time since birth. There are 10 lives observed from birth. At time 10, 2 of the lives die and the other 8 are withdrawn from observation. Determine the maximum likelihood estimate of k.

12.30 (*) 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.

12.31 (*) One hundred people are observed at age 35. Of them, 15 leave the study at age 35.6, 10 die sometime between ages 35 and 35.6, and 3 die sometime after age 35.6 but before age 36. The remaining 72 people survive to age 36. Determine the product-limit estimate of q_{35} and the maximum likelihood estimate of q_{35}. For the latter, assume the time to death is uniform between ages 35 and 36.

12.32 (*) The survival function is $S(t) = 1 - t/w$, $0 \le t \le w$. Five claims were studied in order to estimate the distribution of the time from reporting to settlement. After five years, four of the claims were settled, the times being 1, 3, 4, and 4. Actuary X then estimates w using maximum likelihood. Actuary Y prefers to wait until all claims are settled. The fifth claim is settled after six years, at which time actuary Y estimates w by maximum likelihood. Determine the two estimates.

12.33 (*) Four automobile engines were first observed when they were three years old. They were then observed for r additional years. By that time three of the engines had failed, with the failure ages being 4, 5, and 7. The fourth engine was still working at age $3 + r$. The survival function has the uniform distribution on the interval 0 to w. The maximum likelihood estimate of w is 13.67. Determine r.

12.34 (*) 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.

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

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

Determine the maximum likelihood estimate of α.

12.36 (*) 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:

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

12.37 (*) 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.

12.38 (*) 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 $\theta_S = 2\theta_P$.

12.39 (*) 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.

12.40 (*) 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.

12.41 (*) 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.

12.42 (*) 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.

12.43 (*) 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.

12.44 A random sample of 10 claims obtained from a gamma distribution is given below:

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.

12.45 A random sample of five claims from a lognormal distribution is given below:

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.

12.46 (*) 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 maximum likelihood estimator of θ.

12.47 (*) 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 1.8 3.4 6.9 4.0

Determine the maximum likelihood estimate of β.

12.48 (*) 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 maximum likelihood estimator of α.

12.49 (*) 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:

43 145 233 396 775

Determine the maximum likelihood estimate of α.

Table 12.5 Data for Exercise 12.51

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

12.50 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 $\Pr(X > 200)$.

12.51 The data in Table 12.5 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.

12.52 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 $(x_1, \cdots, x_n)$, show that the maximum likelihood estimates of μ and θ are

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

and

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

12.53 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 $m = \sum_{j=1}^{n} m_j$.

12.3 VARIANCE AND INTERVAL ESTIMATION

In general, it is not easy to determine the variance of complicated estimators such as the maximum likelihood estimator. 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 multi-parameter generalization is taken from [112] 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 12.13 *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 implies that the derivative may be taken outside the integral and so we are just differentiating the constant* 1.[7]

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

[7]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 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* $\int H(x)f(x;\theta)\,dx < \infty$ *with* $\left|\dfrac{\partial^3}{\partial\theta^3}\ln f(x;\theta)\right| < H(x)$. *This 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 maximum likelihood estimator* $\hat{\theta}_n$ *converges to a normal distribution with mean* θ *and variance such that* $I(\theta)\,\mathrm{Var}(\hat{\theta}_n) \to 1$, *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, the last statement 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 $\mathrm{Var}(\hat{\theta}_n)$. The quantity $I(\theta)$ is called the **information** (sometimes more specifically, **Fisher's information**). It follows from this result that the maximum likelihood estimator 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.[8]

The results stated above assume that the sample consists of independent and identically distributed random observations. A more general version of

[8] For an example of a situation where these conditions do not hold, see Exercise 12.55.

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.

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[9] of this distribution is obtained from the inverse of the matrix with (r,s)th element,

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

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. 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 maximum likelihood estimator.

Example 12.14 *Estimate the covariance matrix of the maximum likelihood estimator 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

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

[9] 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 second partial derivatives are

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

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

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

Changing the signs and inverting produce an estimate of the covariance matrix (it is an estimate because Theorem 12.13 only provides the covariance matrix in the limit). It is

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

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

$$\widehat{\mathrm{Var}}(\hat{\mu},\hat{\sigma})=\begin{bmatrix}0.0965 & 0\\ 0 & 0.0483\end{bmatrix}. \tag{12.1}$$

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 estimates are asymptotically uncorrelated. For the particular case of the lognormal distribution, that is also true 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:

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

To obtain the information matrix, it is necessary to take both derivatives and expected values. This 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 12.15 *Estimate the covariance in the previous example using the observed information.*

Substituting the observations into the second derivatives produces

$$\begin{aligned}\frac{\partial^2 l}{\partial \mu^2} &= -\frac{n}{\sigma^2}=-\frac{20}{\sigma^2},\\ \frac{\partial^2 l}{\partial \sigma\,\partial \mu} &= -2\sum_{j=1}^{n}\frac{\ln x_j-\mu}{\sigma^3}=-2\frac{122.7576-20\mu}{\sigma^3},\\ \frac{\partial^2 l}{\partial \sigma^2} &= \frac{n}{\sigma^2}-3\sum_{j=1}^{n}\frac{(\ln x_j-\mu)^2}{\sigma^4}=\frac{20}{\sigma^2}-3\frac{792.0801-245.5152\mu+20\mu^2}{\sigma^4}.\end{aligned}$$

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 (12.1). 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

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

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 12.16 *Repeat the previous example 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

$$\begin{aligned}\frac{\partial^2 l}{\partial \mu^2} &\doteq \frac{l(6.13796, 1.3894) - 2l(6.1379, 1.3894) + l(6.13784, 1.3894)}{(0.00006)^2} \\ &= \frac{-157.71389308198 - 2(-157.71389304968) + (-157.71389305468)}{(0.00006)^2} \\ &= -10.3604.\end{aligned}$$

The other two approximations are

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

We see that here the approximation works very well. □

The information matrix provides a method for assessing the quality of the maximum likelihood estimators of a distribution's parameters. However, we are often more interested in a quantity 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 is very difficult to evaluate the mean and variance of this random variable because it is a complex function of two variables that already have complex distributions. The following theorem (from [108]) can help. The method is often called the **delta method**.

Theorem 12.17 *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 $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}/n$, where neither $\boldsymbol{\theta}$ 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 $g(\boldsymbol{\theta})$ and variance $(\partial \mathbf{g})^T \boldsymbol{\Sigma} (\partial \mathbf{g})/n$, where $\partial \mathbf{g}$ is the vector of first derivatives, that is, $\partial \mathbf{g} = (\partial g/\partial \theta_1, \ldots, \partial g/\partial \theta_k)^T$ and it is to be evaluated at $\boldsymbol{\theta}$, 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 σ^2/n. Then $g(\hat{\theta})$ has an asymptotic normal distribution with mean $g(\theta)$ and asymptotic variance $[g'(\theta)](\sigma^2/n)[g'(\theta)] = g'(\theta)^2\sigma^2/n$.

Example 12.18 *Use the delta method to approximate the variance of the maximum likelihood estimator of the probability that an observation from an exponential distribution exceeds 200. Apply this result to Data Set B.*

From Example 12.8 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 12.19 *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 12.14 we have $\hat{\mu} = 6.1379$ and $\hat{\sigma} = 1.3894$ and an estimated covariance matrix of

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

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 is 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 maximum likelihood estimator is asymptotically UMVUE. □

12.3.1 Exercises

12.54 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 12.8.

12.55 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 9.7 and 9.10 to show that this estimator is asymptotically unbiased and to obtain its variance. Show that Theorem 12.13 yields a negative estimate of the variance and that item (ii) in the conditions does not hold.

12.56 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 12.54.

12.57 (*) 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.

12.58 (*) 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 maximum likelihood estimator $(\hat{\alpha}, \hat{\beta})$.

12.59 In Exercise 12.39 two maximum likelihood estimates were obtained for the same model. The second estimate was based on more information than the first one. It would be reasonable to expect that the second estimate is more accurate. Confirm this by estimating the variance of each of the two estimators. Do your calculations using the observed likelihood.

12.60 This is a continuation of Exercise 12.43. 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 maximum likelihood estimator of p.

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

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

12.61 This is a continuation of Exercise 12.46. 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 maximum likelihood estimator of θ.

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

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

12.62 (*) 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 maximum likelihood estimate of θ is $\hat{\theta} = 2$. Approximate the MSE of $\hat{\theta}$.

12.63 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 maximum likelihood estimator of λ.

(b) Approximate the variance of the maximum likelihood estimator of λ.

12.64 Estimate the covariance matrix of the maximum likelihood estimators for the data in Exercise 12.44 with both α and θ unknown. Do this by computing approximate derivatives of the loglikelihood. Then construct a 95% confidence interval for the mean.

12.65 Estimate the variance of the maximum likelihood estimator for Exercise 12.49 and use it to construct a 95% confidence interval for $\mathrm{E}(X \wedge 500)$.

12.66 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

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

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 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.

12.4 BAYESIAN ESTIMATION

All of the previous discussion on estimation has assumed a frequentist approach. That is, the population distribution is fixed but unknown, and our decisions are concerned not only with the sample we obtained from the population but also with the possibilities attached to other samples that might have been obtained. The Bayesian approach assumes that only the data actually observed are relevant and it is the population that is variable. For parameter estimation the following definitions describe the process and then Bayes' theorem provides the solution.

12.4.1 Definitions and Bayes' theorem

Definition 12.20 *The **prior distribution** is a probability distribution over the space of possible parameter values. It is denoted $\pi(\theta)$ and represents our opinion concerning the relative chances that various values of θ are the true value of the parameter.*

As before, the parameter θ may be scalar or vector valued. Determination of the prior distribution has always been one of the barriers to the widespread acceptance of Bayesian methods. It is almost certainly the case that your experience has provided some insights about possible parameter values before the first data point has been observed. (If you have no such opinions, perhaps the wisdom of the person who assigned this task to you should be questioned.) The difficulty is translating this knowledge into a probability distribution. An excellent discussion about prior distributions and the foundations of Bayesian analysis can be found in Lindley [83], and for a discussion about issues surrounding the choice of Bayesian versus frequentist methods, see Efron [32]. The book by Klugman [77] contains more detail on the Bayesian approach along with several actuarial applications. More recent articles applying Bayesian methods to actuarial problems include [25], [101], [119], and [133]. A good source for a thorough mathematical treatment of Bayesian methods is the text by Berger [13]. In recent years many advancements in Bayesian calculations have occurred. A good resource is [22]. Scollnik [118] has demonstrated how the computer program WINBUGS can be used to provide Bayesian solutions to actuarial problems.

Due to the difficulty of finding a prior distribution that is convincing (you will have to convince others that your prior opinions are valid) and the possibility that you may really have no prior opinion, the definition of prior distribution can be loosened.

Definition 12.21 *An **improper prior distribution** is one for which the probabilities (or pdf) are nonnegative but their sum (or integral) is infinite.*

A great deal of research has gone into the determination of a so-called **noninformative** or **vague** prior. Its purpose is to reflect minimal knowledge. Universal agreement on the best way to construct a vague prior does not exist. However, there is agreement that the appropriate noninformative prior for a scale parameter is $\pi(\theta) = 1/\theta, \theta > 0$. Note that this is an improper prior.

For a Bayesian analysis, the model is no different than before.

Definition 12.22 *The **model distribution** is the probability distribution for the data as collected given a particular value for the parameter. Its pdf is denoted $f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)$, where vector notation for $\mathbf{x}$ is used to remind us that all the data appear here. Also note that this is identical to the likelihood function and so that name may also be used at times.*

If the vector of observations $\mathbf{x} = (x_1, \ldots, x_n)^T$ consists of independent and identically distributed random variables, then

$$f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = f_{X|\Theta}(x_1|\theta)\cdots f_{X|\Theta}(x_n|\theta).$$

We use concepts from multivariate statistics to obtain two more definitions. In both cases, as well as in the following, integrals should be replaced by sums if the distributions are discrete.

Definition 12.23 *The **joint distribution** has pdf*

$$f_{\mathbf{X},\Theta}(\mathbf{x},\theta) = f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\pi(\theta).$$

Definition 12.24 *The **marginal distribution** of* $\mathbf{x}$ *has pdf*

$$f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\pi(\theta)\,d\theta.$$

Compare this definition to that of a mixture distribution given by (4.4) on page 59. The final two quantities of interest are the following.

Definition 12.25 *The **posterior distribution** is the conditional probability distribution of the parameters given the observed data. It is denoted* $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$.

Definition 12.26 *The **predictive distribution** is the conditional probability distribution of a new observation* y *given the data* $\mathbf{x}$. *It is denoted* $f_{Y|\mathbf{X}}(y|\mathbf{x})$.[10]

These last two items are the key output of a Bayesian analysis. The posterior distribution tells us how our opinion about the parameter has changed once we have observed the data. The predictive distribution tells us what the next observation might look like given the information contained in the data (as well as, implicitly, our prior opinion). Bayes' theorem tells us how to compute the posterior distribution.

Theorem 12.27 *The posterior distribution can be computed as*

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\pi(\theta)}{\int f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\pi(\theta)\,d\theta} \tag{12.2}$$

while the predictive distribution can be computed as

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \int f_{Y|\Theta}(y|\theta)\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\,d\theta, \tag{12.3}$$

where $f_{Y|\Theta}(y|\theta)$ *is the pdf of the new observation, given the parameter value.*

[10] In this section and in any subsequent Bayesian discussions, we reserve $f(\cdot)$ for distributions concerning observations (such as the model and predictive distributions) and $\pi(\cdot)$ for distributions concerning parameters (such as the prior and posterior distributions). The arguments will usually make it clear which particular distribution is being used. To make matters explicit, we also employ subscripts to enable us to keep track of the random variables.

The predictive distribution can be interpreted as a mixture distribution where the mixing is with respect to the posterior distribution. The following example illustrates the above definitions and results. The setting, though not the data, is taken from Meyers [92].

Example 12.28 *The following amounts were paid on a hospital liability policy:*

125 132 141 107 133 319 126 104 145 223

The amount of a single payment has the single-parameter Pareto distribution with $\theta = 100$ *and* α *unknown. The prior distribution has the gamma distribution with* $\alpha = 2$ *and* $\theta = 1$. *Determine all of the relevant Bayesian quantities.*

The prior density has a gamma distribution and is

$$\pi(\alpha) = \alpha e^{-\alpha}, \quad \alpha > 0,$$

while the model is (evaluated at the data points)

$$f_{\mathbf{X}|A}(\mathbf{x}|\alpha) = \frac{\alpha^{10}(100)^{10\alpha}}{\left(\prod_{j=1}^{10} x_j^{\alpha+1}\right)} = \alpha^{10}e^{-3.801121\alpha-49.852823}.$$

The joint density of $\mathbf{x}$ and A is (again evaluated at the data points)

$$f_{\mathbf{X},A}(\mathbf{x},\alpha) = \alpha^{11}e^{-4.801121\alpha-49.852823}.$$

The posterior distribution of α is

$$\pi_{A|\mathbf{X}}(\alpha|\mathbf{x}) = \frac{\alpha^{11}e^{-4.801121\alpha-49.852823}}{\int_0^\infty \alpha^{11}e^{-4.801121\alpha-49.852823}\,d\alpha} = \frac{\alpha^{11}e^{-4.801121\alpha}}{(11!)(1/4.801121)^{12}}. \tag{12.4}$$

There is no need to evaluate the integral in the denominator. Because we know that the result must be a probability distribution, the denominator is just the appropriate normalizing constant. A look at the numerator reveals that we have a gamma distribution with $\alpha = 12$ and $\theta = 1/4.801121$.

The predictive distribution is

$$\begin{aligned} f_{Y|\mathbf{X}}(y|\mathbf{x}) &= \int_0^\infty \frac{\alpha 100^\alpha}{y^{\alpha+1}} \frac{\alpha^{11}e^{-4.801121\alpha}}{(11!)(1/4.801121)^{12}}\,d\alpha \\ &= \frac{1}{y(11!)(1/4.801121)^{12}} \int_0^\infty \alpha^{12}e^{-(0.195951+\ln y)\alpha}\,d\alpha \\ &= \frac{1}{y(11!)(1/4.801121)^{12}} \frac{(12!)}{(0.195951+\ln y)^{13}} \\ &= \frac{12(4.801121)^{12}}{y(0.195951+\ln y)^{13}}, \quad y > 100. \end{aligned} \tag{12.5}$$

While this density function may not look familiar, you are asked to show in Exercise 12.67 that $\ln Y - \ln 100$ has the Pareto distribution. □

12.4.2 Inference and prediction

In one sense the analysis is complete. We begin with a distribution that quantifies our knowledge about the parameter and/or the next observation and we end with a revised distribution. But we have a suspicion that your boss may not be satisfied if you produce a distribution in response to his or her request. No doubt a specific number, perhaps with a margin for error, is what is desired. The usual Bayesian solution is to pose a loss function.

Definition 12.29 *A* ***loss function*** $l_j(\hat{\theta}_j, \theta_j)$ *describes the penalty paid by the investigator when* $\hat{\theta}_j$ *is the estimate and* θ_j *is the true value of the* j*th parameter.*

It would also be possible to have a multidimensional loss function $l(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ which allowed the loss to depend simultaneously on the errors in the various parameter estimates.

Definition 12.30 *The* ***Bayes estimate*** *for a given loss function is the one that minimizes the expected loss given the posterior distribution of the parameter in question.*

The three most commonly used loss functions are defined as follows.

Definition 12.31 *For* ***squared-error loss*** *the loss function is (all subscripts are dropped for convenience)* $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. *For* ***absolute loss*** *it is* $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$. *For* ***zero–one loss*** *it is* $l(\hat{\theta}, \theta) = 0$ *if* $\hat{\theta} = \theta$ *and is* 1 *otherwise.*

The following theorem indicates the Bayes estimates for these three common loss functions.

Theorem 12.32 *For squared-error loss the Bayes estimate is the mean of the posterior distribution, for absolute loss it is a median, and for zero–one loss it is a mode.*

Note that there is no guarantee that the posterior mean exists or that the posterior median or mode will be unique. When not otherwise specified, the term *Bayes estimate* will refer to the posterior mean.

Example 12.33 (Example 12.28 continued) *Determine the three Bayes estimates of* α.

The mean of the posterior gamma distribution is $\alpha\theta = 12/4.801121 = 2.499416$. The median of 2.430342 must be determined numerically while the mode is $(\alpha - 1)\theta = 11/4.801121 = 2.291132$. Note that the α used here is the parameter of the posterior gamma distribution, not the α for the single-parameter Pareto distribution that we are trying to estimate. □

For forecasting purposes, the expected value of the predictive distribution is often of interest. It can be thought of as providing a point estimate of the $(n+1)$th observation given the first n observations and the prior distribution. It is

$$\begin{aligned}
\mathrm{E}(Y|\mathbf{x}) &= \int y f_{Y|\mathbf{x}}(y|\mathbf{x})dy \\
&= \int y \int f_{Y|\Theta}(y|\theta)\pi_{\Theta|\mathbf{x}}(\theta|\mathbf{x})d\theta dy \\
&= \int \pi_{\Theta|\mathbf{x}}(\theta|\mathbf{x}) \int y f_{Y|\Theta}(y|\theta)dyd\theta \\
&= \int \mathrm{E}(Y|\theta)\pi_{\Theta|\mathbf{x}}(\theta|\mathbf{x})d\theta. \qquad (12.6)
\end{aligned}$$

Equation (12.6) can be interpreted as a weighted average using the posterior distribution as weights.

Example 12.34 (Example 12.28 continued) *Determine the expected value of the* 11*th observation, given the first* 10.

For the single-parameter Pareto distribution, $\mathrm{E}(Y|\alpha) = 100\alpha/(\alpha - 1)$ for $\alpha > 1$. Because the posterior distribution assigns positive probability to values of $\alpha \leq 1$, the expected value of the predictive distribution is not defined. □

The Bayesian equivalent of a confidence interval is easy to construct. The following definition will suffice.

Definition 12.35 *The points* $a < b$ *define a* $100(1-\alpha)\%$ ***credibility interval*** *for* θ_j *provided that* $\Pr(a \leq \Theta_j \leq b|\mathbf{x}) \geq 1 - \alpha$.

The use of the term credibility has no relationship to its use in actuarial analyses as developed in Chapter 16. The inequality is present for the case where the posterior distribution of θ_j is discrete. Then it may not be possible for the probability to be exactly $1 - \alpha$. This definition does not produce a unique solution. The following theorem indicates one way to produce a unique interval.

Theorem 12.36 *If the posterior random variable* $\theta_j|\mathbf{x}$ *is continuous and unimodal, then the* $100(1 - \alpha)\%$ *credibility interval with smallest width* $b - a$ *is the unique solution to*

$$\begin{aligned}
\int_a^b \pi_{\Theta_j|\mathbf{x}}(\theta_j|\mathbf{x})\, d\theta_j &= 1 - \alpha, \\
\pi_{\Theta|\mathbf{x}}(a|\mathbf{x}) &= \pi_{\Theta|\mathbf{x}}(b|\mathbf{x}).
\end{aligned}$$

This interval is a special case of a highest posterior density (HPD) credibility set.

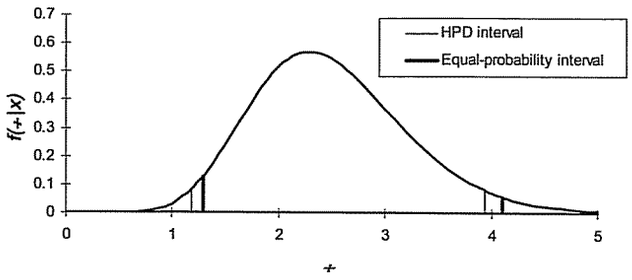


Fig. 12.1 Two Bayesian credibility intervals.

The following example may clarify the theorem.

Example 12.37 (Example 12.28 continued) *Determine the shortest* 95% *credibility interval for the parameter* α. *Also determine the interval that places* 2.5% *probability at each end.*

The two equations from Theorem 12.36 are

$$\begin{aligned} \Pr(a \le A \le b|\mathbf{x}) &= \Gamma(12; 4.801121b) - \Gamma(12; 4.801121a) = 0.95, \\ a^{11}e^{-4.801121a} &= b^{11}e^{-4.801121b}. \end{aligned}$$

and numerical methods can be used to find the solution $a = 1.1832$ and $b = 3.9384$. The width of this interval is 2.7552.

Placing 2.5% probability at each end yields the two equations

$$\Gamma(12; 4.801121b) = 0.975, \quad \Gamma(12; 4.801121a) = 0.025.$$

This solution requires either access to the inverse of the incomplete gamma function or the use of root-finding techniques with the incomplete gamma function itself. The solution is $a = 1.2915$ and $b = 4.0995$. The width is 2.8080, wider than the first interval. Figure 12.1 shows the difference in the two intervals. The solid vertical bars represent the HPD interval. The total area to the left and right of these bars is 0.05. Any other 95% interval must also have this probability. To create the interval with 0.025 probability on each side, both bars must be moved to the right. To subtract the same probability on the right end that is added on the left end, the right limit must be moved a greater distance because the posterior density is lower over that interval than it is on the left end. This must lead to a wider interval. □

The following definition provides the equivalent result for any posterior distribution.

Definition 12.38 *For any posterior distribution the* $100(1-\alpha)\%$ ***HPD credibility set*** *is the set of parameter values* C *such that*

$$\Pr(\theta_j \in C) \ge 1 - \alpha \tag{12.7}$$

and

$$C = \{\theta_j : \pi_{\Theta_j|\mathbf{X}}(\theta_j|\mathbf{x}) \ge c\} \text{ for some } c,$$

where c *is the largest value for which the inequality* (12.7) *holds.*

This set may be the union of several intervals (which can happen with a multimodal posterior distribution). This definition produces the set of minimum total width that has the required posterior probability. Construction of the set is done by starting with a high value of c and then lowering it. As it decreases, the set C gets larger, as does the probability. The process continues until the probability reaches $1 - \alpha$. It should be obvious to see how the definition can be extended to the construction of a simultaneous credibility set for a vector of parameters, $\boldsymbol{\theta}$.

Sometimes it is the case that, while computing posterior probabilities is difficult, computing posterior moments may be easy. We can then use the Bayesian central limit theorem. The following is a paraphrase from Berger [13], p. 224.

Theorem 12.39 *If* $\pi(\theta)$ *and* $f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)$ *are both twice differentiable in the elements of* θ *and other commonly satisfied assumptions hold, then the posterior distribution of* Θ *given* $\mathbf{X} = \mathbf{x}$ *is asymptotically normal.*

The "commonly satisfied assumptions" are like those in Theorem 12.13. As in that theorem, it is possible to do further approximations. In particular, the asymptotic normal distribution also results if the posterior mode is substituted for the posterior mean and/or if the posterior covariance matrix is estimated by inverting the matrix of second partial derivatives of the negative logarithm of the posterior density.

Example 12.40 (Example 12.28 continued) *Construct a* 95% *credibility interval for* α *using the Bayesian central limit theorem.*

The posterior distribution has a mean of 2.499416 and a variance of $\alpha\theta^2 = 0.520590$. Using the normal approximation, the credibility interval is $2.499416 \pm 1.96(0.520590)^{1/2}$, which produces $a = 1.0852$ and $b = 3.9136$. This interval (with regard to the normal approximation) is HPD due to the symmetry of the normal distribution.

The approximation is centered at the posterior mode of 2.291132 (see Example 12.33). The second derivative of the negative logarithm of the posterior density [from (12.4)] is

$$-\frac{d^2}{d\alpha^2}\ln\left[\frac{\alpha^{11}e^{-4.801121\alpha}}{(11!)(1/4.801121)^{12}}\right] = \frac{11}{\alpha^2}.$$

The variance estimate is the reciprocal. Evaluated at the modal estimate of α we get $(2.291132)^2/11 = 0.477208$ for a credibility interval of $2.29113 \pm 1.96(0.477208)^{1/2}$, which produces $a = 0.9372$ and $b = 3.6451$. □

The same concepts can apply to the predictive distribution. However, the Bayesian central limit theorem does not help here because the predictive sample has only one member. The only potential use for it is that for a large original sample size we can replace the true posterior distribution in (12.3) with a multivariate normal distribution.

Example 12.41 (Example 12.28 continued) *Construct a 95% highest density prediction interval for the next observation.*

It is easy to see that the predictive density function (12.5) is strictly decreasing. Therefore the region with highest density runs from $a = 100$ to b. The value of b is determined from

$$\begin{aligned} 0.95 &= \int_{100}^{b} \frac{12(4.801121)^{12}}{y(0.195951 + \ln y)^{13}}\, dy \\ &= \int_{0}^{\ln(b/100)} \frac{12(4.801121)^{12}}{(4.801121 + x)^{13}}\, dx \\ &= 1 - \left[\frac{4.801121}{4.801121 + \ln(b/100)}\right]^{12} \end{aligned}$$

and the solution is $b = 390.1840$. It is interesting to note that the mode of the predictive distribution is 100 (because the pdf is strictly decreasing) while the mean is infinite (with $b = \infty$ and an additional y in the integrand, after the transformation, the integrand is like $e^x x^{-13}$, which goes to infinity as x goes to infinity). □

The following example revisits a calculation done in Section 4.6.3. There the negative binomial distribution was derived as a gamma mixture of Poisson variables. The following example shows how the same calculations arise in a Bayesian context.

Example 12.42 *The number of claims in one year on a given policy is known to have a Poisson distribution. The parameter is not known, but the prior distribution has a gamma distribution with parameters α and θ. Suppose in the past year the policy had x claims. Use Bayesian methods to estimate the number of claims in the next year. Then repeat these calculations assuming claim counts for the past n years, $x_1, \ldots, x_n$.*

The key distributions are (where $x = 0, 1, \ldots$, $\lambda, \alpha, \theta > 0$):

$$\text{Prior:} \quad \pi(\lambda) = \frac{\lambda^{\alpha-1} e^{-\lambda/\theta}}{\Gamma(\alpha)\theta^{\alpha}}$$

$$\text{Model:} \quad p(x|\lambda) = \frac{\lambda^{x} e^{-\lambda}}{x!}$$

$$\text{Joint:} \quad p(x, \lambda) = \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}}{x!\Gamma(\alpha)\theta^{\alpha}}$$

$$\begin{aligned} \text{Marginal:} \quad p(x) &= \int_0^{\infty} \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}}{x!\Gamma(\alpha)\theta^{\alpha}}\, d\lambda \\ &= \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)\theta^{\alpha}(1+1/\theta)^{x+\alpha}} \\ &= \binom{x+\alpha-1}{x} \left(\frac{1}{1+\theta}\right)^{\alpha} \left(\frac{\theta}{1+\theta}\right)^{x} \end{aligned}$$

$$\begin{aligned} \text{Posterior:} \quad \pi(\lambda|x) &= \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}}{x!\Gamma(\alpha)\theta^{\alpha}} \Big/ \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)\theta^{\alpha}(1+1/\theta)^{x+\alpha}} \\ &= \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}(1+1/\theta)^{x+\alpha}}{\Gamma(x+\alpha)} \end{aligned}$$

The marginal distribution is negative binomial with $r = \alpha$ and $\beta = \theta$. The posterior distribution is gamma with shape parameter "α" equal to $x+\alpha$ and scale parameter "θ" equal to $(1 + 1/\theta)^{-1} = \theta/(1 + \theta)$. The Bayes estimate of the Poisson parameter is the posterior mean, $(x + \alpha)\theta/(1 + \theta)$. For the predictive distribution, (12.3) gives

$$\begin{aligned} p(y|x) &= \int_0^{\infty} \frac{\lambda^{y} e^{-\lambda}}{y!} \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}(1+1/\theta)^{x+\alpha}}{\Gamma(x+\alpha)}\, d\lambda \\ &= \frac{(1+1/\theta)^{x+\alpha}}{y!\Gamma(x+\alpha)} \int_0^{\infty} \lambda^{y+x+\alpha-1} e^{-(2+1/\theta)\lambda}\, d\lambda \\ &= \frac{(1+1/\theta)^{x+\alpha}\Gamma(y+x+\alpha)}{y!\Gamma(x+\alpha)(2+1/\theta)^{y+x+\alpha}}, \quad y = 0, 1, \ldots, \end{aligned}$$

and some rearranging shows this to be a negative binomial distribution with $r = x + \alpha$ and $\beta = \theta/(1 + \theta)$. The expected number of claims for the next year is $(x + \alpha)\theta/(1 + \theta)$. Alternatively, from (12.6),

$$\mathrm{E}(Y|x) = \int_0^{\infty} \lambda \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}(1+1/\theta)^{x+\alpha}}{\Gamma(x+\alpha)}\, d\lambda = \frac{(x+\alpha)\theta}{1+\theta}.$$

For a sample of size n, the key change is that the model distribution is now

$$p(\mathbf{x}|\lambda) = \frac{\lambda^{x_1+\cdots+x_n} e^{-n\lambda}}{x_1! \cdots x_n!}.$$

Following this through, the posterior distribution is still gamma, now with shape parameter $x_1 + \cdots + x_n + \alpha = n\bar{x} + \alpha$ and scale parameter $\theta/(1+n\theta)$. The predictive distribution is still negative binomial, now with $r = n\bar{x} + \alpha$ and $\beta = \theta/(1+n\theta)$. □

When only moments are needed, the double-expectation formulas can be very useful. Provided the moments exist, for any random variables X and Y,

$$\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y|X)], \tag{12.8}$$

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathrm{E}(Y|X)]. \tag{12.9}$$

For the predictive distribution,

$$\begin{aligned}\mathrm{E}(Y|\mathbf{x}) &= \mathrm{E}_{\Theta|\mathbf{x}}[\mathrm{E}(Y|\Theta,\mathbf{x})] \\ &= \mathrm{E}_{\Theta|\mathbf{x}}[\mathrm{E}(Y|\Theta)]\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(Y|\mathbf{x}) &= \mathrm{E}_{\Theta|\mathbf{x}}[\mathrm{Var}(Y|\Theta,\mathbf{x})] + \mathrm{Var}_{\Theta|\mathbf{x}}[\mathrm{E}(Y|\Theta,\mathbf{x})] \\ &= \mathrm{E}_{\Theta|\mathbf{x}}[\mathrm{Var}(Y|\Theta)] + \mathrm{Var}_{\Theta|\mathbf{x}}[\mathrm{E}(Y|\Theta)].\end{aligned}$$

The simplification on the inner expected value and variance results from the fact that, if Θ is known, the value of $\mathbf{x}$ provides no additional information about the distribution of Y. This is simply a restatement of (12.6).

Example 12.43 *Apply these formulas to obtain the predictive mean and variance for the previous example. Then anticipate the credibility formulas of Chapter* 16.

The predictive mean uses $\mathrm{E}(Y|\lambda) = \lambda$. Then,

$$\mathrm{E}(Y|\mathbf{x}) = \mathrm{E}(\lambda|\mathbf{x}) = \frac{(n\bar{x}+\alpha)\theta}{1+n\theta}.$$

The predictive variance uses $\mathrm{Var}(Y|\lambda) = \lambda$, and then

$$\begin{aligned}\mathrm{Var}(Y|\mathbf{x}) &= \mathrm{E}(\lambda|\mathbf{x}) + \mathrm{Var}(\lambda|\mathbf{x}) \\ &= \frac{(n\bar{x}+\alpha)\theta}{1+n\theta} + \frac{(n\bar{x}+\alpha)\theta^2}{(1+n\theta)^2} \\ &= (n\bar{x}+\alpha)\frac{\theta}{1+n\theta}\left(1+\frac{\theta}{1+n\theta}\right).\end{aligned}$$

These agree with the mean and variance of the known negative binomial distribution for y. However, these quantities were obtained from moments of the model (Poisson) and posterior (gamma) distributions. The predictive mean can be written as

$$\frac{n\theta}{1+n\theta}\bar{x} + \frac{1}{1+n\theta}\alpha\theta,$$

which is a weighted average of the mean of the data and the mean of the prior distribution. Note that as the sample size increases more weight is placed on the data and less on the prior opinion. The variance of the prior distribution can be increased by letting θ become large. As it should, this also increases the weight placed on the data. The credibility formulas in Chapter 16 generally consist of weighted averages of an estimate from the data and a prior opinion. □

12.4.3 Conjugate prior distributions and the linear exponential family

A large parametric family that includes many of the distributions we have encountered so far has a special use in Bayesian analysis. The definition is as follows:

Definition 12.44 *A random variable X (discrete or continuous) has a distribution which is from the* ***linear exponential family*** *if its pf may be parameterized in terms of a parameter θ and expressed as*

$$f(x;\theta) = \frac{p(x)e^{-\theta x}}{q(\theta)}. \tag{12.10}$$

The function $p(x)$ depends only on x (not on θ), and the function $q(\theta)$ is a normalizing constant. Also, the support of the random variable must not depend on θ. The parameter θ is called the ***natural parameter*** *of the distribution.*

Example 12.45 *Show that the exponential distribution is of the form* (12.10).

The pdf is

$$f(x;\beta) = \beta^{-1}e^{-\beta^{-1}x}.$$

If we let $\theta = 1/\beta$, then the pdf is

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

which is of the form (12.10) with $p(x) = 1$ and $q(\theta) = 1/\theta$. □

Example 12.46 *Show that the Poisson distribution is a member of the linear exponential family.*

The pf is

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(1/x!)e^{-(-\ln\lambda)x}}{e^{\lambda}}.$$

If we let $\theta = -\ln\lambda$, then the pf is

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

which is of the form (12.10) with $p(x) = 1/x!$ and $q(\theta) = e^{e^{-\theta}}$. Note that in this parameterization the Poisson mean is $e^{-\theta}$. □

Example 12.47 *Show that the normal distribution with mean μ and known variance v is a member of the linear exponential family.*

The pdf is

$$\begin{aligned} f(x;\mu,v) &= (2\pi v)^{-1/2}\exp\left[-\frac{1}{2v}(x-\mu)^2\right] \\ &= (2\pi v)^{-1/2}\exp\left(-\frac{x^2}{2v}+\frac{\mu}{v}x-\frac{\mu^2}{2v}\right) \\ &= \frac{(2\pi v)^{-1/2}\exp\left(-\frac{x^2}{2v}\right)\exp\left(\frac{\mu}{v}x\right)}{\exp\left(\frac{\mu^2}{2v}\right)}. \end{aligned}$$

If we let $\theta = -\mu/v$, the pdf is

$$f(x;\theta,v) = \frac{(2\pi v)^{-1/2}\exp\left(-\frac{x^2}{2v}\right)\exp(-\theta x)}{\exp\left(\frac{\theta^2 v}{2}\right)},$$

which is of the form (12.10) with $p(x) = (2\pi v)^{-1/2}\exp[-x^2/(2v)]$, and $q(\theta) = \exp(\theta^2 v/2)$. □

We now find the mean and variance of the distribution defined by (12.10). First, note that

$$\ln f(x;\theta) = \ln p(x) - \theta x - \ln q(\theta).$$

Differentiate with respect to θ to obtain

$$\frac{\partial}{\partial\theta}f(x;\theta) = \left[-x - \frac{q'(\theta)}{q(\theta)}\right]f(x;\theta). \tag{12.11}$$

Integrate (or sum) over the range of x (known not to depend on θ) to obtain

$$\int \frac{\partial}{\partial\theta}f(x;\theta)\,dx = -\int x f(x;\theta)\,dx - \frac{q'(\theta)}{q(\theta)}\int f(x;\theta)\,dx.$$

On the left-hand side, interchange the order of differentiation and integration (or summation) to obtain

$$\frac{\partial}{\partial\theta}\left[\int f(x;\theta)\,dx\right] = -\int x f(x;\theta)\,dx - \frac{q'(\theta)}{q(\theta)}\int f(x;\theta)\,dx.$$

We know that $\int f(x;\theta)\,dx = 1$ and $\int x f(x;\theta)\,dx = \mathrm{E}(X)$ and thus

$$\frac{\partial}{\partial\theta}(1) = -\mathrm{E}(X) - \frac{q'(\theta)}{q(\theta)}.$$

In other words, the mean is

$$\mathrm{E}(X) = \mu(\theta) = -\frac{q'(\theta)}{q(\theta)} = -\frac{d}{d\theta}\ln q(\theta). \tag{12.12}$$

To obtain the variance, (12.11) may first be rewritten as

$$\frac{\partial}{\partial\theta}f(x;\theta) = -[x-\mu(\theta)]f(x;\theta).$$

Differentiate again with respect to θ to obtain

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2}f(x;\theta) &= \mu'(\theta)f(x;\theta) - [x-\mu(\theta)]\frac{\partial}{\partial\theta}f(x;\theta) \\ &= \mu'(\theta)f(x;\theta) + [x-\mu(\theta)]^2 f(x;\theta). \end{aligned}$$

Again, integrate over the range of x to obtain

$$\int \frac{\partial^2}{\partial\theta^2}f(x,\theta)\,dx = \mu'(\theta)\int f(x;\theta)\,dx + \int [x-\mu(\theta)]^2 f(x;\theta)\,dx.$$

In other words

$$\int [x-\mu(\theta)]^2 f(x;\theta)\,dx = -\mu'(\theta) + \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)\,dx.$$

Because $\mu(\theta)$ is the mean, the left-hand side is the variance (by definition), and then because the second term on the right-hand side is zero, we obtain

$$\mathrm{Var}(X) = v(\theta) = -\mu'(\theta) = \frac{d^2}{d\theta^2}\ln q(\theta). \tag{12.13}$$

In Example 12.42 it turned out the posterior distribution was of the same type as the prior distribution (gamma). This makes calculations relatively easy. A definition of this concept follows.

Definition 12.48 *A prior distribution is said to be a* ***conjugate prior distribution*** *for a given model if the resulting posterior distribution is of the same type as the prior (but perhaps with different parameters).*

The following theorem shows that, if the model is a member of the linear exponential family, a conjugate prior distribution is easy to find.

Theorem 12.49 *Suppose that given $\Theta = \theta$ the random variables $X_1, \ldots, X_n$ are independent and identically distributed with pf*

$$f_{X_j|\Theta}(X_j|\theta) = \frac{p(x_j)e^{-\theta x_j}}{q(\theta)},$$

where Θ has pdf

$$\pi(\theta) = \frac{[q(\theta)]^{-k}e^{-\theta\mu k}}{c(\mu,k)},$$

where k and μ are parameters of the distribution and $c(\mu,k)$ is the normalizing constant. Then the posterior pf $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ is of the same form as $\pi(\theta)$.

Proof: The posterior distribution is

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto \frac{\left[\prod_{j=1}^n p(x_j)\right] e^{-\theta\Sigma x_j}}{q(\theta)^n} \frac{[q(\theta)]^{-k}e^{-\theta\mu k}}{c(\mu,k)} \\
&\propto [q(\theta)]^{-(k+n)} \exp\left[\left(-\theta\frac{\mu k+\sum x_j}{k+n}\right)(k+n)\right] \\
&\propto [q(\theta)]^{-k^*} \exp(-\theta\mu^* k^*),
\end{aligned}$$

which is of the same form as $\pi(\theta)$ with parameters

$$\begin{aligned}
k^* &= k+n, \\
\mu^* &= \frac{\mu k+\sum x_j}{k+n} = \frac{k}{k+n}\mu + \frac{n}{k+n}\overline{x}.
\end{aligned}$$

□

Example 12.50 *Show that for the Poisson model the conjugate prior as given in Theorem 12.49 is the gamma distribution.*

From Example 12.46 we have that $q(\theta) = \exp(e^{-\theta})$ and $\lambda = \exp(-\theta)$. The prior as given by the theorem is

$$\pi(\theta) \propto \left[\exp(e^{-\theta})\right]^{-k} \exp(-\theta\mu k).$$

Then the prior density for λ is

$$\pi(\lambda) \propto [\exp(\lambda)]^{-k}\lambda^{\mu k}\lambda^{-1} = \lambda^{\mu k-1}e^{-\lambda k},$$

which is a gamma distribution with $\alpha = \mu k$ and $\theta = 1/k$. The term λ^{-1} appears because it is $|d\theta/d\lambda|$, which is needed for the change of variable. □

12.4.4 Computational issues

It should be obvious by now that all Bayesian analyses proceed by taking integrals or sums. So at least conceptually it is always possible to do a Bayesian analysis. However, only in rare cases are the integrals or sums easy to do, and that means most Bayesian analyses will require numerical integration. While one-dimensional integrations are easy to do to a high degree of accuracy, multidimensional integrals are much more difficult to approximate. A great deal of effort has been expended with regard to solving this problem. A number of ingenious methods have been developed. Some of them are summarized in Klugman [77]. However, the one that is widely used today is called Markov chain Monte Carlo simulation. A good discussion of this method can be found in [118] and actuarial applications can be found in [21] and [119].

There is another way which completely avoids computational problems. This is illustrated using the example (in an abbreviated form) from Meyers [92], which also employed this technique. The example also shows how a Bayesian analysis is used to estimate a function of parameters.

Example 12.51 *Data were collected on 100 losses in excess of 100,000. The single-parameter Pareto distribution is to be used with $\theta = 100{,}000$ and α unknown. The objective is to estimate the layer average severity for the layer from 1,000,000 to 5,000,000. For the observations, $\sum_{j=1}^{100} \ln x_j = 1{,}208.4354$.*

The model density is

$$\begin{aligned}
f_{\mathbf{X}|A}(\mathbf{x}|\alpha) &= \prod_{j=1}^{100} \frac{\alpha(100{,}000)^\alpha}{x_j^{\alpha+1}} \\
&= \exp\left[100\ln\alpha + 100\alpha\ln 100{,}000 - (\alpha+1)\sum_{j=1}^{100}\ln x_j\right] \\
&= \exp\left(100\ln\alpha - \frac{100\alpha}{1.75} - 1{,}208.4354\right).
\end{aligned}$$

The density appears in column 3 of Table 12.6. To prevent computer overflow, the value 1,208.4354 was not subtracted prior to exponentiation. This makes the entries proportional to the true density function. The prior density is given in the second column. It was chosen based on a belief that the true value is in the range 1–2.5 and is more likely to be near 1.5 than at the ends. The posterior density is then obtained using (12.2). The elements of the numerator are found in column 4. The denominator is no longer an integral but a sum. The sum is at the bottom of column 4 and then the scaled values are in column 5.

We can see from column 5 that the posterior mode is at $\alpha = 1.7$, as compared to the maximum likelihood estimate of 1.75 (see Exercise 12.69). The posterior mean of α could be found by adding the product of columns 1 and 5. Here we are interested in a layer average severity. For this problem it

Table 12.6 Bayesian estimation of a layer average severity

α	$\pi(\alpha)$	$f(\mathbf{x}\|\alpha)$	$\pi(\alpha)f(\mathbf{x}\|\alpha)$	$\pi(\alpha\|\mathbf{x})$	LAS(α)	$\pi\times$L*	$\pi(\alpha\|\mathbf{x})\mathrm{l}(\alpha)^2$
1.0	0.0400	1.52×10^{-25}	6.10×10^{-27}	0.0000	160,944	0	6,433
1.1	0.0496	6.93×10^{-24}	3.44×10^{-25}	0.0000	118,085	2	195,201
1.2	0.0592	1.37×10^{-22}	8.13×10^{-24}	0.0003	86,826	29	2,496,935
1.3	0.0688	1.36×10^{-21}	9.33×10^{-23}	0.0038	63,979	243	15,558,906
1.4	0.0784	7.40×10^{-21}	5.80×10^{-22}	0.0236	47,245	1,116	52,737,840
1.5	0.0880	2.42×10^{-20}	2.13×10^{-21}	0.0867	34,961	3,033	106,021,739
1.6	0.0832	5.07×10^{-20}	4.22×10^{-21}	0.1718	25,926	4,454	115,480,050
1.7	0.0784	7.18×10^{-20}	5.63×10^{-21}	0.2293	19,265	4,418	85,110,453
1.8	0.0736	7.19×10^{-20}	5.29×10^{-21}	0.2156	14,344	3,093	44,366,353
1.9	0.0688	5.29×10^{-20}	3.64×10^{-21}	0.1482	10,702	1,586	16,972,802
2.0	0.0640	2.95×10^{-20}	1.89×10^{-21}	0.0768	8,000	614	4,915,383
2.1	0.0592	1.28×10^{-20}	7.57×10^{-22}	0.0308	5,992	185	1,106,259
2.2	0.0544	4.42×10^{-21}	2.40×10^{-22}	0.0098	4,496	44	197,840
2.3	0.0496	1.24×10^{-21}	6.16×10^{-23}	0.0025	3,380	8	28,650
2.4	0.0448	2.89×10^{-22}	1.29×10^{-23}	0.0005	2,545	1	3,413
2.5	0.0400	5.65×10^{-23}	2.26×10^{-24}	0.0001	1,920	0	339
	1.0000		2.46×10^{-20}	1.0000		18,827	445,198,597

*$\pi(\alpha|\mathbf{x})\mathrm{LAS}(\alpha)$

is

$$\begin{aligned}\mathrm{LAS}(\alpha) &= \mathrm{E}(X\wedge 5{,}000{,}000)-\mathrm{E}(X\wedge 1{,}000{,}000)\\ &= \frac{100{,}000^{\alpha}}{\alpha-1}\left(\frac{1}{1{,}000{,}000^{\alpha-1}}-\frac{1}{5{,}000{,}000^{\alpha-1}}\right), \quad \alpha\neq 1,\\ &= 100{,}000\left(\ln 5{,}000{,}000-\ln 1{,}000{,}000\right), \quad \alpha=1.\end{aligned}$$

Values of LAS(α) for the 16 possible values of α appear in column 6. The last two columns are then used to obtain the posterior expected values of the layer average severity. The point estimate is the posterior mean, 18,827. The posterior standard deviation is

$$\sqrt{445{,}198{,}597-18{,}827^2}=9{,}526.$$

We can also use columns 5 and 6 to construct a credibility interval. Discarding the first five rows and the last four rows eliminates 0.0406 of posterior probability. That leaves (5,992, 34,961) as a 96% credibility interval for the layer average severity. Part of Meyers' paper was the observation that even with a fairly large sample the accuracy of the estimate is poor.

The discrete approximation to the prior distribution could be refined by using many more than 16 values. This adds little to the spreadsheet effort. The number was kept small here only for display purposes. □

12.4.5 Exercises

12.67 Show that, if Y is the predictive distribution in Example 12.28, then $\ln Y - \ln 100$ has the Pareto distribution.

12.68 Determine the posterior distribution of α in Example 12.28 if the prior distribution is an arbitrary gamma distribution. To avoid confusion, denote the first parameter of this gamma distribution by γ. Next determine a particular combination of gamma parameters so that the posterior mean is the maximum likelihood estimate of α regardless of the specific values of $x_1,\ldots,x_n$. Is this prior improper?

12.69 For Example 12.51 demonstrate that the maximum likelihood estimate of α is 1.75.

12.70 Let $x_1,\ldots,x_n$ be a random sample from a lognormal distribution with unknown parameters μ and σ. Let the prior density be $\pi(\mu,\sigma)=\sigma^{-1}$.

(a) Write the posterior pdf of μ and σ up to a constant of proportionality.

(b) Determine Bayesian estimators of μ and σ by using the posterior mode.

(c) Fix σ at the posterior mode as determined in part (b) and then determine the exact (conditional) pdf of μ. Then use it to determine a 95% HPD credibility interval for μ.

12.71 A random sample of size 100 has been taken from a gamma distribution with α known to be 2, but θ unknown. For this sample, $\sum_{j=1}^{100} x_j = 30{,}000$. The prior distribution for θ is inverse gamma with β taking the role of α and λ taking the role of θ.

(a) Determine the exact posterior distribution of θ. At this point the values of β and λ have yet to be specified.

(b) The population mean is 2θ. Determine the posterior mean of 2θ using the prior distribution first with $\beta=\lambda=0$ [this is equivalent to $\pi(\theta)=\theta^{-1}$] and then with $\beta=2$ and $\lambda=250$ (which is a prior mean of 250). Then, in each case, determine a 95% credibility interval with 2.5% probability on each side.

(c) Determine the posterior variance of 2θ and use the Bayesian central limit theorem to construct a 95% credibility interval for 2θ using each of the two prior distributions given in part (b).

(d) Determine the maximum likelihood estimate of θ and then use the estimated variance to construct a 95% confidence interval for 2θ.

12.72 Suppose that given $\Theta = \theta$ the random variables $X_1, \ldots, X_n$ are independent and binomially distributed with pf

$$f_{X_j|\Theta}(x_j|\theta) = \binom{K_j}{x_j}\theta^{x_j}(1-\theta)^{K_j - x_j}, \quad x_j = 0, 1, \ldots, K_j,$$

and Θ itself is beta distributed with parameters a and b and pdf

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1.$$

(a) Verify that the marginal pf of X_j is

$$f_{X_j}(x_j) = \frac{\binom{-a}{x_j}\binom{-b}{K_j - x_j}}{\binom{-a-b}{K_j}}, \quad x_j = 0, 1, \ldots, K_j,$$

and $\mathrm{E}(X_j) = aK_j/(a+b)$. This distribution is termed the binomial–beta or negative hypergeometric distribution.

(b) Determine the posterior pdf $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ and the posterior mean $\mathrm{E}(\Theta|\mathbf{x})$.

12.73 Suppose that given $\Theta = \theta$ the random variables $X_1, \ldots, X_n$ are independent and identically exponentially distributed with pdf

$$f_{X_j|\Theta}(x_j|\theta) = \theta e^{-\theta x_j}, \quad x_j > 0,$$

and Θ is itself gamma distributed with parameters $\alpha > 1$ and $\beta > 0$,

$$\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \quad \theta > 0.$$

(a) Verify that the marginal pdf of X_j is

$$f_{X_j}(x_j) = \alpha\beta^{-\alpha}(\beta^{-1} + x_j)^{-\alpha-1}, \quad x_j > 0,$$

and

$$\mathrm{E}(X_j) = \frac{1}{\beta(\alpha-1)}.$$

This distribution is one form of the Pareto distribution.

(b) Determine the posterior pdf $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ and the posterior mean $\mathrm{E}(\Theta|\mathbf{x})$.

12.74 Suppose that given $\Theta = \theta$ the random variables $X_1, \ldots, X_n$ are independent and identically negative binomially distributed with parameters r and θ with pf

$$f_{X_j|\Theta}(x_j|\theta) = \binom{r + x_j - 1}{x_j}\theta^{r}(1-\theta)^{x_j}, \quad x_j = 0, 1, 2, \ldots,$$

and Θ itself is beta distributed with parameters a and b and pdf

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1.$$

(a) Verify that the marginal pf of X_j is

$$f_{X_j}(x_j) = \frac{\Gamma(r+x_j)}{\Gamma(r)x_j!}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+r)\Gamma(b+x_j)}{\Gamma(a+r+b+x_j)}, \quad x_j = 0, 1, 2, \ldots,$$

and

$$\mathrm{E}(X_j) = \frac{rb}{a-1}.$$

This distribution is termed the **generalized Waring distribution**. The special case where $b = 1$ is the **Waring distribution** and the **Yule distribution** if $r = 1$ and $b = 1$.

(b) Determine the posterior pdf $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ and the posterior mean $\mathrm{E}(\Theta|\mathbf{x})$.

12.75 Suppose that given $\Theta = \theta$ the random variables $X_1, \ldots, X_n$ are independent and identically normally distributed with mean μ and variance θ^{-1} and Θ is gamma distributed with parameters α and (θ replaced by) $1/\beta$.

(a) Verify that the marginal pdf of X_j is

$$f_{X_j}(x_j) = \frac{\Gamma(\alpha + \frac{1}{2})}{\sqrt{2\pi\beta}\,\Gamma(\alpha)}\left[1 + \frac{1}{2\beta}(x_j - \mu)^2\right]^{-\alpha-1/2}, \quad -\infty < x_j < \infty,$$

which is a form of the t-distribution.

(b) Determine the posterior pdf $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ and the posterior mean $\mathrm{E}(\theta|\mathbf{x})$.

12.76 Prove that the binomial distribution with pf

$$f(x;p) = \binom{n}{x}p^{x}(1-p)^{n-x}$$

is of the form (12.10) and identify θ, $p(x)$, and $q(\theta)$.

12.77 Consider the negative binomial distribution with pf

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)x!}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{x}.$$

If α is fixed, show that $f(x;\alpha,\beta)$ is of the form (12.10) and identify θ, $p(x)$, and $q(\theta)$.

12.78 Suppose $X_1, \ldots, X_n$ are independent and identically distributed with distribution (12.10). Prove that the maximum likelihood estimate of the mean is the sample mean. In other words, if $\hat{\theta}$ is the maximum likelihood estimator of θ, prove that

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

12.79 Consider the generalization of (12.10) given by

$$f(x;\theta) = \frac{p(m,x)e^{-m\theta x}}{[q(\theta)]^m},$$

where m is a known parameter. Prove that the mean is still given by (12.12) but the variance is given by $v(\theta)/m$, where $v(\theta)$ is given by (12.13).

12.80 Let $X_1, \ldots, X_n$ be independent and identically distributed conditional on Θ with pf

$$f_{X_j|\Theta}(x_j|\theta) = \frac{p(x_j)e^{-\theta x_j}}{q(\theta)}.$$

Let $S = X_1 + \cdots + X_n$.

(a) Show that, conditional on Θ, S has pf of the form

$$f_{S|\Theta}(s|\theta) = \frac{p_n(s)e^{-\theta s}}{[q(\theta)]^n},$$

where $p_n(s)$ does not depend on θ.

(b) Prove that the posterior distribution $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ is the same as the (conditional) distribution of $\Theta|S$,

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{S|\Theta}(s|\theta)\pi(\theta)}{f_S(s)},$$

where $\pi(\theta)$ is the pf of Θ and $f_S(s)$ is the marginal pf of S.

12.81 Suppose that given N the random variable X is binomially distributed with parameters N and p.

(a) Show that, if N is Poisson distributed, so is X (unconditionally) and identify the parameters.

(b) Show that, if N is binomially distributed, so is X (unconditionally) and identify the parameters.

(c) Show that, if N is negative binomially distributed, so is X (unconditionally) and identify the parameters.

12.82 (*) A die is selected at random from an urn that contains two six-sided dice. Die number 1 has three faces with the number 2 while one face each has the numbers 1, 3, and 4. Die number 2 has three faces with the number 4 while one face each has the numbers 1, 2, and 3. The first five rolls of the die yielded the numbers 2, 3, 4, 1, and 4 in that order. Determine the probability that the selected die was die number 2.

12.83 (*) The number of claims in a year, Y, has a distribution which depends on a parameter θ. As a random variable, Θ has the uniform distribution on the interval $(0,1)$. The unconditional probability that Y is 0 is greater than 0.35. For each conditional pf given below, determine if it is possible that it is the true conditional pf of Y.

(a) $\Pr(Y = y|\theta) = e^{-\theta}\theta^y/y!$.

(b) $\Pr(Y = y|\theta) = (y+1)\theta^2(1-\theta)^y$.

(c) $\Pr(Y = y|\theta) = \binom{2}{y}\theta^y(1-\theta)^{2-y}$.

12.84 (*) Your prior distribution concerning the unknown value of H is $\Pr(H = \frac{1}{4}) = \frac{4}{5}$ and $\Pr(H = \frac{1}{2}) = \frac{1}{5}$. The observation from a single experiment has distribution $\Pr(D = d|H = h) = h^d(1-h)^{1-d}$ for $d = 0, 1$. The result of a single experiment is $d = 1$. Determine the posterior distribution of H.

12.85 (*) The number of claims in one year, Y, has the Poisson distribution with parameter θ. The parameter θ has the exponential distribution with pdf $\pi(\theta) = e^{-\theta}$. A particular insured had no claims in one year. Determine the posterior distribution of θ for this insured.

12.86 (*) The number of claims in one year, Y, has the Poisson distribution with parameter θ. The prior distribution has the gamma distribution with pdf $\pi(\theta) = \theta e^{-\theta}$. There was one claim in one year. Determine the posterior pdf of θ.

12.87 (*) Each individual car's claim count has a Poisson distribution with parameter λ. All individual cars have the same parameter. The prior distribution is gamma with parameters $\alpha = 50$ and $\theta = 1/500$. In a two-year period, the insurer covers 750 and 1,100 cars in years 1 and 2, respectively. There were 65 and 112 claims in years one and two, respectively. Determine the coefficient of variation of the posterior gamma distribution.

12.88 (*) The number of claims, r, made by an individual in one year has the binomial distribution with pf $f(r) = \binom{3}{r}\theta^r(1-\theta)^{3-r}$. The prior distribution for θ has pdf $\pi(\theta) = 6(\theta - \theta^2)$. There was one claim in a one-year period. Determine the posterior pdf of θ.

12.89 (*) The number of claims for an individual in one year has a Poisson distribution with parameter λ. The prior distribution for λ has the gamma distribution with mean 0.14 and variance 0.0004. During the past two years a

total of 110 claims has been observed. In each year there were 310 policies in force. Determine the expected value and variance of the posterior distribution of λ.

12.90 (*) The number of claims for an individual in one year has a Poisson distribution with parameter λ. The prior distribution for λ is exponential with an expected value of 2. There were three claims in the first year. Determine the posterior distribution of λ.

12.91 (*) The number of claims in one year has the binomial distribution with $n = 3$ and θ unknown. The prior distribution for θ is beta with pdf $\pi(\theta) = 280\theta^3(1-\theta)^4$, $0 < \theta < 1$. Two claims were observed. Determine each of the following.

(a) The posterior distribution of θ.

(b) The expected value of θ from the posterior distribution.

12.92 (*) An individual risk has exactly one claim each year. The amount of the single claim has an exponential distribution with pdf $f(x) = te^{-tx}$, $x > 0$. The parameter t has a prior distribution with pdf $\pi(t) = te^{-t}$. A claim of 5 has been observed. Determine the posterior pdf of t.

12.93 Suppose that given $\Theta_1 = \theta_1$ and $\Theta_2 = \theta_2$ the random variables $X_1, \ldots, X_n$ are independent and identically normally distributed with mean θ_1 and variance θ_2^{-1}. Suppose also that the conditional distribution of Θ_1 given $\Theta_2 = \theta_2$ is a normal distribution with mean μ and variance σ^2/θ_2 and Θ_2 is gamma distributed with parameters α and $\theta = 1/\beta$.

(a) Show that the posterior conditional distribution of Θ_1 given $\Theta_2 = \theta_2$ is normally distributed with mean

$$\mu_* = \frac{1}{1+n\sigma^2}\mu + \frac{n\sigma^2}{1+n\sigma^2}\overline{x}$$

and variance

$$\sigma_*^2 = \frac{\sigma^2}{\theta_2(1+n\sigma^2)}$$

and the posterior marginal distribution of Θ_2 is gamma distributed with parameters

$$\alpha_* = \alpha + \frac{n}{2}$$

and

$$\beta_* = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{n(\overline{x}-\mu)^2}{2(1+n\sigma^2)}.$$

(b) Find the posterior marginal means $\mathrm{E}(\Theta_1|\mathbf{x})$ and $\mathrm{E}(\Theta_2|\mathbf{x})$.

Table 12.7 Number of hospital liability claims by year

Year	Number of claims
1985	6
1986	2
1987	3
1988	0
1989	2
1990	1
1991	2
1992	5
1993	1
1994	3

Table 12.8 Hospital liability claims by frequency

Frequency (k)	Number of observations (n_k)
0	1
1	2
2	3
3	2
4	0
5	1
6	1
7+	0

12.5 ESTIMATION FOR DISCRETE DISTRIBUTIONS

12.5.1 Poisson

The principles of estimation discussed earlier in this chapter for continuous models can be applied equally to frequency distributions. We will now illustrate the methods of estimation by fitting a Poisson model.

Example 12.52 *A hospital liability policy has experienced the number of claims over a 10-year period given in Table 12.7. Estimate the Poisson parameter using the method of moments and the method of maximum likelihood.*

These data can be summarized in a different way. We can count the number of years in which exactly zero claims occurred, one claim occurred, and so on, as in Table 12.8.

The total number of claims for the period 1985–1994 is 25. Hence, the average number of claims per year is 2.5. The average can also be computed

from Table 12.8. Let n_k denote the number of years in which a frequency of exactly k claims occurred. The expected frequency (sample mean) is

$$\bar{x} = \frac{\sum_{k=0}^{\infty} k n_k}{\sum_{k=0}^{\infty} n_k},$$

where n_k represents the number of observed values at frequency k. Hence the method-of-moments estimate of the Poisson parameter is $\hat{\lambda} = 2.5$.

Maximum likelihood estimation can easily be carried out on these data. The likelihood contribution of an observation of k is p_k. Then the likelihood for the entire set of observations is

$$L = \prod_{k=0}^{\infty} p_k^{n_k}$$

and the loglikelihood is

$$l = \sum_{k=0}^{\infty} n_k \ln p_k.$$

The likelihood and loglikelihood functions are considered to be functions of the unknown parameters. In the case of the Poisson distribution, there is only one parameter, making the maximization easy.

For the Poisson distribution,

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}$$

and

$$\ln p_k = -\lambda + k \ln \lambda - \ln k!.$$

The loglikelihood is

$$\begin{aligned} l &= \sum_{k=0}^{\infty} n_k(-\lambda + k \ln \lambda - \ln k!) \\ &= -\lambda n + \sum_{k=0}^{\infty} k\, n_k \ln \lambda - \sum_{k=0}^{\infty} n_k \ln k!, \end{aligned}$$

where $n = \sum_{k=0}^{\infty} n_k$ is the sample size. Differentiating the loglikelihood with respect to λ, we obtain

$$\frac{dl}{d\lambda} = -n + \sum_{k=0}^{\infty} k\, n_k \frac{1}{\lambda}.$$

By setting the derivative of the loglikelihood to zero, the maximum likelihood estimate is obtained as the solution of the resulting equation. The estimator is then

$$\hat{\lambda} = \frac{\sum_{k=0}^{\infty} k n_k}{n} = \bar{x}.$$

From this it can be seen that for the Poisson distribution the maximum likelihood and the method-of-moments estimators are identical.

If N has a Poisson distribution with mean λ, then

$$\mathrm{E}(\hat{\lambda}) = \mathrm{E}(N) = \lambda$$

and

$$\mathrm{Var}(\hat{\lambda}) = \frac{\mathrm{Var}(N)}{n} = \frac{\lambda}{n}.$$

Hence, $\hat{\lambda}$ is unbiased and consistent. From Theorem 12.13, the maximum likelihood estimator is asymptotically normally distributed with mean λ and variance

$$\begin{aligned} \mathrm{Var}(\hat{\lambda}) &= \left\{-n\mathrm{E}\left[\frac{d^2}{d\lambda^2}\ln p_N\right]\right\}^{-1} \\ &= \left\{-n\mathrm{E}\left[\frac{d^2}{d\lambda^2}(-\lambda + N\ln\lambda - \ln N!)\right]\right\}^{-1} \\ &= \left[n\mathrm{E}(N/\lambda^2)\right]^{-1} \\ &= \left(n\lambda^{-1}\right)^{-1} = \frac{\lambda}{n}. \end{aligned}$$

In this case the asymptotic approximation to the variance is equal to its true value. From this information, we can construct an approximate 95% confidence interval for the true value of the parameter. The interval is $\hat{\lambda} \pm 1.96(\hat{\lambda}/n)^{1/2}$. For this example, the interval becomes (1.52, 3.48). This confidence interval is only an approximation because it relies on large sample theory. The sample size is very small and such a confidence interval should be used with caution. □

The formulas presented so far have assumed that the counts at each observed frequency are known. Occasionally, data are collected so that this is not given. The most common example is to have a final entry given as $k+$, where the count is the number of times k or more claims were observed. If n_{k+} is the number of times this was observed, the contribution to the likelihood function is

$$(p_k + p_{k+1} + \cdots)^{n_{k+}} = (1 - p_0 - \cdots - p_{k-1})^{n_{k+}}.$$

The same adjustments apply to grouped frequency data of any kind. Suppose there were five observations at frequencies 3–5. The contribution to the likelihood function is

$$(p_3 + p_4 + p_5)^5.$$

Table 12.9 Data for Example 12.53

No. of claims/day	Observed no. of days
0	47
1	97
2	109
3	62
4	25
5	16
6+	9

Example 12.53 *For the data in Table* 12.9[11] *determine the maximum likelihood estimate for the Poisson distribution.*

The likelihood function is

$$L = p_0^{47} p_1^{97} p_2^{109} p_3^{62} p_4^{25} p_5^{16} (1 - p_0 - p_1 - p_2 - p_3 - p_4 - p_5)^9,$$

and when written as a function of λ, it becomes somewhat complicated. While the derivative can be taken, solving the equation when it is set equal to zero will require numerical methods. It may be just as easy to use a numerical method to directly maximize the function. A reasonable starting value can be obtained by assuming that all nine observations were exactly at 6 and then using the sample mean. Of course, this will understate the true maximum likelihood estimate, but should be a good place to start. For this particular example, the maximum likelihood estimate is $\hat{\lambda} = 2.0226$, which is very close to the value obtained when all the counts were recorded. □

12.5.2 Negative binomial

The moment equations are

$$r\beta = \frac{\sum_{k=0}^{\infty} k n_k}{n} = \bar{x} \tag{12.14}$$

and

$$r\beta(1+\beta) = \frac{\sum_{k=0}^{\infty} k^2 n_k}{n} - \left(\frac{\sum_{k=0}^{\infty} k n_k}{n}\right)^2 = s^2 \tag{12.15}$$

with solutions $\hat{\beta} = (s^2/\bar{x}) - 1$ and $\hat{r} = \bar{x}/\hat{\beta}$. Note that this variance estimate is obtained by dividing by n, not $n-1$. This is a common, though not required,

[11] This is the same data as will be analyzed in Example 13.15 except the observations at 6 or more have been combined.

approach when using the method of moments. Also note that, if $s^2 < \bar{x}$, the estimate of β will be negative, an inadmissible value.

Example 12.54 (Example 12.52 continued) *Estimate the negative binomial parameters by the method of moments.*

The sample mean and the sample variance are 2.5 and 3.05 (verify this), respectively, and the estimates of the parameters are $\hat{r} = 11.364$ and $\hat{\beta} = 0.22$. □

When compared to the Poisson distribution with the same mean, it can be seen that β is a measure of "extra-Poisson" variation. A value of $\beta = 0$ means no extra-Poisson variation, while a value of $\beta = 0.22$ implies a 22% increase in the variance when compared to the Poisson distribution with the same mean.

We now examine maximum likelihood estimation. The loglikelihood for the negative binomial distribution is

$$\begin{aligned} l &= \sum_{k=0}^{\infty} n_k \ln p_k \\ &= \sum_{k=0}^{\infty} n_k \left[\ln \binom{r+k-1}{k} - r\ln(1+\beta) + k\ln\beta - k\ln(1+\beta)\right]. \end{aligned}$$

The loglikelihood is a function of the two parameters β and r. In order to find the maximum of the loglikelihood, we differentiate with respect to each of the parameters, set the derivatives equal to zero, and solve for the parameters. The derivatives of the loglikelihood are

$$\frac{\partial l}{\partial \beta} = \sum_{k=0}^{\infty} n_k \left(\frac{k}{\beta} - \frac{r+k}{1+\beta}\right) \tag{12.16}$$

and

$$\begin{aligned} \frac{\partial l}{\partial r} &= -\sum_{k=0}^{\infty} n_k \ln(1+\beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \ln \frac{(r+k-1)\cdots r}{k!} \\ &= -n\ln(1+\beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \ln \prod_{m=0}^{k-1} (r+m) \\ &= -n\ln(1+\beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \sum_{m=0}^{k-1} \ln(r+m) \\ &= -n\ln(1+\beta) + \sum_{k=1}^{\infty} n_k \sum_{m=0}^{k-1} \frac{1}{r+m}. \end{aligned} \tag{12.17}$$

Setting these equations to zero yields

$$\hat{\mu} = \hat{r}\hat{\beta} = \frac{\sum_{k=0}^{\infty} kn_k}{n} = \bar{x} \tag{12.18}$$

and

$$n\ln(1+\hat{\beta}) = \sum_{k=1}^{\infty} n_k \left(\sum_{m=0}^{k-1} \frac{1}{\hat{r}+m} \right). \tag{12.19}$$

Note that the maximum likelihood estimator of the mean is the sample mean (as, by definition, in the method of moments). Equations (12.18) and (12.19) can be solved numerically. Replacing $\hat{\beta}$ in (12.19) by $\hat{\mu}/\hat{r}$ yields the equation

$$H(\hat{r}) = n\ln\left(1+\frac{\bar{x}}{\hat{r}}\right) - \sum_{k=1}^{\infty} n_k \left(\sum_{m=0}^{k-1} \frac{1}{\hat{r}+m} \right) = 0. \tag{12.20}$$

If the right-hand side of (12.15) is greater than the right-hand side of (12.14), it can be shown that there is a unique solution of (12.20). If not, then the negative binomial model is probably not a good model to use because the sample variance does not exceed the sample mean.[12]

Equation (12.20) can be solved numerically for $\hat{r}$ using the Newton–Raphson method. The required equation for the kth iteration is

$$r_k = r_{k-1} - \frac{H(r_{k-1})}{H'(r_{k-1})}.$$

A useful starting value for r_0 is the moment-based estimator of r. Of course, any numerical root-finding method (e.g., bisection, secant) may be used.

The loglikelihood is a function of two variables. It can be maximized directly using methods like those described in Appendix F. For the case of the negative binomial distribution with complete data, because we know the estimator of the mean must be the sample mean, setting $\beta = \bar{x}/r$ reduces this to a one-dimensional problem.

Example 12.55 *Determine the maximum likelihood estimates of the negative binomial parameters for the data in Example* 12.52.

The maximum occurs at $\hat{r} = 10.9650$ and $\hat{\beta} = 0.227998$. □

Example 12.56 *Tröbliger* [130] *studied the driving habits of* 23,589 *automobile drivers in a class of automobile insurance by counting the number of accidents per driver in a one-year time period. The data as well as fitted Poisson and negative binomial distributions are given in Table* 12.10. *Based on the information presented, which distribution appears to provide a better model?*

[12] In other words, when the sample variance is less than or equal to the mean, the loglikelihood function will not have a maximum. The function will keep increasing as r goes to infinity and β goes to zero with the product remaining constant. This effectively says that the negative binomial distribution that best matches the data is the Poisson distribution that is a limiting case.

Table 12.10 Two models for automobile claims frequency

No. of claims/year	No. of drivers	Poisson expected	Negative binomial expected
0	20,592	20,420.9	20,596.8
1	2,651	2,945.1	2,631.0
2	297	212.4	318.4
3	41	10.2	37.8
4	7	0.4	4.4
5	0	0.0	0.5
6	1	0.0	0.1
7+	0	0.0	0.0
Parameters		$\lambda = 0.144220$	$r = 1.11790$ $\beta = 0.129010$
Loglikelihood		−10,297.84	−10,223.42

The expected counts are found by multiplying the sample size (23,589) by the probability assigned by the model. It is clear that the negative binomial probabilities produce expected counts that are much closer to those that were observed. In addition, the loglikelihood function is maximized at a significantly higher value. Formal procedures for model selection (including what it means to be *significantly higher*) are discussed in Chapter 13. However, in this case, the superiority of the negative binomial model is apparent. □

12.5.3 Binomial

The binomial distribution has two parameters, m and q. Frequently, the value of m is known and fixed. In this case, only one parameter, q, needs to be estimated. In many insurance situations, q is interpreted as the probability of some event such as death or disability. In such cases the value of q is usually estimated as

$$\hat{q} = \frac{\text{number of observed events}}{\text{maximum number of possible events}},$$

which is the method-of-moments estimator when m is known.

In situations where frequency data are in the form of the previous examples in this chapter, the value of the parameter m, the largest possible observation, may be known and fixed or unknown. In any case, m must be no smaller than

the largest observation. The loglikelihood is

$$\begin{aligned} l &= \sum_{k=0}^{m} n_k \ln p_k \\ &= \sum_{k=0}^{m} n_k \left[\ln \binom{m}{k} + k \ln q + (m-k) \ln(1-q) \right]. \end{aligned}$$

When m is known and fixed, one needs only maximize l with respect to q.

$$\frac{\partial l}{\partial q} = \frac{1}{q} \sum_{k=0}^{m} k n_k - \frac{1}{1-q} \sum_{k=0}^{m} (m-k) n_k.$$

Setting this equal to zero yields

$$\hat{q} = \frac{1}{m} \frac{\sum_{k=0}^{m} k n_k}{\sum_{k=0}^{m} n_k},$$

which is the sample proportion of observed events. For the method of moments, with m fixed, the estimator of q is the same as the maximum likelihood estimator because the moment equation is

$$mq = \frac{\sum_{k=0}^{m} k n_k}{\sum_{k=0}^{m} n_k}.$$

When m is unknown, the maximum likelihood estimator of q is

$$\hat{q} = \frac{1}{\hat{m}} \frac{\sum_{k=0}^{\infty} k n_k}{\sum_{k=0}^{\infty} n_k}, \tag{12.21}$$

where $\hat{m}$ is the maximum likelihood estimate of m. An easy way to approach the maximum likelihood estimation of m and q is to create a *likelihood profile* for various possible values of m as follows:

Step 1: Start with $\hat{m}$ equal to the largest observation.
Step 2: Obtain $\hat{q}$ using (12.21).
Step 3: Calculate the loglikelihood at these values.
Step 4: Increase $\hat{m}$ by 1.
Step 5: Repeat steps 2–4 until a maximum is found.

As with the negative binomial, there need not be a pair of parameters that maximizes the likelihood function. In particular, if the sample mean is less than or equal to the sample variance, the procedure above will lead to ever increasing loglikelihood values as the value of $\hat{m}$ is increased. Once again, the trend is toward a Poisson model. This can be checked out using the data from Example 12.52.

Table 12.11 Number of claims per policy

No. of claims/policy	No. of policies
0	5,367
1	5,893
2	2,870
3	842
4	163
5	23
6	1
7	1
8+	0

Example 12.57 *The number of claims per policy during a one-year period for a block of* 15,160 *insurance policies are given in Table* 12.11. *Obtain moment-based and maximum likelihood estimators.*

The sample mean and variance are 0.985422 and 0.890355, respectively. The variance is smaller than the mean, suggesting the binomial as a reasonable distribution to try. The method of moments leads to

$$mq = 0.985422$$

and

$$mq(1-q) = 0.890355.$$

Hence, $\hat{q} = 0.096474$ and $\hat{m} = 10.21440$. However, m can only take on integer values. We choose $\hat{m} = 10$ by rounding. Then we adjust the estimate of $\hat{q}$ to 0.0985422 from the first moment equation. Doing this will result in a model variance which differs from the sample variance because $10(0.0985422)(1 - 0.0985422) = 0.888316$. This shows one of the pitfalls of using the method of moments with integer-valued parameters.

We now turn to maximum likelihood estimation. From the data $m \geq 7$. If m is known, then only q needs to be estimated. If m is unknown, then we can produce a likelihood profile by maximizing the likelihood for fixed values of m starting at 7 and increasing until a maximum is found. The results are in Table 12.12.

The largest loglikelihood value occurs at $m = 10$. If, a priori, the value of m is unknown, then the maximum likelihood estimates of the parameters are $\hat{m} = 10$ and $\hat{q} = 0.0985422$. This is the same as the adjusted moment estimates. This is not necessarily the case for all data sets. □

Table 12.12 Binomial likelihood profile

$\hat{m}$	$\hat{q}$	−Loglikelihood
7	0.140775	19,273.56
8	0.123178	19,265.37
9	0.109491	19,262.02
10	0.098542	19,260.98
11	0.089584	19,261.11
12	0.082119	19,261.84

12.5.4 The $(a, b, 1)$ class

Estimation of the parameters for the $(a, b, 1)$ class follows the same general principles that were used in connection with the $(a, b, 0)$ class.

Assuming that the data are in the same form as the previous examples, the likelihood is, using (4.13),

$$L = \left(p_0^M\right)^{n_0} \prod_{k=1}^{\infty} \left(p_k^M\right)^{n_k} = \left(p_0^M\right)^{n_0} \prod_{k=1}^{\infty} \left[(1-p_0^M)p_k^T\right]^{n_k}.$$

The loglikelihood is,

$$\begin{aligned} l &= n_0 \ln p_0^M + \sum_{k=1}^{\infty} n_k [\ln(1-p_0^M) + \ln p_k^T] \\ &= n_0 \ln p_0^M + \sum_{k=1}^{\infty} n_k \ln(1-p_0^M) + \sum_{k=1}^{\infty} n_k [\ln p_k - \ln(1-p_0)], \end{aligned}$$

where the last statement follows from $p_k^T = p_k/(1-p_0)$. The three parameters of the $(a, b, 1)$ class are p_0^M, a, and b, where a and b determine $p_1, p_2, \ldots$.

Then it can be seen that

$$l = l_0 + l_1$$

with

$$\begin{aligned} l_0 &= n_0 \ln p_0^M + \sum_{k=1}^{\infty} n_k \ln(1-p_0^M), \\ l_1 &= \sum_{k=1}^{\infty} n_k [\ln p_k - \ln(1-p_0)], \end{aligned}$$

where l_0 depends only on the parameter p_0^M and l_1 is independent of p_0^M, depending only on a and b. This simplifies the maximization because

$$\frac{\partial l}{\partial p_0^M} = \frac{\partial l_0}{\partial p_0^M} = \frac{n_0}{p_0^M} - \sum_{k=1}^{\infty} \frac{n_k}{1-p_0^M} = \frac{n_0}{p_0^M} - \frac{n-n_0}{1-p_0^M},$$

resulting in

$$\hat{p}_0^M = \frac{n_0}{n},$$

the proportion of observations at zero. This is the natural estimator because p_0^M represents the probability of an observation of zero.

Similarly, because the likelihood factors conveniently, the estimation of a and b is independent of p_0^M. Note that although a and b are parameters maximization should not be done with respect to them. That is because not all values of a and b produce admissible probability distributions.[13] For the zero-modified Poisson distribution, the relevant part of the loglikelihood is

$$\begin{aligned} l_1 &= \sum_{k=1}^{\infty} n_k \left[\ln \frac{e^{-\lambda}\lambda^k}{k!} - \ln(1-e^{-\lambda})\right] \\ &= -(n-n_0)\lambda + \left(\sum_{k=1}^{\infty} k\, n_k\right) \ln\lambda - (n-n_0)\ln(1-e^{-\lambda}) + c \\ &= -(n-n_0)[\lambda + \ln(1-e^{-\lambda})] + n\bar{x}\ln\lambda + c, \end{aligned}$$

where $\bar{x} = \frac{1}{n}\sum_{k=0}^{\infty} k n_k$ is the sample mean, $n = \sum_{k=0}^{\infty} n_k$, and c is independent of λ. Hence,

$$\begin{aligned} \frac{\partial l_1}{\partial \lambda} &= -(n-n_0) - (n-n_0)\frac{e^{-\lambda}}{1-e^{-\lambda}} + n\frac{\bar{x}}{\lambda} \\ &= -\frac{n-n_0}{1-e^{-\lambda}} + \frac{n\bar{x}}{\lambda}. \end{aligned}$$

Setting this to zero yields

$$\bar{x}(1-e^{-\lambda}) = \frac{n-n_0}{n}\lambda. \tag{12.22}$$

By graphing each side as a function of λ, it is clear that, if $n_0 > 0$, there exist exactly two roots: one is $\lambda = 0$, the other is $\lambda > 0$. Equation (12.22) can be solved numerically to obtain $\hat{\lambda}$. Note that, because $\hat{p}_0^M = n_0/n$ and $p_0 = e^{-\lambda}$, (12.22) can be rewritten as

$$\bar{x} = \frac{1-\hat{p}_0^M}{1-p_0}\lambda. \tag{12.23}$$

Because the right-hand side of (12.23) is the theoretical mean of the zero-modified Poisson distribution (when $\hat{p}_0^M$ is replaced with p_0^M), (12.23) is a

[13] Maximization can be done with respect to any parameterization because maximum likelihood estimation is invariant under parameter transformations. However, it is more difficult to maximize over bounded regions because numerical methods are difficult to constrain and analytic methods will fail due to a lack of differentiability. Therefore, estimation is usually done with respect to particular class members, such as the Poisson.

moment equation. Hence, an alternative estimation method yielding the same results as the maximum likelihood method is to equate p_0^M to the sample proportion at zero and the theoretical mean to the sample mean. This suggests that, by fixing the zero probability to the observed proportion at zero and equating the low order moments, a modified moment method can be used to get starting values for numerical maximization of the likelihood function. Because the maximum likelihood method has better asymptotic properties, it is preferable to use the modified moment method only to obtain starting values.

For the purpose of obtaining estimates of the asymptotic variance of the maximum likelihood estimator of λ, it is easy to obtain

$$\frac{\partial^2 l_1}{\partial \lambda^2} = (n-n_0)\frac{e^{-\lambda}}{(1-e^{-\lambda})^2} - \frac{n\bar{x}}{\lambda^2},$$

and the expected value is obtained by observing that $\mathrm{E}(\bar{x}) = (1-p_0^M)\lambda/(1-e^{-\lambda})$. Finally, p_0^M may be replaced by its estimator, n_0/n. The variance of $\hat{p}_0^M$ is obtained by observing that the numerator, n_0, has a binomial distribution and therefore the variance is $p_0^M(1-p_0^M)/n$.

For the zero-modified binomial distribution,

$$\begin{aligned} l_1 &= \sum_{k=1}^{m} n_k \left\{ \ln\left[\binom{m}{k} q^k (1-q)^{m-k}\right] - \ln[1-(1-q)^m] \right\} \\ &= \left(\sum_{k=1}^{m} k n_k\right) \ln q + \sum_{k=1}^{m} (m-k) n_k \ln(1-q) \\ &\quad - \sum_{k=1}^{m} n_k \ln[1-(1-q)^m] + c \\ &= n\bar{x} \ln q + m(n-n_0)\ln(1-q) - n\bar{x}\ln(1-q) \\ &\quad -(n-n_0)\ln[1-(1-q)^m] + c \end{aligned}$$

where c does not depend on q and

$$\frac{\partial l_1}{\partial q} = \frac{n\bar{x}}{q} - \frac{m(n-n_0)}{1-q} + \frac{n\bar{x}}{1-q} - \frac{(n-n_0)m(1-q)^{m-1}}{1-(1-q)^m}.$$

Setting this to zero yields

$$\bar{x} = \frac{1-\hat{p}_0^M}{1-p_0} mq, \tag{12.24}$$

where we recall that $p_0 = (1-q)^m$. This equation matches the theoretical mean with the sample mean.

If m is known and fixed, the maximum likelihood estimator of p_0^M is still

$$\hat{p}_0^M = \frac{n_0}{n}.$$

However, even with m known, (12.24) must be solved numerically for q. When m is unknown and also needs to be estimated, the above procedure can be followed for different values of m until the maximum of the likelihood function is obtained.

The zero-modified negative binomial (or extended truncated negative binomial) distribution is a bit more complicated because three parameters need to be estimated. Of course, the maximum likelihood estimator of p_0^M is $\hat{p}_0^M = n_0/n$ as before, reducing the problem to the estimation of r and β. The part of the loglikelihood relevant to r and β is

$$l_1 = \sum_{k=1}^{\infty} n_k \ln p_k - (n-n_0)\ln(1-p_0). \tag{12.25}$$

Hence

$$\begin{aligned} l_1 &= \sum_{k=1}^{\infty} n_k \ln\left[\binom{k+r-1}{k}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^k\right] \\ &\quad -(n-n_0)\ln\left[1-\left(\frac{1}{1+\beta}\right)^r\right]. \end{aligned} \tag{12.26}$$

This function needs to be maximized over the (r,β) plane to obtain the maximum likelihood estimates. This can be done numerically using maximization procedures such as those described in Appendix F. Starting values can be obtained by the modified moment method by setting $\hat{p}_0^M = n_0/n$ and equating the first two moments of the distribution to the first two sample moments. It is generally easier to use raw moments (moments about the origin) than central moments for this purpose. In practice, it may be more convenient to maximize (12.25) rather than (12.26) because one can take advantage of the recursive scheme

$$p_k = p_{k-1}\left(a + \frac{b}{k}\right)$$

in evaluating (12.25). This makes computer programming a bit easier.

For zero-truncated distributions there is no need to estimate the probability at zero because it is known to be zero. The remaining parameters are estimated using the same formulas developed for the zero-modified distributions.

Example 12.58 *The data set in Table* 12.13 *comes from Beard et al.* [12]. *Determine a model that adequately describes the data.*

When a Poisson distribution is fitted to it, the resulting fit is very poor. There is too much probability for one accident and two little at subsequent values. The geometric distribution is tried as a one-parameter alternative. It has loglikelihood

Table 12.13 Fitted distributions to Beard data

Accidents	Observed	Poisson	Geometric	ZM Poisson	ZM geom.
0	370,412	369,246.9	372,206.5	370,412.0	370,412.0
1	46,545	48,643.6	43,325.8	46,432.1	46,555.2
2	3,935	3,204.1	5,043.2	4,138.6	3,913.6
3	317	140.7	587.0	245.9	329.0
4	28	4.6	68.3	11.0	27.7
5	3	0.1	8.0	0.4	2.3
6+	0	0.0	1.0	0.0	0.2
Parameters		λ: 0.13174	β: 0.13174	p_0^M: 0.87934 λ: 0.17827	p_0^M: 0.87934 β: 0.091780
Loglikelihood		−171,373	−171,479	−171,160	−171,133

$$\begin{aligned} l &= -n\ln(1+\beta) + \sum_{k=1}^{\infty} n_k \ln\left(\frac{\beta}{1+\beta}\right)^k \\ &= -n\ln(1+\beta) + \sum_{k=1}^{\infty} k n_k [\ln\beta - \ln(1+\beta)] \\ &= -n\ln(1+\beta) + n\bar{x}[\ln\beta - \ln(1+\beta)] \\ &= -(n+n\bar{x})\ln(1+\beta) + n\bar{x}\ln\beta, \end{aligned}$$

where $\bar{x} = \sum_{k=1}^{\infty} k\, n_k/n$ and $n = \sum_{k=0}^{\infty} n_k$.

Differentiation reveals that the loglikelihood has a maximum at

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

A qualitative look at the numbers indicates that the zero-modified geometric distribution matches the data better than the other three models considered. A formal analysis is done in Example 13.16. □

12.5.5 Compound models

For the method of moments, the first few moments can be matched with the sample moments. The system of equations can be solved to obtain the moment based estimators. Note that the number of parameters in the compound model is the sum of the number of parameters in the primary and secondary distributions. The first two theoretical moments for compound distributions are

$$\begin{aligned} \mathrm{E}(S) &= \mathrm{E}(N)\mathrm{E}(M) \\ \mathrm{Var}(S) &= \mathrm{E}(N)\,\mathrm{Var}(M) + \mathrm{E}(M)^2\,\mathrm{Var}(N). \end{aligned}$$

These results were developed in Chapter 6. The first three moments for the compound Poisson distribution are given in (4.27).

Maximum likelihood estimation is also carried out as before. The loglikelihood to be maximized is

$$l = \sum_{k=0}^{\infty} n_k \ln g_k.$$

When g_k is the probability of a compound distribution, the loglikelihood can be maximized numerically. The first and second derivatives of the loglikelihood can be obtained by using approximate differentiation methods as applied directly to the loglikelihood function at the maximum value.

Example 12.59 *Determine various properties of the Poisson–zero-truncated geometric distribution. This distribution is also called the Polya–Aeppli distribution.*

For the zero-truncated geometric distribution the pgf is

$$P_2(z) = \frac{[1-\beta(z-1)]^{-1} - (1+\beta)^{-1}}{1-(1+\beta)^{-1}}$$

and therefore the pgf of the Polya–Aeppli distribution is

$$\begin{aligned} P(z) &= P_1[P_2(z)] = \exp\left(\lambda\left\{\frac{[1-\beta(z-1)]^{-1} - (1+\beta)^{-1}}{1-(1+\beta)^{-1}} - 1\right\}\right) \\ &= \exp\left\{\lambda\frac{[1-\beta(z-1)]^{-1} - 1}{1-(1+\beta)^{-1}}\right\}. \end{aligned}$$

The mean is

$$P'(1) = \lambda(1+\beta)$$

and the variance is

$$P''(1) + P'(1) - [P'(1)]^2 = \lambda(1+\beta)(1+2\beta).$$

Alternatively, $\mathrm{E}(N) = \mathrm{Var}(N) = \lambda$, $\mathrm{E}(M) = 1+\beta$, and $\mathrm{Var}(M) = \beta(1+\beta)$. Then,

$$\begin{aligned} \mathrm{E}(S) &= \lambda(1+\beta), \\ \mathrm{Var}(S) &= \lambda\beta(1+\beta) + \lambda(1+\beta)^2 = \lambda(1+\beta)(1+2\beta). \end{aligned}$$

From Theorem 4.51, the probability at zero is

$$g_0 = P_1(0) = e^{-\lambda}.$$

The successive values of g_k are computed easily using the compound Poisson recursion

$$g_k = \frac{\lambda}{k}\sum_{j=1}^{k} j f_j g_{k-j}, \qquad k = 1, 2, 3, \ldots, \tag{12.27}$$

Table 12.14 Automobile claims by year

Year	Exposure	Claims
1986	2,145	207
1987	2,452	227
1988	3,112	341
1989	3,458	335
1990	3,698	362
1991	3,872	359

where $f_j = \beta^{j-1}/(1+\beta)^j$, $j = 1, 2, \ldots$. For any values of λ and β, the loglikelihood function can be easily evaluated. □

Example 13.17 provides a data set for which the Polya–Aeppli distribution is a good choice.

Another useful compound Poisson distribution is the Poisson–extended truncated negative binomial (Poisson–ETNB) distribution. Although it does not matter if the secondary distribution is modified or truncated, we prefer the truncated version here so that the parameter r may be extended.[14] Special cases are: $r = 1$, which is the Poisson–geometric (also called Polya–Aeppli); $r \to 0$, which is the Poisson–logarithmic (negative binomial); and $r = -0.5$, which is called the Poisson–inverse Gaussian. This name is not consistent with the others. Here the inverse Gaussian distribution is a mixing distribution (see Section 4.6.9). Example 13.18 provides a data set for which the Poisson–inverse Gaussian distribution is a good choice.

12.5.6 Effect of exposure on maximum likelihood estimation

In Section 4.6.11 the effect of exposure on discrete distributions was discussed. When aggregate data from a large group of insureds is obtained, maximum likelihood estimation is still possible. The following example illustrates this for the Poisson distribution.

Example 12.60 *Determine the maximum likelihood estimate of the Poisson parameter for the data in Table* 12.14.

Let λ be the Poisson parameter for a single exposure. If year k has e_k exposures, then the number of claims has a Poisson distribution with parameter λe_k. If n_k is the number of claims in year k, the likelihood function is

$$L = \prod_{k=1}^{6} \frac{e^{-\lambda e_k}(\lambda e_k)^{n_k}}{n_k!}.$$

The maximum likelihood estimate is found by

$$l = \ln L = \sum_{k=1}^{6} [-\lambda e_k + n_k \ln(\lambda e_k) - \ln(n_k!)],$$

$$\frac{\partial l}{\partial \lambda} = \sum_{k=1}^{6} \left(-e_k + n_k \lambda^{-1}\right) = 0,$$

$$\hat{\lambda} = \frac{\sum_{k=1}^{6} n_k}{\sum_{k=1}^{6} e_k} = \frac{1{,}831}{18{,}737} = 0.09772.$$

□

In this example the answer is what we expected it to be, the average number of claims per exposure. This technique will work for any distribution in the $(a, b, 0)$[15] and compound classes. But care must be taken in the interpretation of the model. For example, if we use a negative binomial distribution, we are assuming that each exposure unit produces claims according to a negative binomial distribution. This is different from assuming that total claims have a negative binomial distribution because they arise from individuals who each have a Poisson distribution but with different parameters.

12.5.7 Exercises

12.94 Assume that the binomial parameter m is known. Consider the maximum likelihood estimator of q.

(a) Show that the maximum likelihood estimator is unbiased.

(b) Determine the variance of the maximum likelihood estimator.

(c) Show that the asymptotic variance as given in Theorem 12.13 is the same as that developed in part (b).

(d) Determine a simple formula for a confidence interval using (9.4) on page 276 that is based on replacing q with $\hat{q}$ in the variance term.

(e) Determine a more complicated formula for a confidence interval using (9.3) that is not based on such a replacement. This should be done in a manner similar to that used in Example 11.12 on page 309.

[14] This does not contradict Theorem 4.54. When $-1 < r < 0$, it is still the case that changing the probability at zero will not produce new distributions. What is true is that there is no probability at zero which will lead to an ordinary $(a, b, 0)$ negative binomial secondary distribution.

[15] For the binomial distribution, the usual problem that m must be an integer remains.

Table 12.15 Data for Exercise 12.96

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

12.95 Use (12.18) to determine the maximum likelihood estimator of β for the geometric distribution. In addition, determine the variance of the maximum likelihood estimator and verify that it matches the asymptotic variance as given in Theorem 12.13.

12.96 A portfolio of 10,000 risks produced the claim counts in Table 12.15.

(a) Determine the maximum likelihood estimate of λ for a Poisson model and then determine a 95% confidence interval for λ.

(b) Determine the maximum likelihood estimate of β for a geometric model and then determine a 95% confidence interval for β.

(c) Determine the maximum likelihood estimate of r and β for a negative binomial model.

(d) Assume that $m = 4$. Determine the maximum likelihood estimate of q of the binomial model.

(e) Construct 95% confidence intervals for q using the methods developed in parts (d) and (e) of Exercise 12.94.

(f) Determine the maximum likelihood estimate of m and q by constructing a likelihood profile.

12.97 An automobile insurance policy provides benefits for accidents caused by both underinsured and uninsured motorists. Data on 1,000 policies revealed the information in Table 12.16.

(a) Determine the maximum likelihood estimate of λ for a Poisson model for each of the variables N_1 = number of underinsured claims and N_2 = number of uninsured claims.

(b) Assume that N_1 and N_2 are independent. Use Theorem 4.37 on page 74 to determine a model for $N = N_1 + N_2$.

12.98 An alternative method of obtaining a model for N in Exercise 12.97 would be to record the total number of underinsured and uninsured claims for each of the 1,000 policies. Suppose this was done and the results were as in Table 12.17.

Table 12.16 Data for Exercise 12.97

No. of claims	Underinsured	Uninsured
0	901	947
1	92	50
2	5	2
3	1	1
4	1	0
5+	0	0

Table 12.17 Data for Exercise 12.98

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

(a) Determine the maximum likelihood estimate of λ for a Poisson model.

(b) The answer to part (a) matched the answer to part (c) of the previous exercise. Demonstrate that this must always be so.

(c) Determine the maximum likelihood estimate of β for a geometric model.

(d) Determine the maximum likelihood estimate of r and β for a negative binomial model.

(e) Assume that $m = 7$. Determine the maximum likelihood estimate of q of the binomial model.

(f) Determine the maximum likelihood estimates of m and q by constructing a likelihood profile.

12.99 The data in Table 12.18 represent the number of prescriptions filled in one year for a group of elderly members of a group insurance plan.

(a) Determine the maximum likelihood estimate of λ for a Poisson model.

(b) Determine the maximum likelihood estimate of β for a geometric model and then determine a 95% confidence interval for β.

Table 12.18 Data for Exercise 12.99

No. of prescriptions	Frequency	No. of prescriptions	Frequency
0	82	16–20	40
1–3	49	21–25	38
4–6	47	26–35	52
7–10	47	36–	91
11–15	57		

(c) Determine the maximum likelihood estimates of r and β for a negative binomial model.

12.6 BIVARIATE MODELS

12.6.1 Introduction

At times a bivariate distribution with dependent variables is the appropriate model. One such situation is a joint life annuity or insurance. Here the timing of the payments depends on the first or second death of two individuals. Because these individuals are often related (typically spouses), the times of death will be dependent. As another example, in casualty insurances it is common to record the expenses that are directly related to the payment of the loss, referred to as the allocated loss adjustment expenses (ALAE). The loss and the ALAE are usually strongly positively correlated.

There are a variety of sources for bivariate and multivariate models. Among them are the books by Hutchinson and Lai ([64]), Kotz, Balakrishnan, and Johnson ([80]), and Mardia ([89]). However, most of the distributions have marginal distributions that are not of interest for actuarial applications or have the parameters related in an unsuitable manner (for example, a bivariate gamma distribution in which both X and Y must have the same value for α). One exception is the bivariate lognormal distribution for which the logarithms of the two variables have a bivariate normal distribution.

Of more interest and practical value are methods which construct bivariate models from known marginal distributions. For example, suppose it were known that losses have the Pareto distribution and that ALAE have the gamma distribution. Then those parameters could be estimated (and the models themselves determined) from the marginal data. Then they could be combined into a bivariate distribution that introduces a degree of association between the two variables. Among the methods available, the copula has received a lot of attention in the actuarial literature and is the only one that will be covered here.

Table 12.19 Twenty-four losses with ALAE

Loss	ALAE	Loss	ALAE
1,500	301	11,750	2,530
2,000	3,043	12,500	165
2,500	415	14,000	175
2,500	4,940	14,750	28,217
4,500	395	15,000	2,072
5,000	25	17,500	6,328
5,750	34,474	19,833	212
7,000	50	30,000	2,172
7,000	10,593	33,033	7,845
7,500	50	44,887	2,178
9,000	406	62,500	12,251
10,000	1,174	210,000	7,357

12.6.2 Copulas

Copula distributions are created using a function, also called a copula. This function must itself be a legitimate bivariate distribution function over the unit square with uniform marginals. Denote the two marginal distribution functions $F_X(x)$ and $F_Y(y)$ and the copula function $C(u, v)$. The bivariate distribution function created by the three is then

$$F_{X,Y}(x, y) = C[F_X(x), F_Y(y)].$$

A simple but fairly useless example is the copula $C(u, v) = uv$. This creates the bivariate distribution function $F_{X,Y}(x, y) = F_X(x)F_Y(y)$, which is true for independent variables.

A good general introduction is [42] and an introduction for actuaries can be found in [40]. The paper by Frees, Carriere, and Valdez, [39] works with Frank's copula for a study of joint lifetimes. An expanded version of the example presented here can be found in [78]. The last two cited papers show how to write the likelihood function under various truncations and censoring.

Example 12.61 *The loss and ALAE were recorded for each of* 24 *claims* (*Table* 12.19). *Determine a model for the joint distribution using Frank's copula with Pareto distributions for both marginals.*

Frank's copula is (where $\log_\alpha$ means the logarithm base α)

$$C(u, v) = \log_\alpha \left[1 + \frac{(\alpha^u - 1)(\alpha^v - 1)}{\alpha - 1}\right], \tag{12.28}$$

where the parameter α controls the degree of association between the two variables. Values of α less than 1 indicate a positive association, values greater

than 1 indicate an inverse association, and 1 indicates independence. If we let β and θ be the parameters of the marginal Pareto distribution for X (where θ is the scale parameter) and let γ and τ be the parameters for Y (with τ the scale parameter), the bivariate distribution function is

$$F(x,y) = \log_\alpha \left\{ 1 + \frac{[\alpha^{1-(1+x/\theta)^{-\beta}} - 1][\alpha^{1-(1+y/\tau)^{-\gamma}} - 1]}{\alpha - 1} \right\}.$$

Taking partial derivatives with respect to x and y provides the joint density function

$$f(x,y) = \frac{(\alpha-1)\dfrac{\beta\gamma}{\theta\tau}\alpha^{2-(1+x/\theta)^{-\beta}-(1+y/\tau)^{-\gamma}} \times (1+x/\theta)^{-\beta-1}(1+y/\tau)^{-\gamma-1}\ln\alpha}{\{\alpha - 1 + [\alpha^{1-(1+x/\theta)^{-\beta}} - 1][\alpha^{1-(1+y/\tau)^{-\gamma}} - 1]\}^2}.$$

Starting values for the four Pareto parameters were obtained by finding the maximum likelihood estimates for the two marginal distributions. Simplex maximization yields the estimates $\hat{\alpha} = 0.133024$, $\hat{\beta} = 2.59889$, $\hat{\theta} = 36{,}141.4$, $\hat{\gamma} = 0.759943$, and $\hat{\tau} = 803.839$. The positive association is apparent and could be tested. One way is to use the likelihood ratio test discussed in Chapter 13. It turns out that with the small sample size there is not sufficient evidence to be sure there is a positive association. □

A number of results concerning Frank's copula can be found in the paper by Genest [41]. Two are presented here. To simulate an (X, Y) pair, begin by simulating the X value from the marginal distribution. This can be done using the standard inversion technique. Follow this by simulating a value of Y from the conditional distribution of Y given $X = x$. To do this, first note that the distribution function is

$$F_{Y|X}(y|x) = \frac{(\partial/\partial x)F(x,y)}{f_X(x)}.$$

For Frank's copula we have

$$\begin{aligned} \frac{\partial}{\partial x}F(x,y) &= f_X(x)\frac{\partial}{\partial u}C(u,v)|_{u=F_X(x),v=F_Y(y)} \\ &= \frac{f_X(x)\alpha^{F_X(x)}[\alpha^{F_Y(y)} - 1]}{\alpha - 1 + [\alpha^{F_X(x)} - 1][\alpha^{F_Y(y)} - 1]}. \end{aligned}$$

To simulate a conditional value of Y using the inversion method discussed in Chapter 17, obtain a uniform(0,1) random number r and solve the equation

$$\frac{\alpha^{F_X(x)}[\alpha^{F_Y(y)} - 1]}{\alpha - 1 + [\alpha^{F_X(x)} - 1][\alpha^{F_Y(y)} - 1]} = r$$

for $F_Y(y)$ to obtain

$$\alpha^{F_Y(y)} = 1 + \frac{r(\alpha-1)}{\alpha^{F_X(x)}(1-r) + r}$$

or

$$F_Y(y) = \frac{1}{\ln\alpha}\ln\left[1 + \frac{r(\alpha-1)}{\alpha^{F_X(x)}(1-r) + r}\right].$$

The right-hand side is a number and then the distribution function of Y can be inverted to solve for the simulated value.

The regression function can be found from

$$\begin{aligned} \mathrm{E}(Y|X=x) &= \int [1 - F_{Y|x}(y|x)]\, dy \\ &= \int \left\{ 1 - \frac{\alpha^{F_X(x)}[\alpha^{F_Y(y)} - 1]}{\alpha - 1 + [\alpha^{F_X(x)} - 1][\alpha^{F_Y(y)} - 1]} \right\} dy, \end{aligned}$$

but it is likely that the integral will have to be done numerically.

12.6.3 Exercise

12.100 Consider the data set in Table 12.19. Fit a bivariate distribution using Frank's copula where each marginal distribution has the inverse exponential distribution.

12.7 MODELS WITH COVARIATES

12.7.1 Introduction

It may be that the distribution of the random variable of interest depends on certain characteristics of the underlying situation. For example, the distribution of time to death may be related to the individual's age, gender, smoking status, blood pressure, height, and weight. Or, consider the number of automobile accidents a vehicle has in a year. The distribution of this variable might be related to the number of miles it is driven, where it is driven, and various characteristics of the primary driver such as age, gender, marital status, and driving history.

Example 12.62 *Suppose we believe that the distribution of the number of accidents a driver has in a year is related to the driver's age and gender. Provide three approaches to modeling this situation.*

Of course there is no limit to the number of models that could be considered. Three that might be used are given below.

1. Construct a model for each combination of gender and age. Collect data separately for each combination and construct each model separately. Either parametric or data-dependent models could be selected.

2. Construct a single, fully parametric model for this situation. As an example, the number of accidents for a given driver could be assumed to have the Poisson distribution with parameter λ. The value of λ is then assumed to depend on the age x and the gender ($g = 1$ for males, $g = 0$ for females) in some way such as

$$\lambda = (\alpha_0 + \alpha_1 x + \alpha_2 x^2)\beta^g.$$

3. Begin with a model for the density, distribution, or hazard rate function that is similar to a data-dependent model. Then use the age and gender to modify this function. For example, select a survival function $S_0(n)$ and then the survival function for a particular driver might be

$$S(n|x, g) = [S_0(n)]^{(\alpha_0+\alpha_1 x+\alpha_2 x^2)\beta^g}.$$ □

While there is nothing wrong with the first approach, it is not very interesting. It just asks us to repeat the modeling process over and over as we move from one combination to another. The second approach is a single parametric model that can also be analyzed with techniques already discussed, but it is clearly more parsimonious. The third model's hybrid nature implies that additional effort will be needed to implement it.

The third model would be a good choice when there is no obvious distributional model for a given individual. In the case of automobile drivers, the Poisson distribution is a reasonable choice and so the second model may be the best approach. If the variable is time to death, a data-dependent model such as a life table may be appropriate.

The advantage of the second and third approaches over the first one is that for some of the combinations there may be very few observations. In this case, the parsimony afforded by the second and third models may allow the limited information to still be useful. For example, suppose our task was to estimate the 80 entries in a life table running from age 20 through age 99 for four gender/smoker combinations. Using the ideas in model 1 above there are 320 items to estimate. Using the ideas in model 3 there would be 83 items to estimate.[16]

[16]There would be 80 items needed to estimate the survival function for one of the four combinations. The other three combinations each add one more item, the power to which the survival function is to be raised.

12.7.2 Proportional hazards models

A particular model that is relatively easy to work with is the Cox proportional hazards model.

Definition 12.63 *Given a baseline hazard rate function $h_0(t)$ and values $z_1, \ldots, z_p$ associated with a particular individual, the* ***Cox proportional hazards model*** *for that person is given by the hazard rate function*

$$h(x|\mathbf{z}) = h_0(x)c(\beta_1 z_1 + \cdots + \beta_p z_p) = h_0(x)c(\boldsymbol{\beta}^T\mathbf{z}),$$

where $c(y)$ is any function that takes on only positive values, $\mathbf{z} = (z_1, \ldots, z_p)^T$ is a column vector of the z values (called ***covariates****), and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a column vector of coefficients.*

The only function that will be used here is $c(y) = e^y$. One advantage of this function is that it must be positive. The name for this model is fitting because if the ratio of the hazard rate functions for two individuals is taken, the ratio will be constant. That is, one person's hazard rate function is proportional to any other person's hazard rate function. Our goal is to estimate the baseline hazard rate function $h_0(t)$ and the vector of coefficients $\boldsymbol{\beta}$.

Example 12.64 *Suppose the size of a homeowner's fire insurance claim as a percentage of the house's value depends upon the age of the house and the type of construction (wood or brick). Develop a Cox proportional hazards model for this situation. Also, indicate the difference between wood and brick houses of the same age.*

Let $z_1 = age$ (a nonnegative whole number) and $z_2 = 1$ if the construction is wood and $z_2 = 0$ if the construction is brick. Then the hazard rate function for a given house is

$$h(x|z_1, z_2) = h_0(x)e^{\beta_1 z_1+\beta_2 z_2}.$$

One consequence of this model is that, regardless of the age, the effect of switching from brick to wood is the same. For two houses of age z_1 we have

$$h_{wood}(x) = h_0(x)e^{\beta_1 z_1+\beta_2} = h_{brick}(x)e^{\beta_2}.$$

The effect on the survival function is

$$\begin{aligned} S_{wood}(x) &= \exp\left[-\int_0^x h_{wood}(y)dy\right] = \exp\left[-\int_0^x h_{brick}(y)e^{\beta_2}dy\right] \\ &= [S_{brick}(x)]^{\exp(\beta_2)}. \end{aligned}$$ □

The baseline hazard rate function can be estimated using either a parametric model or a data-dependent model. The remainder of the model is

Table 12.20 Fire insurance payments

z_1	z_2	Payment
10	0	70
20	0	22
30	0	90*
40	0	81
50	0	8
10	1	51
20	1	95*
30	1	55
40	1	85*
50	1	93

*The payment was made at the policy limit.

parametric. In the spirit of this text, we will use maximum likelihood for estimation of β_1 and β_2. We will begin with a fully parametric example.

Example 12.65 *For the fire insurance example,* 10 *payments are in Table* 12.20. *All values are expressed as a percentage of the house's value. Estimate the parameters of the Cox proportional hazards model using maximum likelihood and both an exponential and a beta distribution for the baseline hazard rate function. There is no deductible on these policies, but there is a policy limit (which differs by policy).*

In order to construct the likelihood function, we need the density and survival functions. Let $c_j = \exp(\boldsymbol{\beta}^T \mathbf{z})$ be the Cox multiplier for the jth observation. Then, as noted in the previous example, $S_j(x) = S_0(x)^{c_j}$, where $S_0(x)$ is the baseline distribution. The density function is

$$\begin{aligned} f_j(x) &= -S_j'(x) = -c_j S_0(x)^{c_j-1} S_0'(x) \\ &= c_j S_0(x)^{c_j-1} f_0(x). \end{aligned}$$

For the exponential distribution,

$$S_j(x) = [e^{-x/\theta}]^{c_j} = e^{-c_j x/\theta} \text{ and } f_j(x) = \left(\frac{c_j}{\theta}\right) e^{-c_j x/\theta}$$

and for the beta distribution,

$$\begin{aligned} S_j(x) &= [1-\beta(a,b;x)]^{c_j}, \text{ and} \\ f_j(x) &= c_j[1-\beta(a,b;x)]^{c_j-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \end{aligned}$$

where $\beta(a,b;x)$ is the distribution function for a beta distribution with parameters a and b [available in Excel® as BETADIST(x,a,b)]. The gamma function is available in Excel® as EXP(GAMMALN(a)). For policies with payments not at the limit, the contribution to the likelihood function is the density function while for those paid at the limit it is the survival function. In both cases, the likelihood function is sufficiently complex that it is not worth writing out. The parameter estimates for the exponential model are $\hat{\beta}_1 = 0.00319$, $\hat{\beta}_2 = -0.63722$, and $\hat{\theta} = 0.74041$. The value of the logarithm of the likelihood function is -6.1379. For the beta model, the estimates are $\hat{\beta}_1 = -0.00315$, $\hat{\beta}_2 = -0.77847$, $\hat{a} = 1.03706$, and $\hat{b} = 0.81442$. The value of the logarithm of the likelihood function is -4.2155. Using the Schwarz Bayesian criterion (see Section 13.5.3), an improvement of $\ln(10)/2 = 1.1513$ is needed to justify a fourth parameter. The beta model is preferred. If an estimate of the information matrix is desired, the only reasonable strategy is to take numerical derivatives of the loglikelihood. □

An alternative is to construct a data-dependent model for the baseline hazard rate. Let $R(y_j)$ be the set of observations that are in the risk set for uncensored observation y_j.[17] Rather than obtain the true likelihood value, it is easier to obtain what is called the *partial likelihood* value. It is a conditional value. Rather than asking, "What is the probability of observing a value of y_j?" we ask, "Given that it is known there is an uncensored observation of y_j, what is the probability that it was the policy that had that value? Do this conditioned on equalling or exceeding that value." This method allows us to estimate the β coefficients separately from the baseline hazard rate. Notation can become a bit awkward here. Let j^* identify the observation that produced the uncensored observation of y_j. Then the contribution to the likelihood function for that policy is

$$\frac{f_{j^*}(y_j)/S_{j^*}(y_j)}{\sum_{i\in R(y_j)} f_i(y_j)/S_i(y_j)} = \frac{c_{j^*} f_0(y_j)/S_0(y_j)}{\sum_{i\in R(y_j)} c_i f_0(y_j)/S_0(y_j)} = \frac{c_{j^*}}{\sum_{i\in R(y_j)} c_i}.$$

Example 12.66 *Use the partial likelihood to estimate β_1 and β_2.*

The ordered, uncensored values are 8, 22, 51, 55, 70, 81, and 93. The calculation of the contribution to the likelihood function is in Table 12.21.

The product is maximized when $\hat{\beta}_1 = -0.00373$ and $\hat{\beta}_2 = -0.91994$ and the logarithm of the partial likelihood is -11.9889. When β_1 is forced to be zero, the maximum is at $\hat{\beta}_2 = -0.93708$ and the logarithm of the partial likelihood is -11.9968. There is no evidence in this sample that age of the house has an impact when using this model. □

Three issues remain. One is to estimate the baseline hazard rate function, one is to deal with the case where there are multiple observations at the same

[17] Recall from Section 11.1 that $y_1, y_2, \ldots$ represent the ordered, unique values from the set of uncensored observations. The risk set was also defined in that section.

Table 12.21 Fire insurance likelihood

Value	y	c	Contribution to L
8	8	$c_1 = \exp(50\beta_1)$	$\frac{c_1}{c_1+\cdots+c_{10}}$
22	22	$c_2 = \exp(20\beta_1)$	$\frac{c_2}{c_2+\cdots+c_{10}}$
51	51	$c_3 = \exp(10\beta_1+\beta_2)$	$\frac{c_3}{c_3+\cdots+c_{10}}$
55	55	$c_4 = \exp(30\beta_1+\beta_2)$	$\frac{c_4}{c_4+\cdots+c_{10}}$
70	70	$c_5 = \exp(10\beta_1)$	$\frac{c_5}{c_5+\cdots+c_{10}}$
81	81	$c_6 = \exp(40\beta_1)$	$\frac{c_6}{c_6+\cdots+c_{10}}$
85		$c_7 = \exp(40\beta_1+\beta_2)$	
90		$c_8 = \exp(30\beta_1)$	
93	93	$c_9 = \exp(50\beta_1+\beta_2)$	$\frac{c_9}{c_9+c_{10}}$
95		$c_{10} = \exp(20\beta_1+\beta_2)$	

value, and the final one is to estimate the variances of estimators. For the second problem, there are a number of approaches in the literature. The question raised earlier could be rephrased as "Given that it is known there are s_j uncensored observations of y_j, what is the probability that it was the s_j policies that actually had that value? Do this conditioned on equalling or exceeding that value." A direct interpretation of this statement would have the numerator reflect the probability of the s_j observations that were observed. The denominator would be based on all subsets of $R(y_j)$ with s_j members. This is a lot of work. A simplified version due to Breslow treats each of the s_j observations separately but, for the denominator, uses the same risk set for all of them. The effect is to require no change from the algorithm introduced above.

Example 12.67 *In the previous example, suppose that the observation of* 81 *had actually been* 70. *Give the contribution to the partial likelihood function for these two observations.*

Using the notation from that example, the contribution for the first observation of 70 would still be $c_5/(c_5+\cdots+c_{10})$. However, the second observation of 70 would now contribute $c_6/(c_5+\cdots+c_{10})$. Note that the numerator has not changed (it is still c_6); however, the denominator reflects the fact that there are six observations in $R(70)$. □

With regard to estimating the hazard rate function, we first note that the cumulative hazard rate function is

$$H(t|\mathbf{z}) = \int_0^t h(u|\mathbf{z})du = \int_0^t h_0(u)c\,du = H_0(t)c.$$

Table 12.22 Fire insurance baseline survival function

Value	y	c	Jump	$\hat{H}_0(y)$	$\hat{S}_0(y)$
8	8	0.8300	$\frac{1}{0.8300+\cdots+0.3699} = 0.1597$	0.1597	0.8524
22	22	0.9282	$\frac{1}{0.9282+\cdots+0.3699} = 0.1841$	0.3438	0.7091
51	51	0.3840	$\frac{1}{0.3840+\cdots+0.3699} = 0.2220$	0.5658	0.5679
55	55	0.3564	$\frac{1}{0.3564+\cdots+0.3699} = 0.2427$	0.8086	0.4455
70	70	0.9634	$\frac{1}{0.9634+\cdots+0.3699} = 0.2657$	1.0743	0.3415
81	81	0.8615	$\frac{1}{0.8615+\cdots+0.3699} = 0.3572$	1.4315	0.2390
85		0.3434			
90		0.8942			
93	93	0.3308	$\frac{1}{0.3308+0.3699} = 1.4271$	2.8586	0.0574
95		0.3699			

To employ an analog of the Nelson–Åalen estimate, we use

$$\hat{H}_0(t) = \sum_{y_j \le t} \frac{s_j}{\sum_{i\in R(y_j)} c_i}.$$

That is, the outer sum is taken over all uncensored observations less than or equal to t. The numerator is the number of observations having an uncensored value equal to y_j and the denominator, rather than having the number in the risk set, adds their c values. As usual, the baseline survival function is estimated as $\hat{S}_0(t) = \exp[-\hat{H}_0(t)]$.

Example 12.68 *For the continuing example (using the original values), estimate the baseline survival function and then estimate the probability that a claim for a* 35*-year-old wood house will exceed* 80% *of the house's value. Compare this to the value obtained from the beta distribution model obtained earlier.*

Using the estimates obtained earlier, the 10 c values are as given in Table 12.22. Also included is the jump in the cumulative hazard estimate, followed by the estimate of the cumulative hazard function itself. Values for that function apply from the given y value up to, but not including, the next y value.

For the house as described, $c = \exp[-0.00373(35) - 0.91994(1)] = 0.34977$. The estimated probability is $0.3415^{0.34977} = 0.68674$. From the beta distribution, $\hat{S}_0(0.8) = 0.27732$ and $c = \exp[-0.00315(35) - 0.77847(1)] = 0.41118$, which gives an estimated probability of $0.27732^{0.41118} = 0.59015$. □

With regard to variance estimates, the logarithm of the partial likelihood function is

$$l(\beta) = \sum_{j^*} \ln \frac{c_{j^*}}{\sum_{i \in R(y_j)} c_i},$$

where the sum is taken over all observations that produced an uncensored value. Taking the first partial derivative with respect to β_g produces

$$\frac{\partial}{\partial \beta_g} l(\beta) = \sum_{j^*} \left[\frac{1}{c_{j^*}} \frac{\partial c_{j^*}}{\partial \beta_g} - \frac{1}{\sum_{i \in R(y_j)} c_i} \frac{\partial}{\partial \beta_g} \sum_{i \in R(y_j)} c_i \right].$$

To simplify this expression, note that

$$\frac{\partial c_{j^*}}{\partial \beta_g} = \frac{\partial e^{\beta_1 z_{j^*1} + \beta_2 z_{j^*2} + \cdots + \beta_p z_{j^*p}}}{\partial \beta_g} = z_{j^*g} c_{j^*},$$

where z_{j^*g} is the value of z_g for subject j^*. The derivative is

$$\frac{\partial}{\partial \beta_g} l(\beta) = \sum_{j^*} \left[z_{j^*g} - \frac{\sum_{i \in R(y_j)} z_{ig} c_i}{\sum_{i \in R(y_j)} c_i} \right].$$

The negative second partial derivative is

$$-\frac{\partial^2}{\partial \beta_h \beta_g} l(\beta) = \sum_{j^*} \left[\frac{\sum_{i \in R(y_j)} z_{ig} z_{ih} c_i}{\sum_{i \in R(y_j)} c_i} - \frac{\left(\sum_{i \in R(y_j)} z_{ig} c_i\right)\left(\sum_{i \in R(y_j)} z_{ih} c_i\right)}{\left(\sum_{i \in R(y_j)} c_i\right)^2} \right].$$

Using the estimated values, these partial derivatives provide an estimate of the information matrix.

Example 12.69 *Obtain the information matrix and estimated covariance matrix for the continuing example. Then use this to produce a 95% confidence interval for the relative risk of a wood house versus a brick house of the same age.*

Consider the entry in the outer sum for the observation with $z_1 = 50$ and $z_2 = 1$. The risk set contains this observation (with a value of 93 and $c = 0.330802$) and the censored observation with $z_1 = 20$ and $z_2 = 1$ (with a value of 95 and $c = 0.369924$). For the derivative with respect to β_1 and β_2 the entry is

$$\frac{50(1)(0.330802) + 20(1)(0.369924)}{0.330802 + 0.369924} - \frac{[50(0.330802) + 20(0.369924)][1(0.330802) + 1(0.369924)]}{(0.330802 + 0.369924)^2} = 0.$$

Summing such items and doing the same for the other partial derivatives yield the information matrix and its inverse, the covariance matrix.

$$I = \begin{bmatrix} 1171.054 & 5.976519 \\ 5.976519 & 1.322283 \end{bmatrix}, \widehat{\text{Var}} = \begin{bmatrix} 0.000874 & -0.00395 \\ -0.00395 & 0.774125 \end{bmatrix}.$$

The relative risk is the ratio of the c values for the two cases. For a house of age x, the relative risk of wood versus brick is $e^{z_1\beta_1 + \beta_2} / e^{z_1\beta_1} = e^{\beta_2}$. A 95% confidence interval for β_2 is $-0.91994 \pm 1.96\sqrt{0.774125}$ or $(-2.6444, 0.80455)$. Exponentiating the endpoints gives the confidence interval for the relative risk, $(0.07105, 2.2357)$. □

12.7.3 The generalized linear and accelerated failure time models

The proportional hazards model requires a particular relationship between survival functions. For actuarial purposes it may not be the most appropriate because it is difficult to interpret the meaning of multiplying a hazard rate function by a constant (or, equivalently, raising a survival function to a power).[18] It may be more useful to relate the covariates to a quantity of direct interest, such as the expected value. Linear models, such as the standard multiple regression model are inadequate because they tend to rely on the normal distribution, a model not suitable for most phenomena of interest to actuaries. The generalized linear model drops that restriction and so may be more useful. A comprehensive reference is [90] and actuarial papers using the model include [47], [61], [93], and [97]. The definition of this model given below is slightly more general than the usual one.

Definition 12.70 *Suppose a parametric distribution has parameters μ and $\boldsymbol{\theta}$, where μ is the mean and $\boldsymbol{\theta}$ is a vector of additional parameters. Let its cdf be $F(x|\mu, \boldsymbol{\theta})$. The mean must not depend on the additional parameters and the additional parameters must not depend on the mean. Let $\mathbf{z}$ be a vector of covariates for an individual, let $\boldsymbol{\beta}$ be a vector of coefficients, and let $\eta(\mu)$ and $c(y)$ be functions. The generalized linear model then states that the random variable, X, has as its distribution function*

$$F(x|\mathbf{z}, \boldsymbol{\theta}) = F(x|\mu, \boldsymbol{\theta}),$$

where μ is such that $\eta(\mu) = c(\boldsymbol{\beta}^T \mathbf{z})$.

The model indicates that the mean of an individual observation is related to the covariates through a particular set of functions. Normally, these functions do not involve parameters, but instead are used to provide a good fit or to ensure that only legitimate values of μ are encountered.

[18] However, it is not uncommon in life insurance to incorporate a given health risk (such as obesity) by multiplying the values of q_x by a constant. This is not much different from multiplying the hazard rate function by a constant.

Example 12.71 *Demonstrate that the ordinary linear regression model is a special case of the generalized linear model.*

For ordinary linear regression, X has a normal distribution with $\mu = \mu$ and $\theta = \sigma^2$. Both η and c are the identity function, resulting in $\mu = \boldsymbol{\beta}^T\mathbf{z}$. □

The model presented here is more general than the one usually used where only a few distributions are allowed for X. The reason is that, for these distributions, it has been possible to develop the full set of regression tools, such as residual analysis. Computer packages that implement the generalized linear model use only these distributions.

For many of the distributions we have been using, the mean is not a parameter. However, it could be. For example, we could parameterize the Pareto distribution by setting $\mu = \theta/(\alpha-1)$ or, equivalently, replacing θ with $\mu(\alpha-1)$. The distribution function is now

$$F(x|\mu,\alpha) = 1 - \left[\frac{\mu(\alpha-1)}{\mu(\alpha-1)+x}\right]^{\alpha}, \quad \mu > 0,\ \alpha > 1.$$

Note the restriction on α in the parameter space.

Example 12.72 *Construct a generalized linear model for the data set in Example* 12.65 *using a beta distribution for the loss model.*

The beta distribution as parameterized in Appendix A has a mean of $\mu = a/(a+b)$. Let the other parameter be $\theta = b$. One way of linking the covariates to the mean is to use $\eta(\mu) = \mu/(1-\mu)$ and $c(\boldsymbol{\beta}^T\mathbf{z}) = \exp(\boldsymbol{\beta}^T\mathbf{z})$. Setting these equal and solving yields

$$\mu = \frac{\exp(\boldsymbol{\beta}^T\mathbf{z})}{1+\exp(\boldsymbol{\beta}^T\mathbf{z})}.$$

Solving the first two equations yields $a = b\mu/(1-\mu) = b\exp(\boldsymbol{\beta}^T\mathbf{z})$. Maximum likelihood estimation proceeds by using a factor of $f(x)$ for each uncensored observation and $1-F(x)$ for each censored observation. For each observation, the beta distribution uses the parameter b directly, and the parameter a from the value of b and the covariates for that observation. Because there is no baseline distribution, the expression $\boldsymbol{\beta}^T\mathbf{z}$ must include a constant term. Maximizing the likelihood function yields the estimates $\hat{b} = 0.5775$, $\hat{\beta}_0 = 0.2130$, $\hat{\beta}_1 = 0.0018$, and $\hat{\beta}_2 = 1.0940$. As in Example 12.65, the impact of age is negligible. One advantage of this model is that the mean is directly linked to the covariates. □

A model that is similar in spirit to the generalized linear model is the accelerated failure time model as described below.

Definition 12.73 *The* ***accelerated failure time model*** *is defined from*

$$S(x|\mathbf{z},\boldsymbol{\beta}) = S_0(xe^{-\boldsymbol{\beta}^T\mathbf{z}}). \tag{12.29}$$

Table 12.23 Data for Example 12.74

	Male (0)			Female (1)		
Age	100	125	150	100	125	150
50	13	12	85	3	12	49
51	11	21	95	7	13	53
52	8	8	105	8	13	69
53	10	20	113	12	16	61
54	8	11	109	12	15	60
55	13	22	126	8	12	68
56	19	16	142	11	11	96
57	9	19	145	5	19	97
58	17	23	155	5	17	93
59	14	28	182	9	14	96

To see that, provided the mean exists, it is a generalized linear model, first note that, (assuming $S(0) = 1$),

$$\mathrm{E}(X|\mathbf{z},\boldsymbol{\beta}) = \int_0^{\infty} S_0(xe^{-\boldsymbol{\beta}^T\mathbf{z}})dx = \int_0^{\infty} e^{\boldsymbol{\beta}^T\mathbf{z}}S_0(y)dy = \exp(\boldsymbol{\beta}^T\mathbf{z})E_0(X),$$

thus relating the mean of the distribution to the covariates. The name comes from the fact that the covariates effectively change the age. A person age x whose covariates are $\mathbf{z}$ has a future lifetime with the same distribution as a person for whom $\mathbf{z} = 0$ and is age $xe^{-\boldsymbol{\beta}^T\mathbf{z}}$. If the baseline distribution has a scale parameter, then the effect of the covariates is to multiply that scale parameter by a constant. So, if θ is the scale parameter for the baseline distribution then a person with covariate $\mathbf{z}$ will have the same distribution, but with scale parameter $\exp(\boldsymbol{\beta}^T\mathbf{z})\theta$. Unlike the generalized linear model, it is not necessary for the mean to exist before this model can be used.

Example 12.74 *A mortality study at ages* 50–59 *included people of both genders and with systolic blood pressure of* 100, 125, *or* 150. *For each of the* 6 *combinations and at each of the* 10 *ages,* 1,000 *people were observed and the number of deaths recorded. The data appear in Table* 12.23. *Develop and estimate the parameters for an accelerated failure time model based on the Gompertz distribution.*

The Gompertz distribution has hazard rate function $h(x) = Bc^x$ which implies a survival function of $S_0(x) = \exp[-B(c^x-1)/\ln c]$ as the baseline distribution. Let the covariates be $z_1 = 0$ for males and $z_1 = 1$ for females and let z_2 be the blood pressure. For an individual insured, let $\gamma = \exp(\beta_1 z_1 + \beta_2 z_2)$. The accelerated failure time model implies that

$$S(x|\gamma) = S_0\left(\frac{x}{\gamma}\right) = \exp\left[-\frac{B(c^{x/\gamma}-1)}{\ln c}\right].$$

Let $c^* = c^{1/\gamma}$ and let $B^* = B/\gamma$. Then,

$$S(x|\gamma) = \exp\left[-\frac{B^*\gamma(c^{*x}-1)}{\gamma \ln c^*}\right] = \exp\left[-\frac{B^*(c^{*x}-1)}{\ln c^*}\right],$$

and so the distribution remains Gompertz with new parameters as indicated. For each age,

$$q_x|\gamma = 1 - \frac{S(x+1|\gamma)}{S(x|\gamma)} = 1 - \exp\left[-\frac{B^*c^{*x}(c^*-1)}{\ln c^*}\right].$$

If there are d_x deaths at age x, the contribution to the loglikelihood function (where a binomial distribution has been assumed for the number of deaths) is

$$d_x \ln q_x + (1000 - d_x)\ln(1 - q_x).$$

The likelihood function is maximized at $B = 0.000243$, $c = 1.00866$, $\beta_1 = 0.110$, and $\beta_2 = -0.0144$. Being female multiplies the expected lifetime (from birth) by a factor of $\exp(0.110) = 1.116$. An increase of 25 in blood pressure lowers the expected lifetime by $1 - \exp[-25(0.0144)] = 0.302$, or a 30.2% decrease. □

12.7.4 Exercises

12.101 Suppose the 40 observations in Data Set D2 in Chapter 10 were from four types of policyholders. Observations 1, 5, ... are from male smokers, observations 2, 6, ... are from male nonsmokers, observations 3, 7, ... are from female smokers, and observations 4, 8, ... are from female nonsmokers. You are to construct a model for the time to surrender and then use the model to estimate the probability of surrendering in the first year for each of the four cases. Construct each of the following three models:

(a) Use four different Nelson–Åalen estimates, keeping the four groups separate.

(b) Use a proportional hazards model where the baseline distribution has the exponential distribution.

(c) Use a proportional hazards model with an empirical estimate of the baseline distribution.

12.102 (*) The duration of a strike follows a Cox proportional hazards model in which the baseline distribution has an exponential distribution. The only variable used is the index of industrial production. When the index has a value of 10, the mean duration of a strike is 0.2060 years. When the index has a value of 25, the median duration is 0.0411 years. Determine the probability that a strike will have a duration of more than one year if the index has a value of 5.

12.103 (*) A Cox proportional hazards model has $z_1 = 1$ for males and $z_1 = 0$ for females and $z_2 = 1$ for adults and $z_2 = 0$ for children. The maximum likelihood estimates of the coefficients are $\hat{\beta}_1 = 0.25$ and $\hat{\beta}_2 = -0.45$. The covariance matrix of the estimators is

$$\begin{bmatrix} 0.36 & 0.10 \\ 0.10 & 0.20 \end{bmatrix}.$$

Determine a 95% confidence interval for the relative risk of a male child subject compared to a female adult subject.

12.104 (*) Four insureds were observed from birth to death. The two from Class A died at times 1 and 9 while the two from Class B died at times 2 and 4. A proportional hazards model uses $z_1 = 1$ for Class B and 0 for Class A. Let $b = \hat{\beta}_1$. Estimate the cumulative hazard rate at time 3 for a member of Class A.

12.105 (*) A Cox proportional hazards model has three covariates. The life that died first has values $1, 0, 0$ for z_1, z_2, z_3. The second to die has values $0, 1, 0$ and the third to die has values $0, 0, 1$. Determine the partial likelihood function (as a function of β_1, β_2, and β_3).

12.106 Repeat Example 12.72 using only construction type and not age.

12.107 Repeat Example 12.74 using a proportional hazards model with a Gompertz baseline distribution.

12.108 Repeat Example 12.74 using an accelerated failure time model with a gamma baseline distribution.

13

Model selection

13.1 INTRODUCTION

When using data to build a model, the process must end with the announcement of a "winner." While qualifications, limitations, caveats, and other attempts to escape full responsibility are appropriate, and often necessary, a commitment to a solution is often required. In this chapter we look at a variety of ways to evaluate a model and compare competing models. But we must also remember that whatever model we select it is only an approximation of reality. This is reflected in the following modeler's motto[1]:

> All models are wrong, but some models are useful.

Thus, our goal is to determine a model that is good enough to use to answer the question. The challenge here is that the definition of *good enough* will depend on the particular application. Another important modeling point is that a solid understanding of the question will guide you to the answer. The following quote from John Tukey [131], pp. 13–14 sums this up:

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

[1] It is usually attributed to George Box.

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

In this chapter, a specific modeling strategy will be considered. Our preference is to have a single approach that can be used for any probabilistic modeling situation. A consequence is that for any particular modeling situation there may be a better (more reliable or more accurate) approach. For example, while maximum likelihood is a good estimation method for most settings, it may not be the best[2] for certain distributions. A literature search will turn up methods that have been optimized for specific distributions, but they will not be mentioned here. Similarly, many of the hypothesis tests used here give approximate results. For specific cases, better approximations, or maybe even exact results, are available. They will also be bypassed. The goal here is to outline a method that will give reasonable answers most of the time and be adaptable to a variety of situations.

This chapter assumes the reader has a basic understanding of statistical hypothesis testing as reviewed in Chapter 9. The remaining sections cover a variety of evaluation and selection tools. Each tool has its own strengths and weaknesses, and it is possible for different tools to lead to different models. This makes modeling as much art as science. At times, in real-world applications, the model's purpose may lead the analyst to favor one tool over another.

13.2 REPRESENTATIONS OF THE DATA AND MODEL

All the approaches to be presented attempt to compare the proposed model to the data or to another model. The proposed model is represented by either its density or distribution function or perhaps some functional of these quantities such as the limited expected value function or the mean residual life function. The data can be represented by the empirical distribution function or a histogram. The graphs are easy to construct when there is individual, complete data. When there is grouping or observations have been truncated or censored, difficulties arise. Here, the only cases to be covered are those where all the data have been truncated at the same value (which could be zero) and are all censored at the same value (which could be infinity). Extensions to the case of multiple truncation or censoring points are detailed in [109].[3] It should be noted that the need for such representations applies only to continuous models. For discrete data, issues of censoring, truncation, and grouping rarely apply. The data can easily be represented by the relative or cumulative frequencies at each possible observation.

[2]There are many definitions of "best." Combining the Cramér-Rao lower bound with Theorem 12.13 indicates that maximum likelihood estimators are asymptotically optimal using unbiasedness and minimum variance as the definition of best.

[3]Because the Kaplan-Meier estimate can be used to represent data with multiple truncation or censoring points, constructing graphical comparisons of the model and data is not difficult. The major challenge is generalizing the hypothesis tests to this situation.

Table 13.1 Data Set B with highest value changed

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1,193	1,340	1,884	2,558	3,476

With regard to representing the data, the empirical distribution function will be used for individual data and the histogram will be used for grouped data.

In order to compare the model to truncated data, we begin by noting that the empirical distribution begins at the truncation point and represents conditional values (that is, they are the distribution and density function given that the observation exceeds the truncation point). In order to make a comparison to the empirical values, the model must also be truncated. Let the truncation point in the data set be t. The modified functions are

$$F^*(x) = \begin{cases} 0, & x < t, \\ \dfrac{F(x) - F(t)}{1 - F(t)}, & x \geq t, \end{cases}$$

$$f^*(x) = \begin{cases} 0, & x < t, \\ \dfrac{f(x)}{1 - F(t)}, & x \geq t. \end{cases}$$

In this chapter, when a distribution function or density function is indicated, a subscript equal to the sample size indicates that it is the empirical model (from Kaplan–Meier, Nelson–Åalen, the ogive, etc.) while no adornment or the use of an asterisk (*), indicates the estimated parametric model. There is no notation for the true, underlying distribution because it is unknown and unknowable.

13.3 GRAPHICAL COMPARISON OF THE DENSITY AND DISTRIBUTION FUNCTIONS

The most direct way to see how well the model and data match up is to plot the respective density and distribution functions.

Example 13.1 *Consider Data Sets B and C. However, for this example and all that follow, in Data Set B replace the value at* 15,743 *with* 3,476 (*this is to allow the graphs to fit comfortably on a page*). *These data sets are reproduced here in Tables* 13.1 *and* 13.2. *Truncate Data Set B at* 50 *and Data Set C at* 7,500. *Estimate the parameter of an exponential model for each data set. Plot the appropriate functions and comment on the quality of the fit of the model. Repeat this for Data Set B censored at* 1,000 (*without any truncation*).

Table 13.2 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

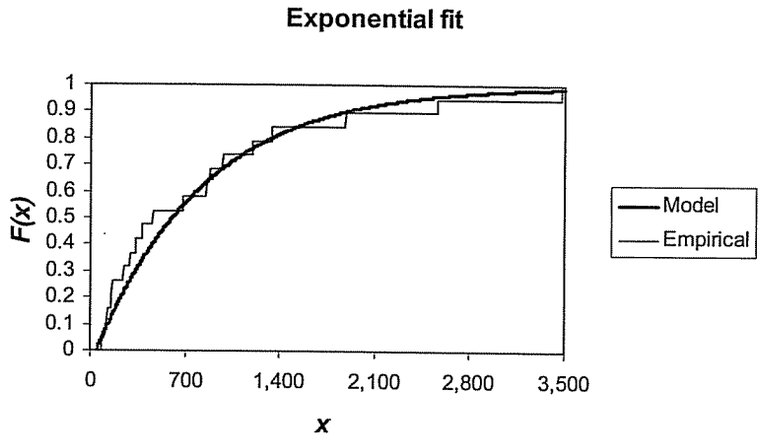


Fig. 13.1 Model vs. data cdf plot for Data Set B truncated at 50.

For Data Set B, there are 19 observations (the first observation is removed due to truncation). A typical contribution to the likelihood function is $f(82)/[1-F(50)]$. The maximum likelihood estimate of the exponential parameter is $\hat{\theta} = 802.32$. The empirical distribution function starts at 50 and jumps 1/19 at each data point. The distribution function, using a truncation point of 50, is

$$F^*(x) = \frac{1-e^{-x/802.32} - (1-e^{-50/802.32})}{1-(1-e^{-50/802.32})} = 1-e^{-(x-50)/802.32}.$$

Figure 13.1 presents a plot of these two functions.

The fit is not as good as we might like because the model understates the distribution function at smaller values of x and overstates the distribution

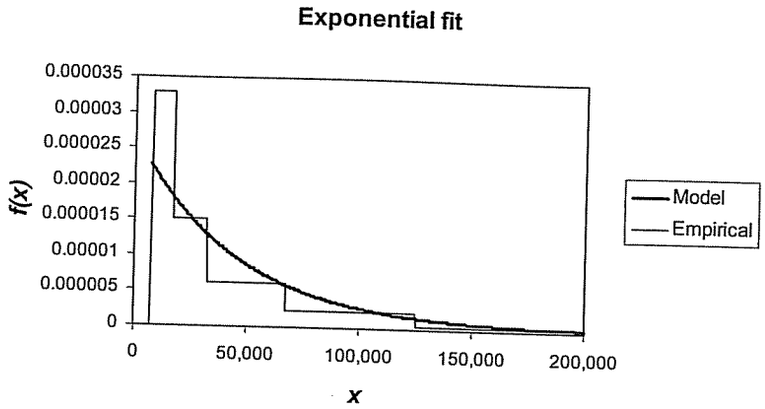


Fig. 13.2 Model vs. data density plot for Data Set C truncated at 7,500.

function at larger values of x. This is not good because it means that tail probabilities are understated.

For Data Set C, the likelihood function uses the truncated values. For example, the contribution to the likelihood function for the first interval is

$$\left[\frac{F(17{,}500)-F(7{,}500)}{1-F(7{,}500)}\right]^{42}.$$

The maximum likelihood estimate is $\hat{\theta} = 44{,}253$. The height of the first histogram bar is

$$\frac{42}{128(17{,}500-7{,}500)} = 0.0000328$$

and the last bar is for the interval from 125,000 to 300,000 (a bar cannot be constructed for the interval from 300,000 to infinity). The density function must be truncated at 7,500 and becomes

$$\begin{aligned} f^*(x) &= \frac{f(x)}{1-F(7{,}500)} = \frac{44{,}253^{-1}e^{-x/44{,}253}}{1-(1-e^{-7{,}500/44{,}253})} \\ &= \frac{e^{-(x-7{,}500)/44{,}253}}{44{,}253}, \quad x > 7{,}500. \end{aligned}$$

The plot of the density function versus the histogram is given Figure 13.2.

The exponential model understates the early probabilities. It is hard to tell from the picture how the curves compare above 125,000.

For Data Set B modified with a limit, the maximum likelihood estimate is $\hat{\theta} = 718.00$. When constructing the plot, the empirical distribution function must stop at 1,000. The plot appears in Figure 13.3.

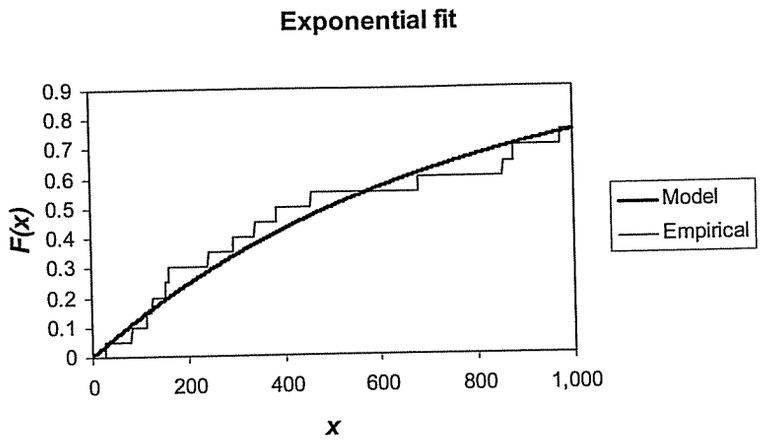


Fig. 13.3 Model vs. data cdf plot for Data Set B censored at 1,000.

Once again, the exponential model does not fit well. □

When the model's distribution function is close to the empirical distribution function, it is difficult to make small distinctions. Among the many ways to amplify those distinctions, two will be presented here. The first is to simply plot the difference of the two functions. That is, if $F_n(x)$ is the empirical distribution function and $F^*(x)$ is the model distribution function, plot $D(x) = F_n(x) - F^*(x)$.

Example 13.2 *Plot $D(x)$ for the previous example.*

For Data Set B truncated at 50, the plot appears in Figure 13.4. The lack of fit for this model is magnified in this plot.

There is no corresponding plot for grouped data. For Data Set B censored at 1,000, the plot must again end at that value. It appears in Figure 13.5. The lack of fit continues to be apparent. □

Another way to highlight any differences is the *p–p* plot, which is also called a probability plot. The plot is created by ordering the observations as $x_1 \le \cdots \le x_n$. A point is then plotted corresponding to each value. The coordinates to plot are $(F_n(x_j), F^*(x_j))$.[4] If the model fits well, the plotted

[4]In the first edition of this text this plot was incorrectly called a *q-q* plot. There is a plot that goes by that name, but it will not be introduced here.

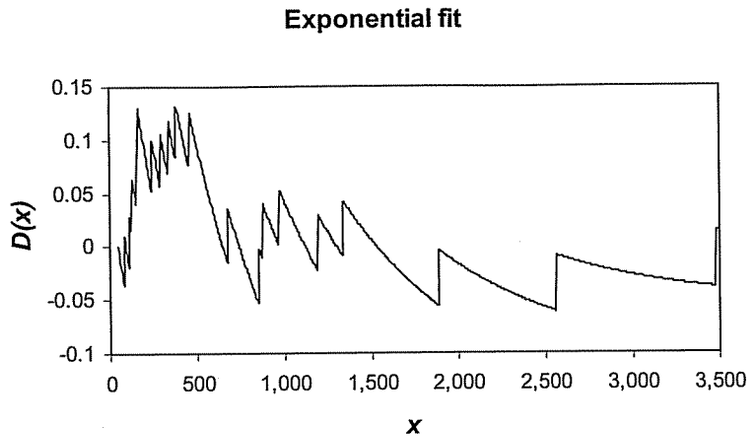


Fig. 13.4 Model vs. data $D(x)$ plot for Data Set B truncated at 50.

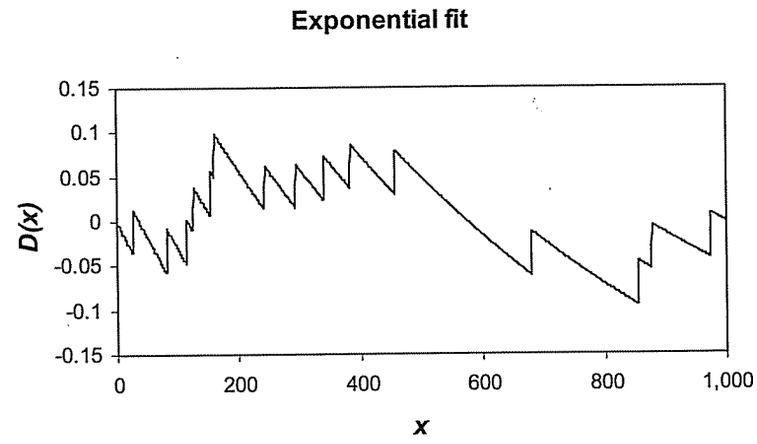


Fig. 13.5 Model vs. data $D(x)$ plot for Data Set B censored at 1,000.

points will be near the 45° line running from (0, 0) to (1, 1). However, for this to be the case, a different definition of the empirical distribution function is needed. It can be shown that the expected value of $F_n(x_j)$ is $j/(n+1)$ and therefore the empirical distribution should be that value and not the usual j/n. If two observations have the same value, either plot both points (they

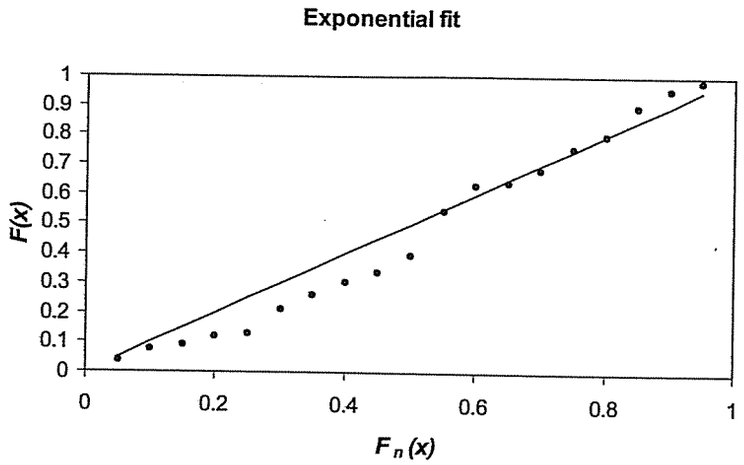


Fig. 13.6 *p–p* for Data Set B truncated at 50.

would have the same "y" value but different "x" values) or plot a single value by averaging the two "x" values.

Example 13.3 *Create a p–p plot for the continuing example.*

For Data Set B truncated at 50, $n = 19$ and one of the observed values is $x = 82$. The empirical value is $F_n(82) = \frac{1}{20} = 0.05$. The other coordinate is

$$F^*(82) = 1 - e^{-(82-50)/802.32} = 0.0391.$$

One of the plotted points will be $(0.05, 0.0391)$. The complete picture appears in Figure 13.6.

From the lower left part of the plot it is clear that the exponential model places less probability on small values than the data call for. A similar plot can be constructed for Data Set B censored at 1,000 and it appears in Figure 13.7.

This plot ends at about 0.75 because that is the highest probability observed prior to the censoring point at 1,000. There are no empirical values at higher probabilities. Again, the exponential model tends to underestimate the empirical values. □

13.3.1 Exercises

13.1 Repeat Example 13.1 using a Weibull model in place of the exponential model.

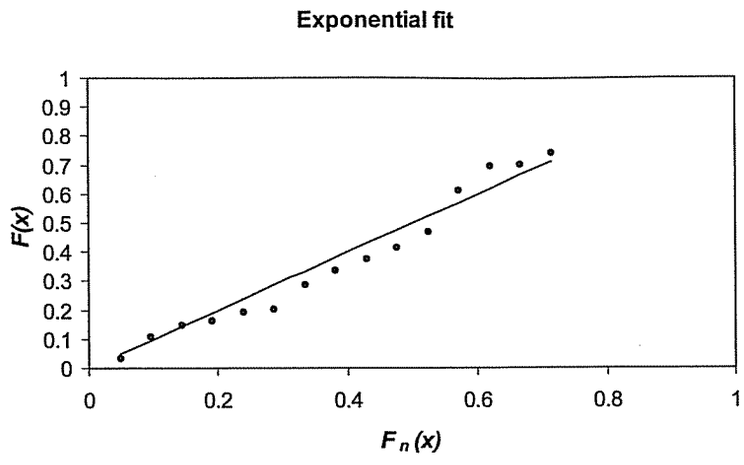


Fig. 13.7 *p–p* plot for Data Set B censored at 1,000.

13.2 Repeat Example 13.2 for a Weibull model.

13.3 Repeat Example 13.3 for a Weibull model.

13.4 HYPOTHESIS TESTS

A picture may be worth many words, but sometimes it is best to replace the impressions conveyed by pictures with mathematical demonstrations. One such demonstration is a test of the hypotheses

H_0 : The data came from a population with the stated model.
H_1 : The data did not come from such a population.

The test statistic is usually a measure of how close the model distribution function is to the empirical distribution function. When the null hypothesis completely specifies the model (for example, an exponential distribution with mean 100), critical values are well known. However, it is more often the case that the null hypothesis states the name of the model but not its parameters. When the parameters are estimated from the data, the test statistic tends to be smaller than it would have been had the parameter values been prespecified. That is because the estimation method itself tries to choose parameters that produce a distribution that is close to the data. In that case, the tests become approximate. Because rejection of the null hypothesis occurs for large values of the test statistic, the approximation tends to increase the probability of a

Type II error while lowering the probability of a Type I error.[5] For actuarial modeling this is likely to be an acceptable trade-off.

One method of avoiding the approximation is to randomly divide the sample in half. Use one half to estimate the parameters and then use the other half to conduct the hypothesis test. Once the model is selected, the full data set could be used to reestimate the parameters.

13.4.1 Kolmogorov–Smirnov test

Let t be the left truncation point ($t = 0$ if there is no truncation) and let u be the right censoring point ($u = \infty$ if there is no censoring). Then, the test statistic is

$$D = \max_{t \le x \le u} |F_n(x) - F^*(x)| .$$

This test should only be used on individual data. This is to ensure that the step function $F_n(x)$ is well defined. Also, the model distribution function $F^*(x)$ is assumed to be continuous over the relevant range.

Example 13.4 *Calculate D for Example* 13.1.

Table 13.3 provides the needed values. Because the empirical distribution function jumps at each data point, the model distribution function must be compared both before and after the jump. The values just before the jump are denoted $F_n(x-)$ in the table. The maximum is $D = 0.1340$.

For Data Set B censored at 1,000, 15 of the 20 observations are uncensored. Table 13.4 illustrates the needed calculations. The maximum is $D = 0.0991$. □

All that remains is to determine the critical value. Commonly used critical values for this test are $1.22/\sqrt{n}$ for $\alpha = 0.10$, $1.36/\sqrt{n}$ for $\alpha = 0.05$, and $1.63/\sqrt{n}$ for $\alpha = 0.01$. When $u < \infty$, the critical value should be smaller because there is less opportunity for the difference to become large. Modifications for this phenomenon exist in the literature (see [125], for example, which also includes tables of critical values for specific null distribution models), and one such modification is given in [109] but will not be introduced here.

Example 13.5 *Complete the Kolmogorov–Smirnov test for the previous example.*

For Data Set B truncated at 50 the sample size is 19. The critical value at a 5% significance level is $1.36/\sqrt{19} = 0.3120$. Because $0.1340 < 0.3120$, the null hypothesis is not rejected and the exponential distribution is a plausible model. While it is unlikely that the exponential model is appropriate for this population, the sample size is too small to lead to that conclusion. For Data Set B censored at 1,000 the sample size is 20 and so the critical value is $1.36/\sqrt{20} = 0.3041$ and the exponential model is again viewed as being plausible. □

[5] Among the tests presented here, only the chi-square test has a built-in correction for this situation. Modifications for the other tests have been developed, but they will not be presented here.

Table 13.3 Calculation of D for Example 13.4

x	$F^*(x)$	$F_n(x-)$	$F_n(x)$	Maximum difference
82	0.0391	0.0000	0.0526	0.0391
115	0.0778	0.0526	0.1053	0.0275
126	0.0904	0.1053	0.1579	0.0675
155	0.1227	0.1579	0.2105	0.0878
161	0.1292	0.2105	0.2632	0.1340
243	0.2138	0.2632	0.3158	0.1020
294	0.2622	0.3158	0.3684	0.1062
340	0.3033	0.3684	0.4211	0.1178
384	0.3405	0.4211	0.4737	0.1332
457	0.3979	0.4737	0.5263	0.1284
680	0.5440	0.5263	0.5789	0.0349
855	0.6333	0.5789	0.6316	0.0544
877	0.6433	0.6316	0.6842	0.0409
974	0.6839	0.6842	0.7368	0.0529
1,193	0.7594	0.7368	0.7895	0.0301
1,340	0.7997	0.7895	0.8421	0.0424
1,884	0.8983	0.8421	0.8947	0.0562
2,558	0.9561	0.8947	0.9474	0.0614
3,476	0.9860	0.9474	1.0000	0.0386

For both this test and the Anderson–Darling test that follows, the critical values are correct only when the null hypothesis completely specifies the model. When the data set is used to estimate parameters for the null hypothesized distribution (as in the example), the correct critical value is smaller. For both tests, the change depends on the particular distribution that is hypothesized and maybe even on the particular true values of the parameters. An indication of how simulation can be used for this situation is presented in Section 17.2.4.

Table 13.4 Calculation of D for Example 13.4 with censoring

x	$F^*(x)$	$F_n(x-)$	$F_n(x)$	Maximum difference
27	0.0369	0.00	0.05	0.0369
82	0.1079	0.05	0.10	0.0579
115	0.1480	0.10	0.15	0.0480
126	0.1610	0.15	0.20	0.0390
155	0.1942	0.20	0.25	0.0558
161	0.2009	0.25	0.30	0.0991
243	0.2871	0.30	0.35	0.0629
294	0.3360	0.35	0.40	0.0640
340	0.3772	0.40	0.45	0.0728
384	0.4142	0.45	0.50	0.0858
457	0.4709	0.50	0.55	0.0791
680	0.6121	0.55	0.60	0.0621
855	0.6960	0.60	0.65	0.0960
877	0.7052	0.65	0.70	0.0552
974	0.7425	0.70	0.75	0.0425
1000	0.7516	0.75	0.75	0.0016

13.4.2 Anderson–Darling test

This test is similar to the Kolmogorov–Smirnov test, but uses a different measure of the difference between the two distribution functions. The test statistic is

$$A^2 = n\int_t^u \frac{[F_n(x) - F^*(x)]^2}{F^*(x)[1 - F^*(x)]} f^*(x)dx.$$

That is, it is a weighted average of the squared differences between the empirical and model distribution functions. Note that when x is close to t or to u the weights might be very large due to the small value of one of the factors in the denominator. This test statistic tends to place more emphasis on good fit in the tails than in the middle of the distribution. Calculating with this formula appears to be challenging. However, for individual data (so this is another test that does not work for grouped data), the integral simplifies to

$$\begin{aligned} A^2 &= -nF^*(u) + n\sum_{j=0}^{k}[1 - F_n(y_j)]^2\{\ln[1 - F^*(y_j)] - \ln[1 - F^*(y_{j+1})]\} \\ &\quad + n\sum_{j=1}^{k} F_n(y_j)^2[\ln F^*(y_{j+1}) - \ln F^*(y_j)], \end{aligned}$$

where the unique noncensored data points are $t = y_0 < y_1 < \cdots < y_k < y_{k+1} = u$. Note that when $u = \infty$ the last term of the first sum is zero

Table 13.5 Anderson–Darling test for Example 13.6

j	y_j	$F^*(x)$	$F_n(x)$	Summand
0	50	0.0000	0.0000	0.0399
1	82	0.0391	0.0526	0.0388
2	115	0.0778	0.1053	0.0126
3	126	0.0904	0.1579	0.0332
4	155	0.1227	0.2105	0.0070
5	161	0.1292	0.2632	0.0904
6	243	0.2138	0.3158	0.0501
7	294	0.2622	0.3684	0.0426
8	340	0.3033	0.4211	0.0389
9	384	0.3405	0.4737	0.0601
10	457	0.3979	0.5263	0.1490
11	680	0.5440	0.5789	0.0897
12	855	0.6333	0.6316	0.0099
13	877	0.6433	0.6842	0.0407
14	974	0.6839	0.7368	0.0758
15	1,193	0.7594	0.7895	0.0403
16	1,340	0.7997	0.8421	0.0994
17	1,884	0.8983	0.8947	0.0592
18	2,558	0.9561	0.9474	0.0308
19	3,476	0.9860	1.0000	0.0141
20	∞	1.0000	1.0000	

[evaluating the formula as written will ask for ln(0)]. The critical values are 1.933, 2.492, and 3.857 for 10, 5, and 1% significance levels, respectively. As with the Kolmogorov–Smirnov test, the critical value should be smaller when $u < \infty$.

Example 13.6 *Perform the Anderson–Darling test for the continuing example.*

For Data Set B truncated at 50, there are 19 data points. The calculation is in Table 13.5, where "summand" refers to the sum of the corresponding terms from the two sums. The total is 1.0226 and the test statistic is $-19(1) + 19(1.0226) = 0.4292$. Because the test statistic is less than the critical value of 2.492, the exponential model is viewed as plausible.

For Data Set B censored at 1,000, the results are in Table 13.6. The total is 0.7602 and the test statistic is $-20(0.7516) + 20(0.7602) = 0.1713$. Because the test statistic does not exceed the critical value of 2.492, the exponential model is viewed as plausible. □

Table 13.6 Anderson–Darling calculation for Example 13.6 with censored data

j	y_j	$F^*(x)$	$F_n^*(x)$	Summand
0	0	0.0000	0.00	0.0376
1	27	0.0369	0.05	0.0718
2	82	0.1079	0.10	0.0404
3	115	0.1480	0.15	0.0130
4	126	0.1610	0.20	0.0334
5	155	0.1942	0.25	0.0068
6	161	0.2009	0.30	0.0881
7	243	0.2871	0.35	0.0493
8	294	0.3360	0.40	0.0416
9	340	0.3772	0.45	0.0375
10	384	0.4142	0.50	0.0575
11	457	0.4709	0.55	0.1423
12	680	0.6121	0.60	0.0852
13	855	0.6960	0.65	0.0093
14	877	0.7052	0.70	0.0374
15	974	0.7425	0.75	0.0092
16	1000	0.7516	0.75	

13.4.3 Chi-square goodness-of-fit test

Unlike the previous two tests, this test allows for some discretion. It begins with the selection of $k-1$ arbitrary values, $t = c_0 < c_1 < \cdots < c_k = \infty$. Let $\hat{p}_j = F^*(c_j) - F^*(c_{j-1})$ be the probability a truncated observation falls in the interval from c_{j-1} to c_j. Similarly, let $p_{nj} = F_n(c_j) - F_n(c_{j-1})$ be the same probability according to the empirical distribution. The test statistic is then

$$\chi^2 = \sum_{j=1}^{k} \frac{n(\hat{p}_j - p_{nj})^2}{\hat{p}_j},$$

where n is the sample size. Another way to write the formula is to let $E_j = n\hat{p}_j$ be the number of expected observations in the interval (assuming that the hypothesized model is true) and $O_j = np_{nj}$ be the number of observations in the interval. Then,

$$\chi^2 = \sum_{j=1}^{k} \frac{(E_j - O_j)^2}{E_j}.$$

The critical value for this test comes from the chi-square distribution with degrees of freedom equal to the number of terms in the sum (k) minus 1 minus the number of estimated parameters. There are a number of rules that have been proposed for deciding when the test is reasonably accurate. They center around the values of $E_j = n\hat{p}_j$. The most conservative states that each must

Table 13.7 Data Set B truncated at 50

Range	$\hat{p}$	Expected	Observed	χ^2
50–150	0.1172	2.227	3	0.2687
150–250	0.1035	1.966	3	0.5444
250–500	0.2087	3.964	4	0.0003
500–1,000	0.2647	5.029	4	0.2105
1,000–2,000	0.2180	4.143	3	0.3152
2,000–∞	0.0880	1.672	2	0.0644
Total	1	19	19	1.4034

be at least 5. Some authors claim that values as low as 1 are acceptable. All agree the test works best when the values are about equal from term to term. If the data are grouped, there is little choice but to use the groups as given, although adjacent groups could be combined to increase E_j. For individual data, the data can be grouped for the purpose of performing this test.[6]

Example 13.7 *Perform the chi-square goodness-of-fit test for the exponential distribution for the continuing example.*

All three data sets can be evaluated with this test. For Data Set B truncated at 50, establish boundaries at 50, 150, 250, 500, 1000, 2000, and infinity. The calculations appear in Table 13.7. The total is $\chi^2 = 1.4034$. With four degrees of freedom (6 rows minus 1 minus 1 estimated parameter) the critical value for a test at a 5% significance level is 9.4877 (this can be obtained with the Excel® function CHIINV(.05,4)) and the p-value is 0.8436 [from CHIDIST(1.4034,4)]. The exponential model is a good fit.

For Data Set B censored at 1,000, the first interval is from 0–150 and the last interval is from 1,000–∞. Unlike the previous two tests, the censored observations can be used. The calculations are in Table 13.8. The total is $\chi^2 = 0.5951$. With three degrees of freedom (5 rows minus 1 minus 1 estimated parameter) the critical value for a test at a 5% significance level is 7.8147 and the p-value is 0.8976. The exponential model is a good fit.

For Data Set C the groups are already in place. The results are given Table 13.9. The test statistic is $\chi^2 = 61.913$. There are four degrees of freedom for a critical value of 9.488. The p-value is about 10^{-12}. There is clear evidence that the exponential model is not appropriate. A more accurate test would

[6] Moore [95] cites a number of rules. Among them are (1) An expected frequency of at least 1 for all cells and and an expected frequency of at least 5 for 80% of the cells; (2) an average count per cell of at least 4 when testing at the 1% significance level and an average count of at least 2 when testing at the 5% significance level; and (3) A sample size of at least 10, at least 3 cells, and the ratio of the square of the sample size to the number of cells at least 10.

Table 13.8 Data Set B censored at 1,000

Range	$\hat{p}$	Expected	Observed	χ^2
0–150	0.1885	3.771	4	0.0139
150–250	0.1055	2.110	3	0.3754
250–500	0.2076	4.152	4	0.0055
500–1,000	0.2500	5.000	4	0.2000
1,000–∞	0.2484	4.968	5	0.0002
Total	1	20	20	0.5951

Table 13.9 Data Set C

Range	$\hat{p}$	Expected	Observed	χ^2
7,500–17,500	0.2023	25.889	42	10.026
17,500–32,500	0.2293	29.356	29	0.004
32,500–67,500	0.3107	39.765	28	3.481
67,500–125,000	0.1874	23.993	17	2.038
125,000–300,000	0.0689	8.824	9	0.003
300,000–∞	0.0013	0.172	3	46.360
Total	1	128	128	61.913

combine the last two groups (because the expected count in the last group is less than 1). The group from 125,000 to infinity has an expected count of 8.997 and an observed count of 12 for a contribution of 1.002. The test statistic is now 16.552 and with three degrees of freedom the p-value is 0.00087. The test continues to reject the exponential model. □

Sometimes, the test can be modified to fit different situations. The following example illustrates this for aggregate frequency data.

Example 13.8 *Conduct an approximate goodness-of-fit test for the Poisson model determined in Example* 12.60. *The data are repeated in Table* 13.10.

For each year we are assuming that the number of claims is the result of the sum of a number (given by the exposure) of independent and identical random variables. In that case the central limit theorem indicates that a normal approximation may be appropriate. The expected count (E_k) is the exposure times the estimated expected value for one exposure unit while the variance (V_k) is the exposure times the estimated variance for one exposure

Table 13.10 Automobile claims by year

Year	Exposure	Claims
1986	2,145	207
1987	2,452	227
1988	3,112	341
1989	3,458	335
1990	3,698	362
1991	3,872	359

unit. The test statistic is then

$$Q = \sum_k \frac{(n_k - E_k)^2}{V_k}$$

and has an approximate chi-square distribution with degrees of freedom equal to the number of data points less the number of estimated parameters. The expected count is $E_k = \lambda e_k$ and the variance is $V_k = \lambda e_k$ also. The test statistic is

$$\begin{aligned} Q &= \frac{(207-209.61)^2}{209.61} + \frac{(227-239.61)^2}{239.61} + \frac{(341-304.11)^2}{304.11} \\ &\quad + \frac{(335-337.92)^2}{337.92} + \frac{(362-361.37)^2}{361.37} + \frac{(359-378.38)^2}{378.38} \\ &= 6.19. \end{aligned}$$

With five degrees of freedom, the 5% critical value is 11.07 and the Poisson hypothesis is accepted. □

There is one important point to note about these tests. Suppose the sample size were to double but sampled values were not much different (imagine each number showing up twice instead of once). For the Kolmogorov–Smirnov test, the test statistic would be unchanged, but the critical value would be smaller. For the Anderson–Darling and chi-square tests, the test statistic would double while the critical value would be unchanged. As a result, for larger sample sizes, it is more likely that the null hypothesis (and thus the proposed model) will be rejected. This should not be surprising. We know that the null hypothesis is false (it is extremely unlikely that a simple distribution using a few parameters can explain the complex behavior that produced the observations) and with a large enough sample size we will have convincing evidence of that truth. When using these tests we must remember that, although all our models are wrong, some may be useful.

13.4.4 Likelihood ratio test

An alternative question to "Could the population have distribution A?" is "Is the population more likely to have distribution B than distribution A?" More formally:

H_0 : The data came from a population with distribution A.
H_1 : The data came from a population with distribution B.

In order to perform a formal hypothesis test distribution A must be a special case of distribution B, for example, exponential versus gamma. An easy way to complete this test is given below.

Definition 13.9 *The **likelihood ratio test** is conducted as follows. First, let the likelihood function be written as $L(\boldsymbol{\theta})$. Let $\boldsymbol{\theta}_0$ be the value of the parameters that maximizes the likelihood function. However, only values of the parameters that are within the null hypothesis may be considered. Let $L_0 = L(\boldsymbol{\theta}_0)$. Let $\boldsymbol{\theta}_1$ be the maximum likelihood estimator where the parameters can vary over all possible values from the alternative hypothesis and then let $L_1 = L(\boldsymbol{\theta}_1)$. The test statistic is $T = 2\ln(L_1/L_0) = 2(\ln L_1 - \ln L_0)$. The null hypothesis is rejected if $T > c$, where c is calculated from $\alpha = \Pr(T > c)$, where T has a chi-square distribution with degrees of freedom equal to the number of free parameters in the model from the alternative hypothesis less the number of free parameters in the model from the null hypothesis.*

This test makes some sense. When the alternative hypothesis is true, forcing the parameter to be selected from the null hypothesis should produce a likelihood value that is significantly smaller.

Example 13.10 *You want to test the hypothesis that the population that produced Data Set B (using the original largest observation) has a mean that is other than* 1,200. *Assume that the population has a gamma distribution and conduct the likelihood ratio test at a* 5% *significance level. Also, determine the p-value.*

The hypotheses are:

H_0 : gamma with $\mu = 1{,}200$.
H_1 : gamma with $\mu \neq 1{,}200$.

From earlier work the maximum likelihood estimates are $\hat{\alpha} = 0.55616$ and $\hat{\theta} = 2{,}561.1$. The loglikelihood at the maximum is $\ln L_1 = -162.293$. Next, the likelihood must be maximized, but only over those values α and θ for which $\alpha\theta = 1{,}200$. That means α can be free to range over all positive numbers but $\theta = 1{,}200/\alpha$. Thus, under the null hypothesis, there is only one free parameter. The likelihood function is maximized at $\hat{\alpha} = 0.54955$ and $\hat{\theta} = 2{,}183.6$. The loglikelihood at this maximum is $\ln L_0 = -162.466$.

Table 13.11 Six useful models for Example 13.11

Model	Number of parameters	Negative loglikelihood	χ^2	p-value
Negative binomial	2	5,348.04	8.77	0.0125
ZM logarithmic	2	5,343.79	4.92	0.1779
Poisson–inverse Gaussian	2	5,343.51	4.54	0.2091
ZM negative binomial	3	5,343.62	4.65	0.0979
Geometric–negative binomial	3	5,342.70	1.96	0.3754
Poisson–ETNB	3	5,342.51	2.75	0.2525

The test statistic is $T = 2(-162.293 + 162.466) = 0.346$. For a chi-square distribution with one degree of freedom, the critical value is 3.8415. Because $0.346 < 3.8415$, the null hypothesis is not rejected. The probability that a chi-square random variable with one degree of freedom exceeds 0.346 is 0.556, a p-value that indicates little support for the alternative hypothesis. □

Example 13.11 (Example 4.42 continued) *Members of the $(a, b, 0)$ class were not sufficient to describe these data. Determine a suitable model.*

Thirteen different distributions were fit to the data. The results of that process revealed six models with p-values above 0.01 for the chi-square goodness-of-fit test. Information about those models is given in Table 13.11. The likelihood ratio test indicates that the three-parameter model with the smallest negative loglikelihood (Poisson–ETNB) is not significantly better than the two-parameter Poisson–inverse Gaussian model. The latter appears to be an excellent choice. □

Example 13.12 *The estimated value of β_1 in Example* 12.65 *is small. Perform a likelihood ratio test using the beta model to see if age has a significant impact on losses.*

The parameters are reestimated forcing β_1 to be zero. When this is done, the estimates are $\hat{\beta}_2 = -0.79193$, $\hat{a} = 1.03118$, and $\hat{b} = 0.74249$. The value of the logarithm of the likelihood function is -4.2209. Adding age improves the likelihood by 0.0054, which is not significant. □

It is tempting to use this test when the alternative distribution simply has more parameters than the null distribution. In such cases the test is not appropriate. For example, it is possible for a two-parameter lognormal model to have a higher loglikelihood value than a three-parameter Burr model. This produces a negative test statistic, indicating that a chi-square distribution is not appropriate. When the null distribution is a limiting (rather than special) case of the alternative distribution, the test may still be used, but the test

statistic's distribution is now a mixture of chi-square distributions (see [120]). Regardless, it is still reasonable to use the "test" to make decisions in these cases, provided it is clearly understood that a formal hypothesis test was not conducted. Further examples and exercises using this test to make decisions appear in both the next section and the next chapter.

13.4.5 Exercises

13.4 Use the Kolmogorov–Smirnov test to see if a Weibull model is appropriate for the data used in Example 13.5.

13.5 (*) Five observations are made from a random variable. They are 1, 2, 3, 5, and 13. Determine the value of the Kolmogorov–Smirnov test statistic for the null hypothesis that $f(x) = 2x^{-2}e^{-2/x}$, $x > 0$.

13.6 (*) You are given the following five observations from a random sample: 0.1, 0.2, 0.5, 1.0, and 1.3. Calculate the Kolmogorov–Smirnov test statistic for the null hypothesis that the population density function is $f(x) = 2(1+x)^{-3}$, $x > 0$.

13.7 Perform the Anderson–Darling test of the Weibull distribution for Example 13.6.

13.8 Repeat Example 13.7 for the Weibull model.

13.9 (*) One hundred and fifty policyholders were observed from the time they arranged a viatical settlement until their death. No observations were censored. There were 21 deaths in the first year, 27 deaths in the second year, 39 deaths in the third year, and 63 deaths in the fourth year. The survival model

$$S(t) = 1 - \frac{t(t+1)}{20}, \quad 0 \le t \le 4,$$

is being considered. At a 5% significance level, conduct the chi-square goodness-of-fit test.

13.10 (*) Each day, for 365 days, the number of claims is recorded. The results were 50 days with no claims, 122 days with one claim, 101 days with two claims, 92 days with three claims, and no days with four or more claims. For a Poisson model determine the maximum likelihood estimate of λ and then perform the chi-square goodness-of-fit test at a 2.5% significance level.

13.11 (*) During a one-year period, the number of accidents per day was distributed as given in Table 13.12. Test the hypothesis that the data are from a Poisson distribution with mean 0.6 using the maximum number of groups such that each group has at least five expected observations. Use a significance level of 5%.

Table 13.12 Data for Exercise 13.11

No. of accidents	Days
0	209
1	111
2	33
3	7
4	3
5	2

13.12 Redo Example 13.8 assuming that each exposure unit has a geometric distribution. Conduct the approximate chi-square goodness-of-fit test. Is the geometric preferable to the Poisson model?

13.13 Using Data Set B (with the original largest value), determine if a gamma model is more appropriate than an exponential model. Recall that an exponential model is a gamma model with $\alpha = 1$. Useful values were obtained in Example 12.8.

13.14 Use Data Set C to choose a model for the population that produced those numbers. Choose from the exponential, gamma, and transformed gamma models. Information for the first two distributions was obtained in Example 12.9 and Exercise 12.21, respectively.

13.15 Conduct the chi-square goodness-of-fit test for each of the models obtained in Exercise 12.96.

13.16 Conduct the chi-square goodness-of-fit test for each of the models obtained in Exercise 12.98.

13.17 Conduct the chi-square goodness-of-fit test for each of the models obtained in Exercise 12.99.

13.18 For the data in Table 13.20 determine the method of moments estimates of the parameters of the Poisson–Poisson distribution where the secondary distribution is the ordinary (not zero-truncated) Poisson distribution. Perform the chi-square goodness-of-fit test using this model.

13.19 You are given the data in Table 13.13 which represent results from 23,589 automobile insurance policies. The third column headed "fitted model" represents the expected number of losses for a fitted (by maximum likelihood) negative binomial distribution.

(a) Perform the chi-squared goodness-of-fit test at a significance level of 5%.

Table 13.13 Data for Exercise 13.19

Number of losses, k	Number of policies, n_k	Fitted model
0	20,592	20,596.76
1	2,651	2,631.03
2	297	318.37
3	41	37.81
4	7	4.45
5	0	0.52
6	1	0.06
≥ 7	0	0.00

(b) Determine the maximum likelihood estimates of the negative binomial parameters r and β. This can be done from the given numbers without actually maximizing the likelihood function.

13.5 SELECTING A MODEL

13.5.1 Introduction

Almost all of the tools are in place for choosing a model. Before outlining a recommended approach, two important concepts must be introduced. The first is **parsimony**. The principle of parsimony states that unless there is considerable evidence to do otherwise a simpler model is preferred. The reason is that a complex model may do a great job of matching the data, but that is no guarantee the model matches the population from which the observations were sampled. For example, given any set of 10 (x, y) pairs with unique x values, there will always be a polynomial of degree 9 or less that goes through all 10 points. But if these points were a random sample, it is highly unlikely that the population values all lie on that polynomial. However, there may be a straight line that comes close to the sampled points as well as the other points in the population. This matches the spirit of most hypothesis tests. That is, do not reject the null hypothesis (and thus claim a more complex description of the population holds) unless there is strong evidence to do so.

The second concept does not have a name. It states that, if you try enough models, one will look good, even if it is not. Suppose I have 900 models at my disposal. For most data sets, it is likely that one of them will fit well, but this does not help us learn about the population.

Thus, in selecting models, there are two things to keep in mind:

1. Use a simple model if at all possible.

2. Restrict the universe of potential models.

The methods outlined in the remainder of this section will help with the first point. The second one requires some experience. Certain models make more sense in certain situations, but only experience can enhance the modeler's senses so that only a short list of quality candidates is considered.

The section is split into two types of selection criteria. The first set is based on the modeler's judgment while the second set is more formal in the sense that most of the time all analysts will reach the same conclusions. That is because the decisions are made based on numerical measurements rather than charts or graphs.

13.5.2 Judgment-based approaches

Using one's own judgment to select models involves one or more of the three concepts outlined below. In all cases, the analyst's experience is critical.

First, the decision can be based on the various graphs (or tables based on the graphs) presented in this chapter.[7] This allows the analyst to focus on aspects of the model that are important for the proposed application. For example, it may be more important to fit the tail well or it may be more important to match the mode or modes. Even if a score-based approach is used, it may be appropriate to present a convincing picture to support the chosen model.

Second, the decision can be influenced by the success of particular models in similar situations or the value of a particular model for its intended use. For example, the 1941 CSO mortality table follows a Makeham distribution for much of its range of ages. In a time of limited computing power, such a distribution allowed for easier calculation of joint life values. As long as the fit of this model was reasonable, this advantage outweighed the use of a different, but better fitting, model. Similarly, if the Pareto distribution has been used to model a particular line of liability insurance both by the analyst's company and by others, it may require more than the usual amount of evidence to change to an alternative distribution.

Third, the situation may completely determine the distribution. For example, suppose a dental insurance contract provides for at most two check-ups per year and suppose that individuals make two independent choices each year as to whether or not to have a check-up. If each time the probability is q, then the distribution must be binomial with $m = 2$.

Finally, it should be noted that the more algorithmic approaches outlined below do not always agree. In that case judgment is most definitely required, if only to decide which algorithmic approach to use.

[7] Besides the ones discussed here, there are other plots/tables that could be used. Other choices are a q-q plot and a comparison of model and empirical limited expected values or mean residual life functions.

13.5.3 Score-based approaches

Some analysts might prefer an automated process for selecting a model. An easy way to do that would be to assign a score to each model and the model with the best value wins. The following scores are worth considering:

1. Lowest value of the Kolmogorov–Smirnov test statistic.
2. Lowest value of the Anderson–Darling test statistic.
3. Lowest value of the chi-square goodness-of-fit test statistic.
4. Highest p-value for the chi-square goodness-of-fit test.
5. Highest value of the likelihood function at its maximum.

All but the chi-square p-value have a deficiency with respect to parsimony. First, consider the likelihood function. When comparing, say, an exponential to a Weibull model, the Weibull model must have a likelihood value that is at least as large as the exponential model. They would only be equal in the rare case that the maximum likelihood estimate of the Weibull parameter τ is equal to 1. Thus, the Weibull model would always win over the exponential model, a clear violation of the principle of parsimony. For the three test statistics, there is no assurance that the same relationship will hold, but it seems likely that, if a more complex model is selected, the fit measure is likely to be better. The only reason the p-value is immune from this problem is that with more complex models the test has fewer degrees of freedom. It is then possible that the more complex model will have a smaller p-value. There is no comparable adjustment for the first two test statistics listed.

With regard to the likelihood value, there are two ways to proceed. One is to perform the likelihood ratio test and the other is to extract a penalty for employing additional parameters. The likelihood ratio test is technically only available when one model is a special case of another (for example, Pareto vs generalized Pareto). The concept can be turned into an algorithm by using the test at a 5% significance level. Begin with the best one-parameter model (the one with the highest loglikelihood value). Add a second parameter only if the two-parameter model with the highest loglikelihood value shows an increase of at least 1.92 (so twice the difference exceeds the critical value of 3.84). Then move to three-parameter models. If the comparison is to a two-parameter model, a 1.92 increase is again needed. If the early comparison led to keeping the one-parameter model, an increase of 3.00 is needed (because the test has two degrees of freedom). To add three parameters requires a 3.91 increase, four parameters a 4.74 increase, and so on. In the spirit of this chapter, this algorithm can be used even for nonspecial cases. However, it would not be appropriate to claim that a likelihood ratio test was being conducted.

Aside from the issue of special cases, the likelihood ratio test has the same problem as the other hypothesis tests. Were the sample size to double, the loglikelihoods would also double, making it more likely that a model with a higher number of parameters will be selected. This tends to defeat the parsimony principle. On the other hand, it could be argued that, if we possess a lot of data, we have the right to consider and fit more complex models. A method that effects a compromise between these positions is the Schwarz Bayesian criterion (SBC) [121], which recommends that when ranking models a deduction of $(r/2)\ln n$ should be made from the loglikelihood value, where r is the number of estimated parameters and n is the sample size.[8] Thus, adding a parameter requires an increase of $0.5\ln n$ in the loglikelihood. For larger sample sizes, a greater increase is needed, but it is not proportional to the sample size itself.[9]

Example 13.13 *For the continuing example in this chapter, choose between the exponential and Weibull models for the data.*

Graphs were constructed in the various examples and exercises. Table 13.14 summarizes the numerical measures. For the truncated version of Data Set B, the SBC is calculated for a sample size of 19, while for the version censored at 1,000 there are 20 observations. For both versions of Data Set B, while the Weibull offers some improvement, it is not convincing. In particular, neither the likelihood ratio test nor the SBC indicates value in the second parameter. For Data Set C it is clear that the Weibull model is superior and provides an excellent fit. □

Example 13.14 *In Example* 4.57 *an ad hoc method was used to demonstrate that the Poisson–ETNB distribution provided a good fit. Use the methods of this chapter to determine a good model.*

The data set is very large and, as a result, requires a very close correspondence of the model to the data. The results are given in Table 13.15.

From Table 13.15, it is seen that the negative binomial distribution does not fit well while the fit of the Poisson–inverse Gaussian is marginal at best ($p = 2.88\%$). The Poisson–inverse Gaussian is a special case ($r = -0.5$) of the Poisson–ETNB. Hence, a likelihood ratio test can be formally applied to determine if the additional parameter r is justified. Because the loglikelihood increases by 5, which is more than 1.92, the three-parameter model is a significantly better fit. The chi-square test shows that the Poisson-ETNB provides an adequate fit. On the other hand, the SBC favors the Poisson-inverse Gaussian distribution. Given the improved fit in the tail for the three parameter model, it seems to be the best choice. □

[8] In the first edition not only was Schwarz' name misspelled, but the formula for the penalty was incorrect. This edition has the correct version.

[9] There are other information-based decision rules. Section 3 of Brockett [17] promotes the Akaike information criterion. In a discussion to that paper, Carlin provides support for the SBC.

Table 13.14 Results for Example 13.13

	B truncated at 50		B censored at 1,000	
Criterion	Exponential	Weibull	Exponential	Weibull
K-S*	0.1340	0.0887	0.0991	0.0991
A-D*	0.4292	0.1631	0.1713	0.1712
χ^2	1.4034	0.3615	0.5951	0.5947
p-value	0.8436	0.9481	0.8976	0.7428
Loglikelihood	−146.063	−145.683	−113.647	−113.647
SBC	−147.535	−148.628	−115.145	−116.643
	C			
χ^2	61.913	0.3698		
p-value	10^{-12}	0.9464		
Loglikelihood	−214.924	−202.077		
SBC	−217.350	−206.929		

*K-S and A-D refer to the Kolmogorov–Smirnov and Anderson–Darling test statistics, respectively.

Example 13.15 *The following example is taken from Douglas* [29], p. 253. *An insurance company's records for one year show the number of accidents per day which resulted in a claim to the insurance company for a particular insurance coverage. The results are in Table* 13.16. *Determine if a Poisson model is appropriate.*

A Poisson model is fitted to these data. The method of moments and the maximum likelihood method both lead to the estimate of the mean,

$$\hat{\lambda} = \frac{742}{365} = 2.0329.$$

The results of a chi-square goodness-of-fit test are in Table 13.17. Any time such a table is made, the expected count for the last group is

$$E_{k+} = n\hat{p}_{k+} = n(1 - \hat{p}_0 - \cdots - \hat{p}_{k-1}).$$

The last three groups were combined to ensure an expected count of at least one for each row. The test statistic is 9.93 with six degrees of freedom. The critical value at a 5% significance level is 12.59 and the p-value is 0.1277. By this test the Poisson distribution is an acceptable model; however, it should be noted that the fit is poorest at the large values, and with the model understating the observed values, this may be a risky choice. □

Example 13.16 *The data set in Table* 12.13 *come from Beard et al.* [12] *and were previously analyzed in Example* 12.58. *Determine a model that adequately describes the data.*

Table 13.15 Results for Example 13.14

		Fitted distributions		
No. of claims	Observed frequency	Negative binomial	Poisson–inverse Gaussian	Poisson–ETNB
0	565,664	565,708.1	565,712.4	565,661.2
1	68,714	68,570.0	68,575.6	68,721.2
2	5,177	5,317.2	5,295.9	5,171.7
3	365	334.9	344.0	362.9
4	24	18.7	20.8	29.6
5	6	1.0	1.2	3.0
6+	0	0.0	0.1	0.4
Parameters		$\beta = 0.0350662$ $r = 3.57784$	$\lambda = 0.123304$ $\beta = 0.0712027$	$\lambda = 0.123395$ $\beta = 0.233862$ $r = -0.846872$
Chi square		12.13	7.09	0.29
Degrees of freedom		2	2	1
p-value		<1%	2.88%	58.9%
−Loglikelihood		251,117	251,114	251,109
SBC		−251,130	−251,127	−251,129

Table 13.16 Data for Example 13.15

No. of claims/day	Observed no. of days
0	47
1	97
2	109
3	62
4	25
5	16
6	4
7	3
8	2
9+	0

Parameter estimates from fitting four models are in Table 12.13. Various fit measures are given in Table 13.18. Only the zero-modified geometric distribution passes the goodness-of-fit test. It is also clearly superior according to the SBC. A likelihood ratio test against the geometric has a test statistic of 2(171,479 − 171,133) = 692, which with one degree of freedom is clearly significant. This confirms the qualitative conclusion in Example 12.58. □

Table 13.17 Chi-square goodness-of-fit test for Example 13.15

Claims/day	Observed	Expected	Chi square
0	47	47.8	0.01
1	97	97.2	0.00
2	109	98.8	1.06
3	62	66.9	0.36
4	25	34.0	2.39
5	16	13.8	0.34
6	4	4.7	0.10
7+	5	1.8	5.66
Totals	365	365	9.93

Table 13.18 Test results for Example 13.16

	Poisson	Geometric	ZM Poisson	ZM geometric
Chi square	543.0	643.4	64.8	0.58
Degrees of freedom	2	4	2	2
p-value	$< 1\%$	$< 1\%$	$< 1\%$	74.9%
Loglikelihood	−171,373	−171,479	−171,160	−171,133
SBC	−171,379.5	−171,485.5	−171,173	−171,146

Example 13.17 *The data in Table* 13.19, *from Simon* [122], *represent the observed number of claims per contract for* 298 *contracts. Determine an appropriate model.*

The Poisson, negative binomial, and Polya–Aeppli distributions are fitted to the data. The Polya–Aeppli and the negative binomial are both plausible distributions. The p-value of the chi-square statistic and the loglikelihood both indicate that the Polya–Aeppli is slightly better than the negative binomial. The SBC verifies that both models are superior to the Poisson distribution. The ultimate choice may depend on familiarity, prior use, and computational convenience of the negative binomial versus the Polya–Aeppli model. □

Example 13.18 *Consider the data in Table* 13.20 *on automobile liability policies in Switzerland taken from Bühlmann* [19]. *Determine an appropriate model.*

Three models are considered in Table 13.20. The Poisson distribution is a very bad fit. Its tail is far too light compared with the actual experience. The negative binomial distribution appears to be much better but cannot be accepted because the p-value of the chi-square statistic is very small. The large sample size requires a better fit. The Poisson–inverse Gaussian distribution

Table 13.19 Fit of Simon data

		Fitted distributions		
Number of claims/contract	Number of contracts	Poisson	Negative binomial	Polya–Aeppli
0	99	54.0	95.9	98.7
1	65	92.2	75.8	70.6
2	57	78.8	50.4	50.2
3	35	44.9	31.3	32.6
4	20	19.2	18.8	20.0
5	10	6.5	11.0	11.7
6	4	1.9	6.4	6.6
7	0	0.5	3.7	3.6
8	3	0.1	2.1	2.0
9	4	0.0	1.2	1.0
10	0	0.0	0.7	0.5
11	1	0.0	0.4	0.3
12+	0	0.0	0.5	0.3
Parameters		$\lambda = 1.70805$	$\beta = 1.15907$ $r = 1.47364$	$\lambda = 1.10551$ $\beta = 0.545039$
Chi square		72.64	4.06	2.84
Degrees of freedom		4	5	5
p-Value		<1%	54.05%	72.39%
Loglikelihood		−577.0	−528.8	−528.5
SBC		−579.8	−534.5	−534.2

provides an almost perfect fit (p-value is large). Note that the Poisson–inverse Gaussian has two parameters, like the negative binomial. The SBC also favors this choice. This example shows that the Poisson–inverse Gaussian can have a much heavier right-hand tail than the negative binomial. □

Example 13.19 *Comprehensive medical claims were studied by Bevan* [15] *in* 1963. *Male* (955 *payments*) *and female* (1,291 *payments*) *claims were studied separately. The data appear in Table* 13.21 *where there was a deductible of* 25. *Can a common model be used?*

When using the combined data set the lognormal distribution is the best two-parameter model. Its negative loglikelihood (NLL) is 4,580.20. This is 19.09 better than the one-parameter inverse exponential model and 0.13 worse than the three-parameter Burr model. Because none of these models is a special case of the other, the likelihood ratio test (LRT) cannot be used, but it is clear that using the 1.92 difference as a standard, the lognormal is preferred. The SBC requires an improvement of $0.5 \ln(2{,}246) = 3.86$ and again

Table 13.20 Fit of Buhlmann data

No. of accidents	Observed frequency	Fitted distributions		
		Poisson	Negative binomial	P.–i.G.[a]
0	103,704	102,629.6	103,723.6	103,710.0
1	14,075	15,922.0	13,989.9	14,054.7
2	1,766	1,235.1	1,857.1	1,784.9
3	255	63.9	245.2	254.5
4	45	2.5	32.3	40.4
5	6	0.1	4.2	6.9
6	2	0.0	0.6	1.3
7+	0	0.0	0.1	0.3
Parameters		$\lambda = 0.155140$	$\beta = 0.150232$ $r = 1.03267$	$\lambda = 0.144667$ $\beta = 0.310536$
Chi square		1,332.3	12.12	0.78
Degrees of freedom		2	2	3
p-Values		<1%	<1%	85.5%
Loglikelihood		−55,108.5	−54,615.3	−54,609.8
SBC		−55,114.3	−54,627.0	−54,621.5

[a]P.–i.G. stands for Poisson–inverse Gaussian.

the lognormal is preferred. The parameters are $\mu = 4.5237$ and $\sigma = 1.4950$. When separate lognormal models are fit to males ($\mu = 3.9686$ and $\sigma = 1.8432$) and females ($\mu = 4.7713$ and $\sigma = 1.2848$), the respective NLLs are 1,977.25 and 2,583.82 for a total of 4,561.07. This is an improvement of 19.13 over a common lognormal model, which is significant by both the LRT (3.00 needed) and SBC (7.72 needed). Sometimes it is useful to be able to use the same nonscale parameter in both models. When a common value of σ is used, the NLL is 4,579.77, which is significantly worse than using separate models. □

Example 13.20 *In* 1958 *Longley-Cook* [86] *examined employment patterns of casualty actuaries. One of his tables listed the number of members of the Casualty Actuarial Society employed by casualty companies in* 1949 (55 *actuaries*) *and* 1957 (78 *actuaries*). *Using the data in Table* 13.22 *determine a model for the number of actuaries per company which employs at least one actuary and find out whether the distribution has changed over the eight-year period.*

Because a value of zero is impossible, only zero-trucated distributions should be considered. In all three cases (1949 data only, 1957 data only, combined data) the ZT logarithmic and ZT (extended) negative binomial distributions have acceptable goodness-of-fit test values. The improvement in NLL is 0.52, 0.02, and 0.94. The LRT can be applied (except that the ZT

Table 13.21 Comprehensive medical losses for Example 13.19

Loss	Male	Female
25–50	184	199
50–100	270	310
100–200	160	262
200–300	88	163
300–400	63	103
400–500	47	69
500–1,000	61	124
1,000–2,000	35	40
2,000–3,000	18	12
3,000–4,000	13	4
4,000–5,000	2	1
5,000–6,667	5	2
6,667–7,500	3	1
7,500–10,000	6	1

Table 13.22 Number of actuaries per company for Example 13.20

Number of actuaries	Number of companies—1949	Number of companies—1957
1	17	23
2	7	7
3–4	3	3
5–9	2	3
10+	0	1

logarithmic distribution is a limiting case of the ZT negative binomial distribution with $r \to 0$), and the improvement is not significant in any of the cases. The same conclusions apply if the SBC is used. The parameter estimates (where β is the only parameter) are 2.0227, 2.8114, and 2.4479, respectively. The NLL for the combined data set is 74.35 while the total for the two separate models is 74.15. The improvement is only 0.20, which is not significant (there is one degree of freedom). Even though the estimated mean has increased from $2.0227/\ln(3.0227) = 1.8286$ to $2.8114/\ln(3.8114) = 2.1012$, there is not enough data to make a convincing case that the true mean has increased. □

13.5.4 Exercises

13.20 (*) One thousand policies were sampled and the number of accidents for each recorded. The results are in Table 13.23. Without doing any for-

Table 13.23 Data for Exercise 13.20

No. of accidents	No. of policies
0	100
1	267
2	311
3	208
4	87
5	23
6	4
Total	1,000

Table 13.24 Results for Exercise 13.23

Model	No. of parameters	Negative loglikelihood
Generalized Pareto	3	219.1
Burr	3	219.2
Pareto	2	221.2
Lognormal	2	221.4
Inverse exponential	1	224.3

mal tests, determine which of the following five models is most appropriate: binomial, Poisson, negative binomial, normal, gamma.

13.21 For Example 13.1, determine if a transformed gamma model is more appropriate than either the exponential model or the Weibull model for each of the three data sets.

13.22 (*) From the data in Exercise 13.11 the maximum likelihood estimates are $\hat{\lambda} = 0.60$ for the Poisson distribution and $\hat{r} = 2.9$ and $\hat{\beta} = 0.21$ for the negative binomial distribution. Conduct the likelihood ratio test for choosing between these two models.

13.23 (*) From a sample of size 100, five models are fit with the results given in Table 13.24. Use the Schwarz Bayesian criterion to select the best model.

13.24 This is a continuation of Exercise 12.38. Use both the likelihood ratio test (at a 5% significance level) and the Schwarz Bayesian criterion to decide if Sylvia's claim is true.

13.25 Using the results from Exercises 12.96 and 13.15, use the chi-square goodness-of-fit test, the likelihood ratio test, and the Schwarz Bayesian criterion to determine the best model from the members of the $(a, b, 0)$ class.

Table 13.25 Data for Exercise 13.28

No. of medical claims	No. of accidents
0	529
1	146
2	169
3	137
4	99
5	87
6	41
7	25
8+	0

13.26 Using the results from Exercises 12.98 and 13.16, use the chi-square goodness-of-fit test, the likelihood ratio test, and the Schwarz Bayesian criterion to determine the best model from the members of the $(a, b, 0)$ class.

13.27 Using the results from Exercises 12.99 and 13.17, use the chi-square goodness-of-fit test, the likelihood ratio test, and the Schwarz Bayesian criterion to determine the best model from the members of the $(a, b, 0)$ class.

13.28 Table 13.25 gives the number of medical claims per reported automobile accident.

(a) Construct a plot similar to Figure 4.8. Does it appear that a member of the $(a, b, 0)$ class will provide a good model? If so, which one?

(b) Determine the maximum likelihood estimates of the parameters for each member of the $(a, b, 0)$ class.

(c) Based on the chi-square goodness-of-fit test, the likelihood ratio test, and the Schwarz Bayesian criterion, which member of the $(a, b, 0)$ class provides the best fit? Is this model acceptable?

13.29 For the four data sets introduced in Exercises 12.96, 12.98, 12.99, and 13.28, you have determined the best model from among members of the $(a, b, 0)$ class. For each data set determine the maximum likelihood estimates of the zero-modified Poisson, geometric, logarithmic, and negative binomial distributions. Use the chi-square goodness-of-fit test and likelihood ratio tests to determine the best of the eight models considered and state whether or not the selected model is acceptable.

13.30 A frequency model that has not been mentioned to this point is the **zeta distribution**. It is a zero-truncated distribution with $p_k^T = k^{-(\rho+1)}/\zeta(\rho+1)$, $k = 1, 2, \ldots, \rho > 0$. The denominator is the zeta function, which must be

Table 13.26 Data for Excercise 13.32(a)

No. of claims	No. of policies
0	96,978
1	9,240
2	704
3	43
4	9
5+	0

evaluated numerically as $\zeta(\rho+1) = \sum_{k=1}^{\infty} k^{-(\rho+1)}$. The zero-modified zeta distribution can be formed in the usual way. More information can be found in Luong and Doray [88].

(a) Determine the maximum likelihood estimates of the parameters of the zero-modified zeta distribution for the data in Example 12.58.

(b) Is the zero-modified zeta distribution acceptable?

13.31 In Exercise 13.29 the best model from among the members of the $(a, b, 0)$ and $(a, b, 1)$ classes was selected for the data sets in Exercises 12.96, 12.98, 12.99, and 13.28. Fit the Poisson–Poisson, Polya–Aeppli, Poisson–inverse Gaussian, and Poisson–ETNB distributions to these data and determine if any of these distributions should replace the one selected in Exercise 13.29. Is the current best model acceptable?

13.32 The five data sets presented in this problem are all taken from Lemaire [82]. For each data set compute the first three moments and then use the ideas in Section 4.6.8 to make a guess at an appropriate model from among the compound Poisson collection (Poisson, geometric, negative binomial, Poisson–binomial (with $m = 2$ and $m = 3$), Polya–Aeppli, Neyman Type A, Poisson–inverse Gaussian, and Poisson–ETNB). From the selected model (if any) and members of the $(a, b, 0)$ and $(a, b, 1)$ classes, determine the best model.

(a) The data in Table 13.26 represent counts from third-party automobile liability coverage in Belgium.

(b) The data in Table 13.27 represent the number of deaths due to horse kicks in the Prussian army between 1875 and 1894. The counts are the number of deaths in a corps (there were 10 of them) in a given year, and thus there are 200 observations. This data set is often cited as the inspiration for the Poisson distribution. For using any of our models, what additional assumption about the data must be made?

(c) The data in Table 13.28 represent the number of major international wars per year from 1500 through 1931.

Table 13.27 Data for Excercise 13.32(b)

No. of deaths	No. of corps
0	109
1	65
2	22
3	3
4	1
5+	0

Table 13.28 Data for Excercise 13.32(c)

No. of wars	No. of years
0	223
1	142
2	48
3	15
4	4
5+	0

Table 13.29 Data for Excercise 13.32(d)

No. of runs	No. of half innings
0	1,023
1	222
2	87
3	32
4	18
5	11
6	6
7+	3

(d) The data in Table 13.29 represent the number of runs scored in each half-inning of World Series baseball games played from 1947 through 1960.

(e) The data in Table 13.30 represent the number of goals per game per team in the 1966–1967 season of the National Hockey League.

13.33 Verify that the estimates presented in Example 4.64 are the maximum likelihood estimates. (Because only two decimals are presented, it is probably sufficient to observe that the likelihood function takes on smaller values at

Table 13.30 Data for Excercise 13.32(e)

No. of goals	No. of games
0	29
1	71
2	82
3	89
4	65
5	45
6	24
7	7
8	4
9	1
10+	3

each of the nearby points.) The negative binomial distribution was fit to these data in Example 12.56. Which of these two models is preferable?

14 Five examples

14.1 INTRODUCTION

In this chapter we present five examples that illustrate many of the concepts discussed to this point. The first is a model for the time to death. The second model is for the time from when a medical malpractice incident occurs to when it is reported. The third model is for the amount of a liability payment. This model is also continuous but most likely has a decreasing failure rate (typical of payment amount variables). On the other hand, time to event variables tend to have an increasing failure rate. The last two examples add aggregate loss calculations from Chapter 6 to the mix.

14.2 TIME TO DEATH

14.2.1 The data

A variety of mortality tables are available from the Society of Actuaries at www.soa.org. The typical mortality table provides values of the survival function at each whole-number age at death. Table 14.1 represents female mortality in 1900, with only some of the data points presented. It is followed by

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

Table 14.1 1900 female mortality

x	$S(x)$	x	$S(x)$	x	$S(x)$
0	1.000	35	0.681	75	0.233
1	0.880	40	0.650	80	0.140
5	0.814	45	0.617	85	0.062
10	0.796	50	0.580	90	0.020
15	0.783	55	0.534	95	0.003
20	0.766	60	0.478	100	0.000
25	0.739	65	0.410		
30	0.711	70	0.328		

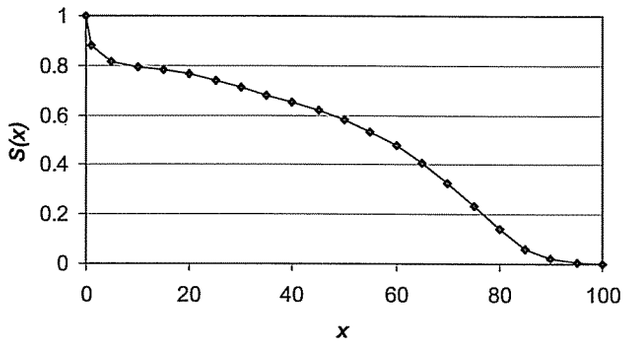


Fig. 14.1 Survival function for Society of Actuaries data.

Figure 14.1, a graph of the survival function obtained by connecting the given points with straight lines.

The mean residual life function can be obtained by assuming that the survival function is indeed a straight line connecting each of the available points. From (3.5) it can be computed as the area under the curve beyond the given age divided by the value of the survival function at that age. Figure 14.2 contains a plot of the mean residual life function. The slight increase shortly after birth indicates that in 1900 infant mortality was high. Surviving the first year after birth adds about five years to one's expected remaining lifetime. After that, the mean residual life steadily decreases, which is the effect of aging that we would have expected.

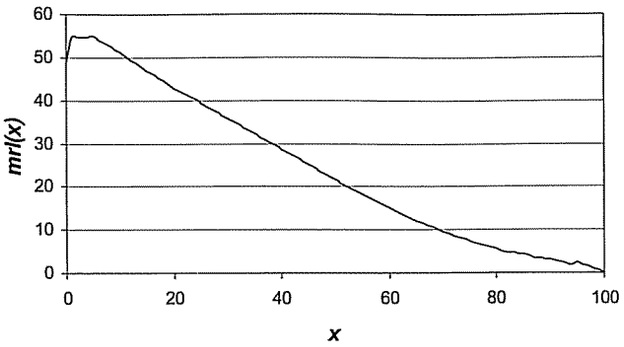


Fig. 14.2 Mean residual life function for Society of Actuaries data.

14.2.2 Some calculations

Items such as deductibles, limits, and coinsurances are not particularly interesting with regard to insurances on human lifetimes. We will consider the following two questions:

1. For a person age 65, determine the expected present value of providing 1,000 at the beginning of each year in which the person is alive. The interest rate is 6%.

2. For a person age 20, determine the expected present value of providing 1,000 at the moment of death. The interest rate is 6%.

For the first problem, the present value random variable Y can be written as $Y = 1{,}000(Y_0 + \cdots + Y_{34})$, where Y_j is the present value of that part of the benefit that pays 1 at age $65 + j$ if the person is alive at that time. Then,

$$Y_j = \begin{cases} 1.06^{-j} & \text{with probability } \dfrac{S(65+j)}{S(65)}, \\ 0 & \text{with probability } 1 - \dfrac{S(65+j)}{S(65)}. \end{cases}$$

The answer is then

$$\begin{aligned} \mathrm{E}(Y) &= 1{,}000 \sum_{j=0}^{34} \frac{1.06^{-j} S(65+j)}{0.410} \\ &= 8{,}408.07, \end{aligned}$$

where linear interpolation was used for intermediate values of the survival function.

For the second problem, let $Z = 1{,}000(1.06^{-T})$ be the present value random variable, where T is the time in years to death of the 20-year old. The calculation is

$$\mathrm{E}(Z) = 1{,}000 \int_0^{80} \frac{1.06^{-t} f(20+t)}{S(20)} dt.$$

When linear interpolation is used to obtain the survival function at intermediate ages, the density function becomes the slope. That is, if x is a multiple of 5, then

$$f(t) = \frac{S(x) - S(x+5)}{5}, \quad x < t < x+5.$$

Breaking the range of integration into 16 pieces gives

$$\begin{aligned} \mathrm{E}(Z) &= \frac{1000}{0.766} \sum_{j=0}^{15} \frac{S(20+5j) - S(25+5j)}{5} \int_{5j}^{5+5j} 1.06^{-t} dt \\ &= \frac{200}{0.766} \sum_{j=0}^{15} [S(20+5j) - S(25+5j)] \frac{1.06^{-5j} - 1.06^{-5-5j}}{\ln 1.06} \\ &= 155.10. \end{aligned}$$

While it is unusual for a parametric model to be used, we will do so anyway. Consider the Makeham distribution with hazard rate function $h(x) = A + Bc^x$. Then

$$S(x) = \exp\left[-Ax - \frac{B(c^x - 1)}{\ln c}\right].$$

Maximum likelihood estimation cannot be used because no sample size is given. Because it is unlikely that this model will be effective below age 20, only information beyond that age will be used. Assume that that there were 1,000 lives at age 0 who died according to the survival function in Table 14.1. Then, for example, the contribution to the likelihood function for the interval from age 30 to 35 is $30 \ln\{[S(30) - S(35)]/S(20)\}$ with the survival function using the Makeham distribution. The sample size comes from $1{,}000(0.711 - 0.681)$ with these survival function values taken from the "data."[1] The values that maximize this likelihood function are $\hat{A} = 0.006698$, $\hat{B} = 0.00007976$, and $\hat{c} = 1.09563$. In Figure 14.3 the diamonds represent the "data" and the solid curve is the Makeham survival function (both have been conditioned on being alive at age 20). The fit is almost too good, suggesting that perhaps this mortality table was already smoothed to follow a Makeham distribution at adult ages.

The same calculations can be done. For the annuity, no interpolation is needed because the Makeham function provides the survival function values

[1] Aside from not knowing the sample size, the values in Table 14.1 are probably not random observations. It is possible the values in the table were smoothed using techniques of the kind discussed in Chapter 15.

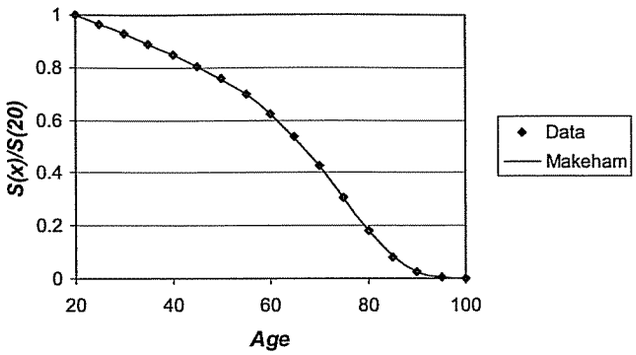


Fig. 14.3 Comparison of "data" and Makeham model.

at each age. The answer is 8,405.24. For the insurance, it is difficult to do the integral analytically. Linear interpolation was used between integral ages to produce an answer of 154.90. The agreement with the answers obtained earlier is not surprising.

14.2.3 Exercise

14.1 From ages 5 through 100 the mean residual life function is essentially linear. Because insurances are rarely sold under age 5, it would be reasonable to extend the graph linearly back to 0. Then a reasonable approximation is $e(x) = 60 - 0.6x$. From this, determine the density and survival function for the age at death and then use this function to solve the two problems.

14.3 TIME FROM INCIDENCE TO REPORT

Consider an insurance contract that provides payment when a certain event (such as death, disability, fire) occurs. There are three key dates. The first is when the event occurs, the second is when it is reported to the insurance company, and the third is when the claim is settled. The time between these dates is important because it affects the amount of interest that can be earned on the premium prior to paying the claim and because it provides a mechanism for estimating unreported claims. This example concerns the time from incidence to report. The particular example used here is based on a paper by Accomando and Weissner [4].

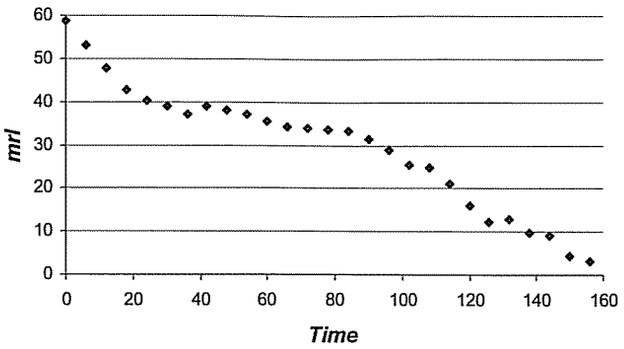


Fig. 14.4 Mean residual life function for report lag data.

14.3.1 The problem and some data

This example concerns medical malpractice claims that occurred in a particular year. One hundred sixty-eight months after the beginning of the year under study, there have been 463 claims reported that were known to have occurred in that year. The distribution of the times from occurrence to report (by month in six month intervals) is given in Table 14.2. A graph of the mean residual life function appears in Figure 14.4.[2]

Your task is to fit a model to these observations and then use the model to estimate the total number of claims that occurred in the year under study. A look at the mean residual life function indicates a decreasing pattern and so a lighter than exponential tail is expected. A Weibull model can have such a tail and so can be used here.

14.3.2 Analysis

Using maximum likelihood to estimate the Weibull parameters, the result is $\hat{\tau} = 1.71268$ and $\hat{\theta} = 67.3002$. According to the Weibull distribution, the probability that a claim is reported by time 168 is

$$F(168) = 1 - e^{-(168/\theta)^{\tau}}.$$

If N is the unknown total number of claims, the number observed by time 168 is the result of binomial sampling, and thus on an expected value basis

[2]Because of the right truncation of the data, there are some items missing for calculation of the mean residual life. It is not clear from the data what the effect will be. This picture gives a guide, but the model ultimately selected should both fit the data and be reasonable based on the analyst's experience and judgment.

Table 14.2 Medical malpractice report lags

Lag in months	No. of claims	Lag in months	No. of claims
0–6	4	84–90	11
6–12	6	90–96	9
12–18	8	96–102	7
18–24	38	102–108	13
24–30	45	108–114	5
30–36	36	114–120	2
36–42	62	120–126	7
42–48	33	126–132	17
48–54	29	132–138	5
54–60	24	138–144	8
60–66	22	144–150	2
66–72	24	150–156	6
72–78	21	156–162	2
78–84	17	162–168	0

we obtain

$$\text{Expected number of reported claims by time } 168 = N[1 - e^{-(168/\theta)^{\tau}}].$$

Setting this expectation equal to the observed number reported of 463 and then solving for N yields

$$N = \frac{463}{1 - e^{-(168/\theta)^{\tau}}}.$$

Inserting the parameter estimates yields the value 466.88. Thus, after 14 years, we expect to have about four more claims reported.

The delta method (Theorem 12.17) can be used to produce a 95% confidence interval. It is 466.88 ± 2.90, indicating that there could reasonably be between one and seven additional claims reported.

14.4 PAYMENT AMOUNT

You are the consulting actuary for a reinsurer and have been asked to determine the expected cost and the risk (as measured by the coefficient of variation) for various coverages. To help you out, losses from 200 claims have been supplied. The reinsurer also estimates (and you may confidently rely on its estimate) that there will be 21 losses per year and the number of losses has a Poisson distribution. The coverages it is interested in are full coverage, 1 million excess of 250,000, and 2 million excess of 500,000. The phrase "z excess of y" is to be interpreted as $d = y$ and $u = y + z$ in the notation of Theorem 5.13.

Table 14.3 Losses up to 200 (thousand)

Loss range (thousands)	Number of losses	Loss range (thousands)	Number of losses
1–5	3	41–50	19
6–10	12	51–75	28
11–15	14	76–100	21
16–20	9	101–125	15
21–25	7	126–150	10
26–30	7	151–200	15
31–40	18		

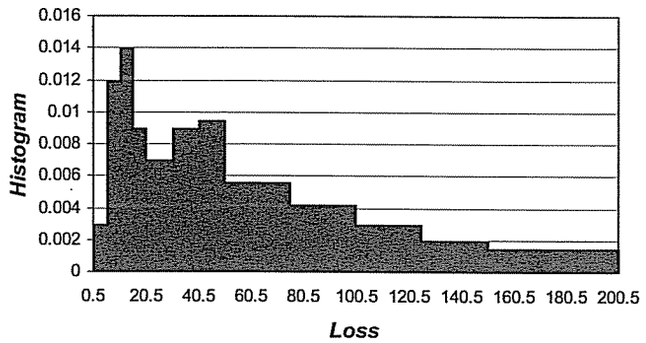


Fig. 14.5 Histogram of losses.

14.4.1 The data

One hundred seventy-eight losses that were 200,000 or below (all expressed in whole numbers of thousands of dollars) that were supplied are summarized in Table 14.3. In addition, there were 22 losses in excess of 200. They are listed below:

206	219	230	235	241	272	283	286	312	319	385
427	434	555	562	584	700	711	869	980	999	1506

Finally, the 178 losses in the table sum to 11,398 and their squares sum to 1,143,164.

To get a feel for the data in the table, the histogram in Figure 14.5 was constructed. Keep in mind that the height of a histogram bar is the count in the cell divided by the sample size (200) and then further divided by the interval width. Therefore, the first bar has a height of $3/[200(5)] = 0.003$.

Table 14.4 Mean residual life for losses above 200 (thousand)

Loss	Mean residual life
200	314
300	367
400	357
500	330
600	361
700	313
800	289
900	262

It can be seen from the histogram that the underlying distribution has a nonzero mode. To check the tail, we can compute the empirical mean residual life function at a number of values. They are presented in Table 14.4. The function appears to be fairly constant and so an exponential model seems reasonable.

14.4.2 The first model

A two-component spliced model was selected. The empirical model is used through 200 (thousand) and an exponential model thereafter. There are (at least) two ways to choose the exponential model. One is to restrict the parameter by forcing the distribution to place 11% (22 out of 200) of probability at points above 200. The other option is to estimate the exponential model independent of the 11% requirement and then multiply the density function to make the area above 200 be 0.11. The latter was selected and the resulting parameter estimate is $\theta = 314$. For values below 200, the empirical distribution places probability 1/200 at each observed value. The resulting exponential density function (for $x > 200$) is

$$f(x) = 0.000662344e^{-x/314}.$$

For a coverage that pays all losses, the kth moment is (where the 200 losses in the sample have been ordered from smallest to largest)

$$\mathrm{E}(X^k) = \frac{1}{200}\sum_{j=1}^{178} x_j^k + \int_{200}^{\infty} x^k f(x)dx.$$

Then,

$$
\begin{aligned}
\mathrm{E}(X) &= \frac{11{,}398}{200} + 0.000662344[314(200) + 314^2]e^{-200/314} = 113.53, \\
\mathrm{E}(X^2) &= \frac{1{,}143{,}164}{200} \\
&\quad +0.000662344[314(200)^2 + 2(314)^2(200) + 2(314)^3]e^{-200/314} \\
&= 45{,}622.93.
\end{aligned}
$$

The variance is $45{,}622.93 - 113.53^2 = 32{,}733.87$ for a coefficient of variation of 1.59. However, these are for one loss only. The distribution of annual losses follows a compound Poisson distribution. The mean is

$$\mathrm{E}(S) = \mathrm{E}(N)\mathrm{E}(X) = 21(113.53) = 2{,}384.13$$

and the variance is

$$
\begin{aligned}
\mathrm{Var}(S) &= \mathrm{E}(N)\,\mathrm{Var}(X) + \mathrm{Var}(N)\mathrm{E}(X)^2 \\
&= 21(32{,}733.87) + 21(113.53)^2 = 958{,}081.53
\end{aligned}
$$

for a coefficient of variation of 0.41.

For the other coverages we need general formulas for the first two limited expected moments. For $u > 200$,

$$
\begin{aligned}
\mathrm{E}(X \wedge u) &= 56.99 + \int_{200}^{u} xf(x)dx + \int_{u}^{\infty} uf(x)dx \\
&= 56.99 + c\int_{200}^{u} xe^{-x/314}dx + c\int_{u}^{\infty} ue^{-x/314}dx \\
&= 56.99 + c\left(-314xe^{-x/314} - 314^2e^{-x/314}\right)\Big|_{200}^{u} \\
&\quad + -cu314e^{-x/314}\Big|_{u}^{\infty} \\
&= 56.99 + c\left(161{,}396e^{-200/314} - 314^2e^{-u/314}\right),
\end{aligned}
$$

where $c = 0.000662344$ and similarly

$$
\begin{aligned}
\mathrm{E}[(X \wedge u)^2] &= 5{,}715.82 + c\int_{200}^{u} x^2e^{-x/314}dx + c\int_{u}^{\infty} u^2e^{-x/314}dx \\
&\quad 5{,}715.82 + c\left[-314x^2 - 314^2(2x) - 314^3(2)\right]e^{-x/314}\Big|_{200}^{u} \\
&\quad - cu^2314e^{-x/314}\Big|_{u}^{\infty} \\
&= 5{,}715.82 + c\left[113{,}916{,}688e^{-200/314}\right. \\
&\quad \left.-(197{,}192u + 61{,}918{,}288)e^{-u/314}\right].
\end{aligned}
$$

Table 14.5 Limited moment calculations

u	$\mathrm{E}(X \wedge u)$	$\mathrm{E}[(X \wedge u)^2]$
250	84.07	12,397.08
500	100.24	23,993.47
1,250	112.31	41,809.37
2,500	113.51	45,494.83

Table 14.5 gives the quantities needed to complete the assignment.

The requested moments for the 1,000 excess of 250 coverage are, for one loss,

$$
\begin{aligned}
\text{Mean} &= 112.31 - 84.07 = 28.24, \\
\text{Second moment} &= 41{,}809.37 - 12{,}397.08 - 2(250)(28.24) \\
&= 15{,}292.29, \\
\text{Variance} &= 15{,}292.29 - 28.24^2 = 14{,}494.79, \\
\text{Coefficient of variation} &= \frac{\sqrt{14{,}494.79}}{28.24} = 4.26.
\end{aligned}
$$

It is interesting to note that while, as expected, the coverage limitations reduce the variance, the risk, as measured by the coefficient of variation, has increased considerably. For a full year, the mean is 593.04, the variance is 321,138.09, and the coefficient of variation is 0.96.

For the 2,000 excess of 500 coverage, we have, for one loss,

$$
\begin{aligned}
\text{Mean} &= 113.51 - 100.24 = 13.27, \\
\text{Second moment} &= 45{,}494.83 - 23{,}993.47 - 2(500)(13.27) \\
&= 8{,}231.36, \\
\text{Variance} &= 8{,}231.36 - 13.27^2 = 8{,}055.27, \\
\text{Coefficient of variation} &= \frac{\sqrt{8{,}055.27}}{13.27} = 6.76.
\end{aligned}
$$

Moving further into the tail increases our risk. For one year, the three items are 278.67, 172,858.56, and 1.49.

14.4.3 The second model

From Figure 14.5, if a single parametric distribution is to be used, one with a nonzero mode should be tried. Because the data were rounded to the nearest 1,000, the intervals should be treated as 0.5–5.5, 5.5–10.5, and so on. After considering lognormal, Weibull, gamma, and mixture models (adding an exponential distribution), the lognormal distribution is clearly superior (using the SBC). The parameters are $\hat{\mu} = 4.0626$ and $\hat{\sigma} = 1.1466$. The chi-square

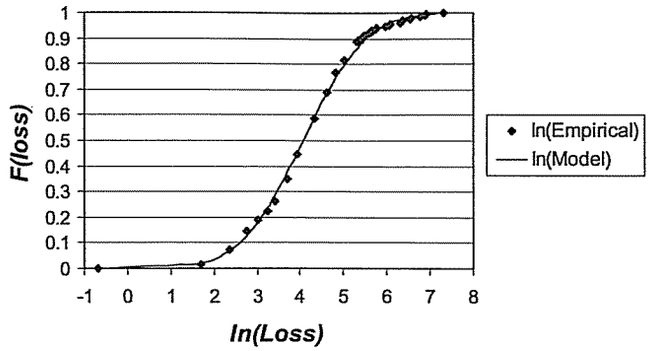


Fig. 14.6 Distribution function plot.

goodness-of-fit test (placing the observations above 200 into a single group) statistic is 7.77 for a p-value of 0.73. Figure 14.6 compares the lognormal model to the empirical model. The graph confirms the good fit.

14.5 AN AGGREGATE LOSS EXAMPLE

The example to be covered in this section summarizes many of the techniques introduced up to this point. The coverage is perhaps more complex than those found in practice, but that gives us a chance to work through a variety of tasks.

Example 14.1 *You are a consulting actuary and have been retained to assist in the pricing of a group hospitalization policy. Your task is to determine the expected payment to be made by the insurer. The terms of the policy (per covered employee) are as follows:*

1. *For each hospitalization of the employee or a member of the employee's family, the employee pays the first* 500 *plus any losses in excess of* 50,500. *On any one hospitalization, the insurance will pay at most* 50,000.
2. *In any calendar year, the employee will pay no more than* 1,000 *in deductibles, but there is no limit on how much the employee will pay in respect of losses exceeding* 50,500.
3. *Any particular hospitalization is assigned to the calendar year in which the individual entered the hospital. Even if hospitalization extends into subsequent years, all payments are made in respect to the policy year assigned.*

Table 14.6 Hospitalizations, per family member, per year

No. of hospitalizations per family member	No. of family members
0	2,659
1	244
2	19
3	2
4 or more	0
Total	2,924

Table 14.7 Number of family members per employee

No. of family members per employee	No. of employees
1	84
2	140
3	139
4	131
5	73
6	42
7	27
8 or more	33
Total	669

4. *The premium is the same, regardless of the number of family members.*

Experience studies have provided the data contained in Tables 14.6 *and* 14.8. *The data in Table* 14.7 *represent the profile of the current set of employees.*

The first step is to fit parametric models to each of the three data sets. For the data in Table 14.6, 12 distributions were fitted. The best one-parameter distribution is the geometric with a negative loglikelihood (NLL) of 969.251 and a chi-square goodness-of-fit p-value of 0.5325. The best two-parameter model is the zero-modified geometric. The NLL improves to 969.058, but by the likelihood ratio test, this is not sufficient to justify the second parameter. The best three-parameter distribution is the zero-modified negative binomial, which has an NLL of 969.056, again not enough to dislodge the geometric as our choice. For the two- and three-parameter models there were not enough degrees of freedom to conduct the chi-square test. We choose the geometric distribution with $\beta = 0.098495$.

Table 14.8 Losses per hospitalization

Loss per hospitalization	No. of hospitalizations
0–250	36
250–500	29
500–1,000	43
1,000–1,500	35
1,500–2,500	39
2,500–5,000	47
5,000–10,000	33
10,000–50,000	24
50,000–	2
Total	288

For the data in Table 14.7, only zero-truncated distributions should be considered. The best one-parameter model is the zero-truncated Poisson with an NLL of 1,298.725 and a p-value near zero. The two-parameter zero-truncated negative binomial has an NLL of 1,292.532, a significant improvement. The p-value is 0.2571, indicating that this is an acceptable choice. The parameters are $r = 13.207$ and $\beta = 0.25884$.

For the data in Table 14.8, 15 continuous distributions were fitted. The four best models for a given number of parameters are listed in Table 14.9. It should be clear that the best choice is the Pareto distribution. The parameters are $\alpha = 1.6693$ and $\theta = 3{,}053.0$.

The remaining calculations were done using the recursive method, but inversion or simulation would work equally well.

The first step is to determine the distribution of payments by the employee per family member with regard to the deductible. The frequency distribution is the geometric distribution while the individual loss distribution is the Pareto distribution, limited to the maximum deductible of 500. That is, any losses in excess of 500 are assigned to the value 500. With regard to discretization for recursion, the span should divide evenly into 500 and then all probability

Table 14.9 Four best models for loss per hospitalization

Name	No. of parameters	NLL	p-value
Inverse exponential	1	632.632	Near 0
Pareto	2	601.642	0.9818
Burr	3	601.612	0.9476
Transformed beta	4	601.553	0.8798

Table 14.10 Discretized Pareto distribution with 500 limit

Loss	Probability
0	0.000273
1	0.000546
2	0.000546
3	0.000545
⋮	⋮
498	0.000365
499	0.000365
500	0.776512

Table 14.11 Probabilities for aggregate deductibles per family member

Loss	Probability
0	0.910359
1	0.000045
2	0.000045
3	0.000045
⋮	⋮
499	0.000031
500	0.063386
501	0.000007
⋮	⋮
999	0.000004
1,000	0.004413
1,001	0.000001
⋮	⋮

not accounted for by the time 500 is reached is placed there. For this example a span of 1 was used. The first few and last few values of the discretized distribution appear in Table 14.10. After applying the recursive formula, it is clear that there is non-zero probability beyond 3,000. However, looking ahead, we know that, with regard to the employee aggregate deductible, payments beyond 1,000 have no impact. A few of these probabilities appear in Table 14.11.

We next must obtain the aggregate distribution of deductibles paid per employee per year. This is another compound distribution. The frequency distribution is the truncated negative binomial and the individual loss distribution is the one for losses per family member that was just obtained.

Table 14.12 Probabilities for aggregate deductibles per employee

Loss	Probability
0	0.725517
1	0.000116
2	0.000115
3	0.000115
⋮	⋮
499	0.000082
500	0.164284
501	0.000047
⋮	⋮
999	0.000031
1,000	0.042343

Recursions can again be used to obtain this distribution. Because there is a 1,000 limit on deductibles, all probability to the right of 1,000 will be placed at 1,000. Selected values from this aggregate distribution are given in Table 14.12. Note that the chance that more than 1,000 in deductibles will be paid is very small. The cost to the insurer of limiting the insured's costs is also small. Using this discrete distribution, it is easy to obtain the mean and standard deviation of aggregate deductibles. They are 150.02 and 274.42, respectively.

We next require the expected value of aggregate costs to the insurer for individual losses below the upper limit of 50,000. This can be found analytically. The expected payment per loss is $\mathrm{E}(X \wedge 50{,}500) = 3{,}890.87$ for the Pareto distribution. The expected number of losses per family member is the mean of the geometric distribution which is the parameter, 0.098495. The expected number of family members per employee comes from the zero-truncated negative binomial distribution and is 3.59015. This implies that the expected number of losses per employee is $0.098495(3.59015) = 0.353612$. Then the expected aggregate dollars in payments up to the individual limit is $0.353612(3{,}890.87) = 1{,}375.86$.

Then the expected cost to the insurer is the difference $1{,}375.86 - 150.02 = 1{,}225.84$. As a final note, it is not possible to use any method other than simulation if the goal is to obtain the probability distribution of the insurer's payments. This situation is similar to that of Example 17.7, where it is easy to get the overall distribution, as well as the distribution for the insured (in this case, if payments for losses over 50,500 are ignored), but not for the insurer. □

14.6 ANOTHER AGGREGATE LOSS EXAMPLE

Careful modeling has revealed that individual losses have the lognormal distribution with $\mu = 10.5430$ and $\sigma = 2.31315$. It has also been determined that the number of losses has the Poisson distribution with $\lambda = 0.0154578$.

Begin by considering excess of loss reinsurance in which the reinsurance pays the excess over a deductible, d, up to a maximum payment $u - d$, where u is the limit established in the primary coverage. There are two approaches available to create the distribution of reinsurer payments. The first is to work with the distribution of payments per payment. On this basis, the severity distribution is mixed, with pdf

$$f_Y(x) = \frac{f_X(x+d)}{1 - F_X(d)}, \quad 0 \le x < u - d,$$

and discrete probability

$$\Pr(Y = u - d) = \frac{1 - F_X(u)}{1 - F_X(d)}.$$

This distribution would then be discretized for use with the recursive formula or the FFT or approximated by a histogram for use with the Heckman–Meyers method. Regardless, the frequency distribution must be adjusted to reflect the distribution of the number of payments as opposed to the number of losses. The new Poisson parameter will be $\lambda[1 - F_X(d)]$.

14.6.1 Distribution for a single policy

We consider the distribution of losses for a single policy for various combinations of d and u. We use the Poisson parameter for the combined group and have employed the recursive algorithm with a discretization interval of 10,000 and the method of rounding. In all cases the 90th and 99th percentiles are zero, indicating that most of the time the excess of loss reinsurance will involve no payments. This is not surprising because the probability there will be no losses is $\exp(-0.0154578) = 0.985$ and with the deductible this probability is even higher. The mean, standard deviation, and coefficient of variation for various combinations of d and u are given in Table 14.13.

It is not surprising that the risk (as measured by the coefficient of variation, C.V.) increases when either the deductible or the limit is increased. It is also clear that the risk of writing one policy is extreme.

14.6.2 One hundred policies—excess of loss

We next consider the possibility of reinsuring 100 policies. If we assume that the same deductible and limit apply to all of them, the aggregate distribution requires only that the frequency be changed. When 100 independent Poisson

Table 14.13 Excess of loss reinsurance, one policy

Deductible (10^6)	Limit (10^6)	Mean	Standard Deviation	C.V.
0.5	1	778	18,858	24.24
0.5	5	2,910	94,574	32.50
0.5	10	3,809	144,731	38.00
0.5	25	4,825	229,284	47.52
0.5	50	5,415	306,359	56.58
1.0	5	2,132	80,354	37.69
1.0	10	3,031	132,516	43.72
1.0	25	4,046	219,475	54.24
1.0	50	4,636	298,101	64.30
5.0	10	899	62,556	69.58
5.0	25	1,914	162,478	84.89
5.0	50	2,504	249,752	99.74
10.0	25	1,015	111,054	109.41
10.0	50	1,605	205,939	128.71

Table 14.14 Excess of loss reinsurance, 100 policies

Deductible (10^6)	Limit (10^6)	Mean (10^3)	Standard deviation (10^3)	C.V.	Percentiles (10^3) 90	99
0.5	5	291	946	3.250	708	4,503
0.5	10	381	1,447	3.800	708	9,498
0.5	25	482	2,293	4.752	708	11,674
1.0	5	213	804	3.769	190	4,002
1.0	10	303	1,325	4.372	190	8,997
1.0	25	405	2,195	5.424	190	11,085
5.0	10	90	626	6.958	0	4,997
5.0	25	191	1,625	8.489	0	6,886
10.0	25	102	1,111	10.941	0	1,854

random variables are added, the sum has a Poisson distribution with the original parameter multiplied by 100. The same process was repeated with the revised Poisson parameter. The results appear in Table 14.14.

As must be the case with independent policies, the mean is 100 times the mean for one policy and the standard deviation is 10 times the standard deviation for one policy. This implies that the coefficient of variation will be one-tenth of its previous value. In all cases, the 99th percentile is now above zero. This may make it appear that there is more risk, but in reality it just indicates that it is now more likely that a claim will be paid.

14.6.3 One hundred policies—aggregate stop-loss

We now turn to aggregate reinsurance. Assume policies have no individual deductible but do have a policy limit of u. There are again 100 policies and this time the reinsurer pays all aggregate losses in excess of an aggregate deductible of a. For a given limit, the severity distribution is modified as before, the Poisson parameter is multiplied by 100, and then some algorithm is used to obtain the aggregate distribution. Let this distribution have cdf $F_S(s)$ or, in the case of a discretized distribution (as will be the output from the recursive algorithm or the FFT), a pf $f_S(s_i)$ for $i = 1, \ldots, n$. For a deductible of a, the corresponding functions for the reinsurance distribution S_r are

$$\begin{aligned} F_{S_r}(s) &= F_S(s+a), \quad s \geq 0, \\ f_{S_r}(0) &= F_S(a) = \sum_{s_i \leq a} f_S(s_i), \\ f_{S_r}(r_i) &= f_S(r_i + a), \quad r_i = s_i - a, \; i = 1, \ldots, n. \end{aligned}$$

Moments and percentiles may be determined in the usual manner.

Using the recursive formula with an interval of 10,000, results for various stop-loss deductibles and individual limits are given in Table 14.15. The results are similar to those for the excess of loss coverage. For the most part, as either the individual limit or the aggregate deductible is increased, the risk, as measured by the coefficient of variation, increases. The exception is when both the limit and the deductible are 5,000,000. This is a risky setting because it is the only one in which two losses are required before the reinsurance will take effect.

Now suppose the 100 policies are known to have different Poisson parameters (but the same severity distribution). Assume 30 have $\lambda = 0.0162249$ and so the number of claims from this subgroup is Poisson with mean

$$30(0.0162249) = 0.486747.$$

For the second group (50 members) the parameter is $50(0.0174087) = 0.870435$ and for the third group (20 members) it is $20(0.0096121) = 0.192242$. There are three methods for obtaining the distribution of the sum of the three separate aggregate distributions.

1. Because the sum of independent Poisson random variables is still Poisson, the total number of losses has the Poisson distribution with parameter 1.549424. The common severity distribution remains the lognormal. This reduces to a single compound distribution which can be evaluated by any method.

2. Obtain the three aggregate distributions separately. If the recursive or FFT algorithms are used, the result will be three discrete distributions. The distribution of their sum can be obtained by using convolutions.

Table 14.15 Aggregate stop-loss reinsurance, 100 policies

Deductible (10^6)	Limit (10^6)	Mean (10^3)	Standard deviation (10^3)	C.V.	Percentiles (10^3) 90	99
0.5	5	322	1,003	3.11	863	4,711
0.5	10	412	1,496	3.63	863	9,504
0.5	25	513	2,331	4.54	863	11,895
1.0	5	241	879	3.64	363	4,211
1.0	10	331	1,389	4.19	363	9,004
1.0	25	433	2,245	5.19	363	11,395
2.5	5	114	556	4.86	0	2,711
2.5	10	204	1,104	5.40	0	7,504
2.5	25	306	2,013	6.58	0	9,895
5.0	5	13	181	13.73	0	211
5.0	10	103	714	6.93	0	5,004
5.0	25	205	1,690	8.26	0	7,395

3. If the FFT or Heckman–Meyers algorithms are used, the three transforms can be found and then multiplied. The inverse transform is then taken of the product.

Each of the methods has advantages and drawbacks. The first method is restricted to those frequency distributions for which the sum has a known form. If the severity distributions are not identical, it may not be possible to combine them to form a single model. The major advantage is that, if it is available, this method requires only one aggregate calculation.

The advantage of method 2 is that there is no restriction on the frequency and severity components of the components. The drawback is the expansion of computer storage. For example, if the first distribution requires 3,000 points, the second one 5,000 points, and the third one 2,000 points (with the same discretization interval being used for the three distributions), the combined distribution will require 10,000 points. More will be said about this at the end of this section.

The third method also has no restriction on the separate models. It has the same drawback as the second method, but here the expansion must be done in advance. That is, in the example, all three components must work with 10,000 points. There is no way to avoid this.

14.6.4 Numerical convolutions

The remaining problem is expansion of the number of points required when performing numerical convolutions. The problem arises when the individual distributions use a large number of discrete points, to the point where the storage capacity of the computer becomes an obstacle. The following example is a small-scale version of the problem and indicates a simple solution.

Example 14.2 *The probability functions for two discrete distributions are given below. Suppose the maximum vector allowed by the computer program being used is of length* 6. *Determine an approximation to the probability function for the sum of the two random variables.*

x	$f_1(x)$	$f_2(x)$
0	0.3	0.4
2	0.2	0.3
4	0.2	0.2
6	0.2	0.1
8	0.1	0.0

The maximum possible value for the sum of the two random variables is 14 and would require a vector of length 8 to store. Usual convolutions produce the answer as given below.

x	0	2	4	6	8	10	12	14
$f(x)$	0.12	0.17	0.20	0.21	0.16	0.09	0.04	0.01

With 6 points available, the span must be increased to $14/5 = 2.8$. We then do a sort of reverse interpolation, taking the probability at each point that is not a multiple of 2.8 and allocating it to the two nearest multiples of 2.8. For example, the probability of 0.16 at $x = 8$ is allocated to the points 5.6 and 8.4. Because 8 is 2.4/2.8 of the way from 5.6 to 8.4, six-sevenths of the probability is placed at 8.4 and the remaining one-seventh is placed at 5.6. The complete allocation process appears in Table 14.16. The probabilities allocated to each multiple of 2.8 are then combined to produce the approximation to the true distribution of the sum. The approximating distribution is given below.

x	0	2.8	5.6	8.4	11.2	14.0
$f(x)$	0.1686	0.2357	0.2886	0.2057	0.0800	0.0214

This method preserves both the total probability of one and the mean (both the true distribution and the approximating distribution have a mean of 5.2). □

One refinement that can eliminate some of the need for storage is to note that when a distribution requires a large vector the probabilities at the end are

Table 14.16 Allocation of probabilities for Example 14.2

x	$f(x)$	Lower point	Probabiity	Upper point	Probability
0	0.12	0	0.1200		
2	0.17	0	0.0486	2.8	0.1214
4	0.20	2.8	0.1143	5.6	0.0857
6	0.21	5.6	0.1800	8.4	0.0300
8	0.16	5.6	0.0229	8.4	0.1371
10	0.09	8.4	0.0386	11.2	0.0514
12	0.04	11.2	0.0286	14.0	0.0114
14	0.01	14.0	0.0100		

likely to be very small. When they are multiplied to create the convolution, the probabilities at the ends of the new, long vector may be so small that they can be ignored. Thus those cells need not be retained and do not add to the storage problem.

Many more refinements are possible. In the appendix to the article by Bailey [9] a method which preserves the first three moments is presented. He also provides guidance with regard to the elimination or combination of storage locations with exceptionally small probability.

14.7 COMPREHENSIVE EXERCISES

The exercises in this section are similar to the examples presented earlier in this chapter. They are based on questions that arose in published papers.

14.2 In New York there were special funds for some infrequent occurrences under workers compensation insurance. One was the event of a case being reopened. Hipp [57] collected data on the time from an accident to when the case was reopened. These covered cases reopened between April 24, 1933 and December 31, 1936. The data appear in Table 14.17. Determine a parametric model for the time from accident to reopening. By definition, at least seven years must elapse before a claim can qualify as a reopening, so the model should be conditioned on the time being at least seven years.

14.3 In the first of two papers by Arthur Bailey [6], written in 1942 and 1943, he observed on page 51 that "Another field where a knowledge of sampling distributions could be used to advantage is that of rating procedures for deductibles and excess coverages." In the second paper [7], he presented some data (Table 14.18) on the distribution of loss ratios. In that paper he made the statement that the popular lognormal model provided a good fit and passed the chi-square test. Does it? Is there a better model?

Table 14.17 Time to reopening of a workers compensation claim for Exercise 14.2

Years	No. reopened	Years	No. reopened
7–8	27	15–16	13
8–9	43	16–17	9
9–10	42	17–18	7
10–11	37	18–19	4
11–12	25	19–20	4
12–13	19	20–21	1
13–14	23	21+	0
14–15	10		
		Total	264

Table 14.18 Loss ratio data for Exercise 14.3

Loss ratio	Number
0.0–0.2	16
0.2–0.4	27
0.4–0.6	22
0.6–0.8	29
0.8–1.0	19
1.0–1.5	32
1.5–2.0	10
2.0–3.0	13
3.0+	5
Total	173

14.4 In 1979, Hewitt and Lefkowitz [56] looked at automobile bodily injury liability data (Table 14.19) and concluded that a two-point mixture of the gamma and loggamma distributions [If X has a gamma distribution, then $Y = \exp(X)$ has the loggamma distribution. Note that its support begins at 1] was superior to the lognormal. Do you agree? Also consider the gamma and loggamma distributions.

14.5 A 1980 paper by Patrik [102] contained many of the ideas recommended in this text. One of his examples was data supplied by the Insurance Services Office on Owners, Landlords, and Tenants bodily injury liability. Policies at two different limits were studied. Both were for policy year 1976 with losses developed to the end of 1978. The groupings in Table 14.20 have been condensed from those in the paper. Can the same model (with or without identical parameters) be used for the two limits?

Table 14.19 Automobile bodily injury liability losses for Exercise 14.4

Loss	Number	Loss	Number
0–50	27	750–1,000	8
50–100	4	1,000–1,500	16
100–150	1	1,500–2 ,000	8
150–200	2	2,000–2 ,500	11
200–250	3	2,500–3 ,000	6
250–300	4	3,000–4 ,000	12
300–400	5	4,000–5 ,000	9
400–500	6	5,000–7 ,500	14
500–750	13	7,500–	40
		Total	189

Table 14.20 OLT bodily injury liability losses for Exercise 14.5

Loss (10^3)	300 Limit	500 Limit	Loss (10^3)	300 Limit	500 Limit
0–0.2	10,075	3,977	11–12	56	22
0.2–0.5	3,049	1,095	12–13	47	23
0.5–1	3,263	1,152	13–14	20	6
1–2	2,690	991	14–15	151	51
2–3	1,498	594	15–20	151	54
3–4	964	339	20–25	109	44
4–5	794	307	25–50	154	53
5–6	261	103	50–75	24	14
6–7	191	79	75–100	19	5
7–8	406	141	100–200	22	6
8–9	114	52	200–300	6	9
9–10	279	89	300–500	10[a]	3
10–11	58	23	500–		0
			Totals	24,411	9,232

[a]losses for 300+

14.6 The data in Table 14.21 were collected by Fisher [37] on coal mining disasters in the United States over 25 years ending about 1910. This particular compilation counted the number of disasters per year that claimed the lives of five to nine miners. In the article, Fisher claimed that a Poisson distribution was a good model. Is it? Is there a better model?

14.7 Harwayne [49] was curious as to the relationship between driving record and number of accidents. His data on California drivers included the number of violations. For each of the six data sets represented by each column in Table 14.22, is a negative binomial distribution appropriate? If so, are the same parameters appropriate? Is it reasonable to conclude that the expected number of accidents increases with the number of violations?

Table 14.21 Mining disasters per year for Exercise 14.6

No. of disasters	No. of years	No. of disasters	No. of years
0	1	7	3
1	1	8	1
2	3	9	0
3	4	10	1
4	5	11	1
5	2	12	1
6	2	13+	0

Table 14.22 Number of accidents by number of violations for Exercise 14.7

Number of Accidents	No. of violations					
	0	1	2	3	4	5+
0	51,365	17,081	6,729	3,098	1,548	1,893
1	3,997	3,131	1,711	963	570	934
2	357	353	266	221	138	287
3	34	41	44	31	34	66
4	4	6	6	6	4	14
5+	0	1	1	1	3	1

Table 14.23 Number of accidents per year for Exercise 14.8

No. of accidents	No. of stretches	No. of accidents	No. of stretches
0	99	6	4
1	65	7	0
2	57	8	3
3	35	9	4
4	20	10	0
5	10	11	1

14.8 In 1961, Simon [122] proposed using the zero-modified negative binomial distribution. His data set was the number of accidents in one year along various one-mile stretches of Oregon highway. The data appear in Table 14.23. Simon claimed that the zero-modified negative binomial distribution was superior to the negative binomial. Is he correct? Is there a better model?

Part V

Adjusted estimates and simulation

15

Interpolation and smoothing

15.1 INTRODUCTION

Methods of model building discussed to this point are based on ideas that came primarily from the fields of probability and statistics. Data are considered to be observations from a sample space associated with a probability distribution. The quantities to be estimated are functions of that probability distribution for example, pdf, cdf, hazard rate (force of mortality), mean, variance.

In contrast, the methods described in this chapter have their origins in the field of numerical analysis, without specific considerations of probabilistic statistical concepts.

In practice, many of these numerical methods have been subsequently adapted to a probability and statistics framework. Although the key ideas of the methods are easy to understand, most of these techniques are computationally demanding, thus requiring computer programs. The techniques described in this chapter are at the lowest end of the complexity scale.

The objective is to fit a smooth curve through a set of data according to some specified criteria. This has many applications in actuarial science as it has in many other fields. We begin with a set of distinct points in the plane. In practice these points represent a sequence of successive observations of some quantity, for example, a series of successive monthly inflation rates, a set of

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

successive average annual claim costs, or a set of successive observed mortality rates by age. The methods in this chapter are considered to be nonparametric in nature in the sense that the underlying model is not prespecified by a simple mathematical function with a small number of parameters. The methods in this chapter allow for great flexibility in the shape of the resulting curve. They are especially useful in situations where the shape is complex.

One such example is the curve representing the probabilities of death within a short period for humans, such as the function q_x. These probabilities decrease sharply at the youngest ages as a result of neonatal deaths, are relatively flat until the early teens, rise slowly during the teens, rise and then fall (especially for males) during the 18–25 age range (as a result of accidents), then continue to rise slowly but at an increasing rate for higher ages. This curve is not captured adequately by a simple function (although there are models with eight or more parameters available).

Historically, the process of smoothing a set of observed irregular points is called graduation. The set of points typically represents observed rates of mortality (probability of death within one year) or rates of some other contingency such as disablement, unemployment, or accident. The methods described in this chapter are not restricted to these kinds of applications. Indeed, they can be applied to any set of successive points.

In graduation theory, it is assumed that there is some underlying, but unobservable, true curve or function that is to be estimated or approximated. Graduation depends on a trade-off between the high degree of fit that is obtained by a "noisy" curve such as a high-degree polynomial that fits the data well and the high degree of smoothness that is obtained by a simple curve such as a straight line or an exponential curve.

There are a number of classical methods described in older actuarial textbooks such as Miller [94]. These include simple graphical methods using an engineering draftsman's French curve or a spline and weights. A French curve is a flat piece of wood with a smooth outside edge, with the diameter of the outside edge changing gradually. This could be used to draw curves through specified points. A spline was a thin rod of flexible metal or plastic that was anchored by attaching lead weights called *ducks* at specified points along the rod. By altering the position of the ducks on the rod and moving the rod relative to the drafting surface, smooth curves could be drawn through successive sets of points. The resulting shape of the rod is the one that minimizes the *energy of deflection* subject to the rod passing through the specified points. In that sense it is a very natural method for developing the shape of a structure so that it has maximal strength. Methods developed by actuaries included mathematical methods based on running averages, methods based on interpolation, and methods based directly on finding a balance between fit and smoothness. All these methods were developed in the early 1900s, some even earlier. They were developed using methods of finite differences, in which it was frequently assumed that fourth and higher differences should be set to zero, implicitly forcing the use of third-degree polynomials. Formulas

involving differences were developed so that an actuary could develop smooth functions using only pencil and paper. Remember these formulas were developed long before calculators (mechanical or electronic!) and very long before computers were developed. A more recent summary of these methods along with some updated variations can be found in London [84].

With the advent of computers in the 1950s and 1960s, many computerized mathematical procedures were developed. Among them was the theory of splines, this time not mechanical in nature. As with graduation, the objective of splines is to find an appropriate balance between fit and smoothness. The solutions that were developed were in terms of linear systems of equations that could be easily solved on a computer. The modern theory of splines dates back to Schoenberg [117].

In this chapter, we focus only on the modern techniques of spline interpolation and smoothing. These techniques are so powerful and flexible that they have largely superseded the older methods.

15.2 POLYNOMIAL INTERPOLATION AND SMOOTHING

Consider $n+1$ distinct points labeled $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ with $x_0 < x_1 < x_2 < \cdots < x_n$. A unique polynomial of degree n can be passed through these points. This polynomial is called a *collocation* polynomial and can be expressed as

$$f(x) = \sum_{j=0}^{n} a_j x^j, \tag{15.1}$$

where

$$f(x_j) = y_j, \quad j = 0, 1, \ldots, n. \tag{15.2}$$

Equations (15.2) form a system of $n+1$ equations in $n+1$ unknowns $\{a_j; j = 0, 1, \ldots, n\}$. However, when n is large, the numerical exercise of solving the system of equations may be difficult.

Fortunately, the solution can be explicitly written without solving the system of equations. The solution is known as Lagrange's formula:

$$\begin{aligned} f(x) &= y_0 \frac{(x-x_1)(x-x_2)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_n)} \\ &\quad + y_1 \frac{(x-x_0)(x-x_2)\ldots(x-x_n)}{(x_1-x_0)(x_1-x_2)\ldots(x_1-x_n)} \\ &\quad + \ldots \\ &\quad + y_n \frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})} \\ &= \sum_{j=0}^{n} y_j \frac{(x-x_0)\ldots(x-x_{j-1})(x-x_{j+1})\ldots(x-x_n)}{(x_j-x_0)\ldots(x_j-x_{j-1})(x_j-x_{j+1})\ldots(x_j-x_n)}. \end{aligned} \tag{15.3}$$

Table 15.1 Mortality rates for Example 15.1

j	Ages	Exposed to Risk	Actual Deaths	Estimated Mortality Rate Per 1,000
0	25-29	35,700	139	3.89
1	30-34	244,066	599	2.45
2	35-39	741,041	1,842	2.49
3	40-44	1,250,601	4,771	3.81
4	45-49	1,746,393	11,073	6.34
5	50-54	2,067,008	21,693	10.49
6	55-59	1,983,710	31,612	15.94
7	60-64	1,484,347	39,948	26.91
8	65-69	988,980	40,295	40.74
9	70-74	559,049	33,292	59.55
10	75-79	241,497	20,773	86.02
11	80-84	78,229	11,376	145.42
12	85-89	15,411	2,653	172.15
13	90-94	2,552	589	230.80
14	95-	162	44	271.60
	Total	11,438,746	220,699	

To verify that (15.3) is the collocation polynomial, note that each term is a polynomial of degree n and that when $x = x_j$ the right-hand side of (15.3) takes on value y_j for each of $j = 0, 1, 2, \ldots, n$.

The n-degree polynomial $f(x)$ provides interpolation between (x_0, y_0) and (x_n, y_n) and passes through all interior points $\{(x_j, y_j); j = 1, \ldots, n-1\}$. However, for large n, the function $f(x)$ can exhibit excessive oscillation; that is to say, it can be very "wiggly." This is particularly problematic when there is some "noise" in the original series $\{(x_j, y_j); j = 0, \ldots, n\}$. Such noise can be caused by measurement error or random fluctuation.

Example 15.1 *The data in Table* 15.1 *are from Miller* [94], *p.* 62. *They are observed mortality rates in five-year age groups. The estimated mortality rates are obtained as the ratio of the dollars of death claims paid to the total dollars exposed to death.*[1] *The rates are plotted in Figure* 15.1.

The estimates of mortality rates at each age are the maximum likelihood estimates of the true rates assuming mutually independent binomial models

[1]Deaths and exposures are in units of \$1,000. It is common in mortality studies to count dollars rather than lives in order to give more weight to the larger policies. The mortality rates in the table are the ratios of the given deaths and exposures. The last entry differs from Miller's table due to rounding.

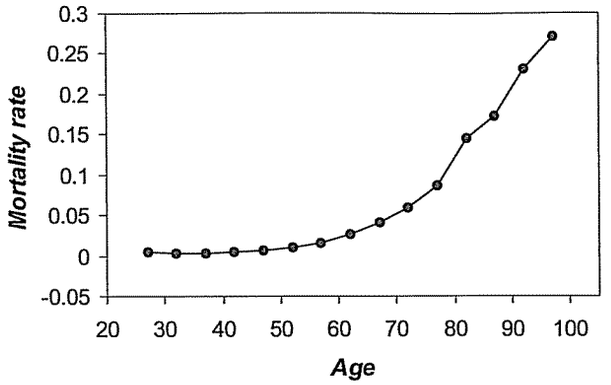


Fig. 15.1 Mortality rates for Example 15.1.

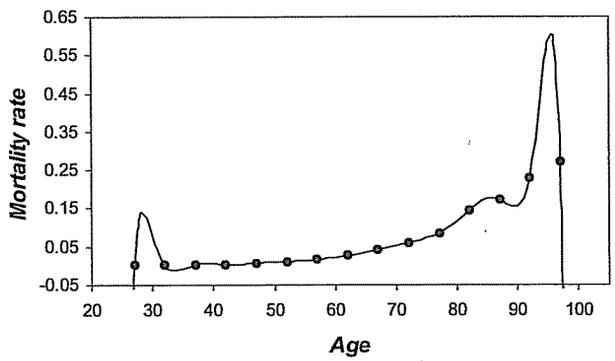


Fig. 15.2 Collocation polynomial for mortality data.

at each age. Note that there is considerable variability in successive estimates. Of course, mortality rates are expected to be relatively smooth from age to age. Figure 15.1 shows the observed mortality rates connected by straight lines while Figure 15.2 shows a collocation polynomial fitted through the observed rates. Notice its wiggly form and its extreme oscillation near the ends. □

To avoid the excessive oscillatory behavior or wiggliness, lower order polynomials could be used for interpolation. For example, successive values could

be joined by straight lines. However, the successive interpolating lines form a jagged series because of the "kinks" at the points of juncture.

Another method is to piece together a sequence of low-degree polynomials. For example, a quadratic function can be collocated with successive points at $(x_0, x_1, x_2), (x_2, x_3, x_4) \ldots$. However, there will not be smoothness at the points of juncture $x_2, x_4, \ldots$ in the sense that the interpolating function will have kinks at these points with slopes and curvature not matching. One way to get rid of the kinks is to force some left-hand and right-hand derivatives to be equal at these points. This creates apparent smoothness at the points of juncture of the successive polynomials. This is the key idea behind splines. Interpolating splines are piecewise polynomial functions that pass through the given data points but that have the added feature that they are smooth at the points of juncture of the successive pieces. The order of the polynomial is kept low to minimize "wiggly" behavior. Interpolation using cubic splines is introduced in Section 15.3.

An alternative to interpolation is *smoothing*, or, more precisely, fitting a smooth function to the observed data but not requiring that the function pass through each data point. Polynomials allow for great flexibility of shapes. However, this flexibility of shape also makes polynomials quite risky to use for extrapolation, especially for polynomials of high degree. This was the case in Figure 15.2, where the extrapolated values, even for one year, were completely unreliable. As with the fitting of other models earlier in this book, a fitting criterion needs to be selected in order to fit a model. We will illustrate the use of polynomial smoothing by using a least squares criterion. Figures 15.3–15.6 show the fits of polynomials of degree 2, 3, 4, and 5 to the data of Example 15.1. It should be noted that the fit improves with each increase in degree because there is one additional degree of freedom in carrying out the fit. However, it can be seen that as each degree is added the behavior of the extrapolated values for only a few years below age 27 and above age 97 changes quite significantly. Smoothing splines provide one solution to this dilemma. Smoothing splines are just like interpolating splines except that the spline is not required to pass through the data points but, rather, should be close to the data points. Cubic splines limit the degree of the polynomial to 3.

15.2.1 Exercises

15.1 Determine the equation of the polynomial that interpolates the points $(2, 50)$, $(4, 25)$, and $(5, 20)$.

15.2 Determine the equation of the straight line that best fits the data of Exercise 15.1 using the least squares criterion.

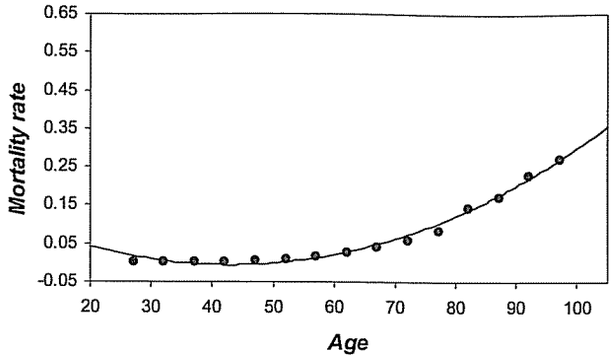


Fig. 15.3 Second-degree polynomial fit.

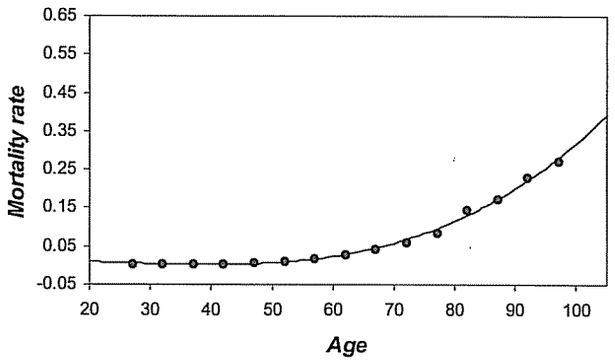


Fig. 15.4 Third-degree polynomial fit.

15.3 CUBIC SPLINE INTERPOLATION

Cubic splines are piecewise cubic functions that have the property that the first and second derivatives can be forced to be continuous, unlike the approach of successive polynomials with jagged points of juncture.

Cubic splines are used extensively in computer-aided design and manufacturing in creating surfaces that are smooth to the touch and to the eye. The cubic spline is fitted to a series of points, called **knots**, that give the basic shape of the object being designed or manufactured.

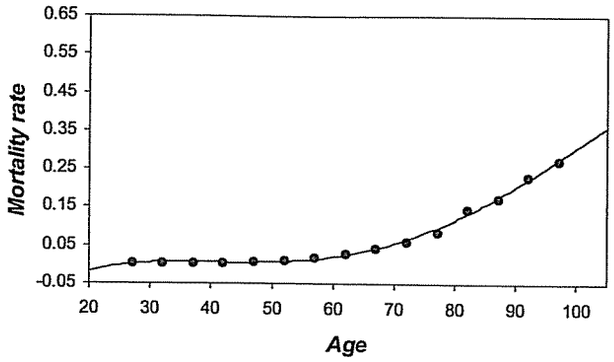


Fig. 15.5 Fourth-degree polynomial fit.

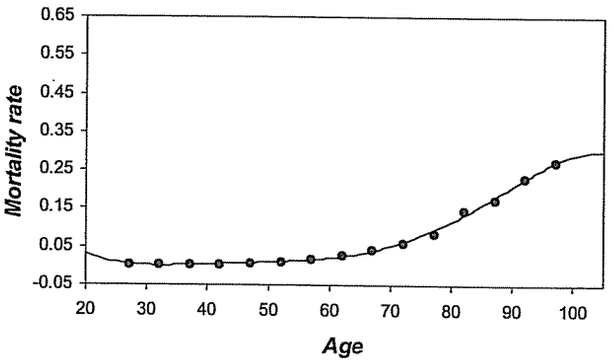


Fig. 15.6 Fifth-degree polynomial fit.

In the terminology of graduation theory as developed by actuaries in the early 1900s, cubic spline interpolation is called **osculatory interpolation.**[2]

Programs for cubic splines are included in many mathematical and engineering software packages. This makes them very easy to apply.

[2]The word **osculation** means "the act of kissing." Successive cubic polynomials exhibit osculatory behavior by "kissing" each other smoothly at the knots!

Definition 15.2 *Suppose that* $\{(x_j, y_j); j = 0, \ldots, n\}$ *are* $n+1$ *distinct knots with* $x_0 < x_1 < x_2 < \cdots < x_n$. *The function* $f(x)$ *is a* ***cubic spline*** *if there exist* n *cubic polynomials* $f_j(x)$ *with coefficients* a_j, b_j, c_j, *and* d_j *that satisfy:*

I. $f(x) = f_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$ *for* $x_j \le x \le x_{j+1}$ *and* $j = 0, 1, \ldots, n-1$.

II. $f(x_j) = y_j, \quad j = 0, 1, \ldots, n$.

III. $f_j(x_{j+1}) = f_{j+1}(x_{j+1}), \quad j = 0, 1, 2, \ldots, n-2$.

IV. $f_j'(x_{j+1}) = f_{j+1}'(x_{j+1}), \quad j = 0, 1, 2, \ldots, n-2$.

V. $f_j''(x_{j+1}) = f_{j+1}''(x_{j+1}), \quad j = 0, 1, 2, \ldots, n-2$.

Property I states that $f(x)$ consists of piecewise cubics. Property II states that the piecewise cubics pass through the given set of data points. Property III requires the spline to be continuous at the interior data points. Properties IV and V provide smoothness at the interior data points by forcing the first and second derivatives to be continuous.

15.3.1 Construction of cubic splines

Each cubic polynomial has four unknown constants: a_j, b_j, c_j, and d_j. Because there are n such cubics, there are $4n$ coefficients to be determined. Properties II–V provide $n+1, n-1, n-1$, and $n-1$ conditions, respectively, for a total of $4n-2$ conditions. In order to determine the $4n$ coefficients, we need exactly two more conditions. This can be done by adding two **endpoint constraints** involving some of $f'(x), f''(x)$, or $f'''(x)$ at x_0 and x_n. Different choices of endpoint constraints lead to different results. Various possible endpoint constraints are discussed in the next section.

In order to construct the cubic segments in the successive intervals, first consider the second derivative $f_j''(x)$. It is a linear function because $f_j(x)$ is cubic. Therefore, the Lagrangian representation of the second derivatives is

$$f_j''(x) = f''(x_j)\frac{x - x_{j+1}}{x_j - x_{j+1}} + f''(x_{j+1})\frac{x - x_j}{x_{j+1} - x_j}. \tag{15.4}$$

To simplify notation, let $m_j = f''(x_j)$ and $h_j = x_{j+1} - x_j$, so that

$$f_j''(x) = \frac{m_j}{h_j}(x_{j+1} - x) + \frac{m_{j+1}}{h_j}(x - x_j) \tag{15.5}$$

for $x_j \le x \le x_{j+1}$ and $j = 0, 1, \ldots, n-1$.

Integrating this twice leads to

$$f_j(x) = \frac{m_j}{6h_j}(x_{j+1} - x)^3 + \frac{m_{j+1}}{6h_j}(x - x_j)^3 + p_j(x_{j+1} - x) + q_j(x - x_j), \tag{15.6}$$

where p_j and q_j are undetermined constants of integration. [To check this, just differentiate (15.6) twice.]

Substituting x_j and x_{j+1} into (15.6) yields

$$y_j = \frac{m_j}{6}h_j^2 + p_j h_j \tag{15.7}$$

and

$$y_{j+1} = \frac{m_{j+1}}{6}h_j^2 + q_j h_j \tag{15.8}$$

because $f_j(x_j) = y_j$ and $f_j(x_{j+1}) = y_{j+1}$.

We now obtain the constants p_j and q_j from (15.7) and (15.8). When they are substituted into (15.6), we obtain

$$\begin{aligned} f_j(x) &= \frac{m_j}{6h_j}(x_{j+1}-x)^3 + \frac{m_{j+1}}{6h_j}(x-x_j)^3 \\ &\quad + \left(\frac{y_j}{h_j} - \frac{m_j h_j}{6}\right)(x_{j+1}-x) \\ &\quad + \left(\frac{y_{j+1}}{h_j} - \frac{m_{j+1}h_j}{6}\right)(x-x_j). \end{aligned} \tag{15.9}$$

Note that the $m_j = f''(x_j)$ terms are still unknown. To obtain them, differentiate (15.9),

$$\begin{aligned} f_j'(x) &= -\frac{m_j}{2h_j}(x_{j+1}-x)^2 + \frac{m_{j+1}}{2h_j}(x-x_j)^2 \\ &\quad - \left(\frac{y_j}{h_j} - \frac{m_j h_j}{6}\right) + \frac{y_{j+1}}{h_j} - \frac{m_{j+1}h_j}{6}. \end{aligned} \tag{15.10}$$

Now setting $x = x_j$ yields, after simplification,

$$f_j'(x_j) = -\frac{m_j}{3}h_j - \frac{m_{j+1}}{6}h_j + \frac{y_{j+1}-y_j}{h_j}. \tag{15.11}$$

Replacing j by $j-1$ in (15.10) and setting $x = x_j$,

$$f_{j-1}'(x_j) = \frac{m_j}{3}h_{j-1} + \frac{m_{j-1}}{6}h_{j-1} + \frac{y_j - y_{j-1}}{h_{j-1}}. \tag{15.12}$$

Now, Property IV forces the slopes to be equal at each knot. This requires us to equate the right-hand sides of (15.11) and (15.12), yielding the following relation between successive values m_{j-1}, m_j and m_{j+1}:

$$h_{j-1}m_{j-1} + 2(h_{j-1}+h_j)m_j + h_j m_{j+1} = 6\left(\frac{y_{j+1}-y_j}{h_j} - \frac{y_j - y_{j-1}}{h_{j-1}}\right) \tag{15.13}$$

for $j = 1, 2, \ldots, n-1$.

The system of equations (15.13) consists of $n-1$ equations in $n+1$ unknowns $m_0, m_1, \ldots, m_n$. Two endpoint constraints can be added to determine m_0 and m_n. Obtaining the $n-1$ remaining unknowns in (15.13) then allows for complete determination of the cubic (15.9) for $j = 0, 1, 2, \ldots, n-1$ and therefore the entire cubic spline.

For purpose of notational simplicity, we can rewrite (15.13) as

$$h_{j-1}m_{j-1} + g_j m_j + h_j m_{j+1} = u_j, \quad j = 1, 2, \ldots, n-1, \tag{15.14}$$

where

$$u_j = 6\left(\frac{y_{j+1}-y_j}{h_j} - \frac{y_j - y_{j-1}}{h_{j-1}}\right) \text{ and } g_j = 2(h_{j-1}+h_j). \tag{15.15}$$

When the endpoints m_0 and m_n are determined externally, the system (15.14) can be rewritten in matrix notation as

$$\underset{(n-1)\times(n-1)}{\begin{bmatrix} g_1 & h_1 & 0 & & \cdots & & & 0 \\ h_1 & g_2 & h_2 & 0 & & \cdots & & 0 \\ 0 & h_2 & g_3 & h_3 & 0 & & \cdots & 0 \\ & 0 & & & & & & \vdots \\ & & \ddots & \ddots & & & h_{n-3} & 0 \\ & & & & h_{n-3} & g_{n-2} & h_{n-2} & \\ 0 & 0 & & \cdots & 0 & h_{n-2} & g_{n-1} & \end{bmatrix}} \underset{(n-1)\times 1}{\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ \\ \vdots \\ m_{n-2} \\ m_{n-1} \end{bmatrix}} = \underset{(n-1)\times 1}{\begin{bmatrix} u_1 - h_0 m_0 \\ u_2 \\ \vdots \\ \\ \vdots \\ u_{n-2} \\ u_{n-1} - h_{n-1}m_n \end{bmatrix}} \tag{15.16}$$

or as

$$\mathbf{Hm} = \mathbf{v}. \tag{15.17}$$

The matrix $\mathbf{H}$ is tridiagonal and invertible. Thus the system (15.17) has a unique solution $\mathbf{m} = \mathbf{H}^{-1}\mathbf{v}$. Alternatively, the system can be solved manually using Gaussian elimination.

Once the values $m_1, m_2, \ldots, m_{n-1}$ are determined, the values of c_j are determined by

$$c_j = \frac{m_j}{2}, \quad j = 1, \ldots, n-1.$$

Property II specifies that

$$a_j = y_j, \quad j = 0, \ldots, n-1.$$

Property V specifies that

$$m_j + 6d_j h_j = m_{j+1}, \quad j = 0, \ldots, n-2,$$

yielding

$$d_j = \frac{m_{j+1} - m_j}{6h_j}, \quad j = 0, \ldots, n-2.$$

Property III specifies that

$$a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 = y_{j+1}, \quad j = 0, \ldots, n-2.$$

Substituting for a_j, c_j, and d_j yields

$$b_j = \frac{y_{j+1} - y_j}{h_j} - \frac{h_j(2m_j + m_{j+1})}{6}, \quad j = 0, \ldots, n-2.$$

Summarizing, the spline coefficients for the first $n-1$ spline segments are computed as

$$\begin{aligned} a_j &= y_j, \\ b_j &= \frac{y_{j+1} - y_j}{h_j} - \frac{h_j(2m_j + m_{j+1})}{6}, \\ c_j &= \frac{m_j}{2}, \\ d_j &= \frac{m_{j+1} - m_j}{6h_j}, \quad j = 0, \ldots, n-2. \end{aligned} \tag{15.18}$$

The only remaining issue in order to obtain the cubic spline is the choice of the two endpoint constraints. There are several possible choices. Once the two endpoint constraints are selected, the n cubics are fully specified. Thus the values of b_{n-1}, c_{n-1}, and d_{n-1} can also be obtained using (15.18).

Case 1: Natural Cubic Spline ($m_0 = m_n = 0$)

The natural spline is obtained by setting m_0 and m_n to zero in (15.16). Because m_0 and m_n are the second derivatives at the endpoints, the choice of zero minimizes the oscillatory behavior at both ends. It also makes the spline linear beyond the boundary knots, a property that minimizes oscillatory behavior beyond both ends of the data. This is probably safest for extrapolation beyond the data points in most applications. Note that the second-derivative endpoint constraints do not in themselves restrict the slopes at the endpoints.

Case 2: Curvature-Adjusted Cubic Spline (m_0 and m_n fixed)

It is similarly possible to fix the endpoint second derivatives m_0 and m_n to prespecified values $f''(x_0)$ and $f''(x_n)$, respectively. Then (15.16) can again be used directly to obtain the values of $m_1, m_2, \ldots, m_{n-1}$. However, in practice, this is difficult to do without some judgment. It is suggested that the natural spline is a good place to start. If more curvature at the ends is wanted, it can be added using this procedure.

Other endpoint constraints may be a bit more complicated and may require modification of the first and last of the system of equations (15.14), which will result in changes in the matrix $\mathbf{H}$ and the vector $\mathbf{v}$ in (15.17).

Case 3: Parabolic Runout Spline ($m_0 = m_1, m_n = m_{n-1}$)

Reducing the cubic functions on the first and last intervals to quadratics adds two more constraints, $d_0 = 0$ and $d_n = 0$. This results in the second derivatives being identical at both ends of the first and last intervals; that is, $m_0 = m_1$ and $m_n = m_{n-1}$. As a result, the first and last equations of (15.14) are replaced by

$$\begin{aligned} (3h_0 + 2h_1)m_1 + h_1 m_2 &= u_1, \\ h_{n-2}m_{n-2} + (2h_{n-2} + 3h_{n-1})m_{n-1} &= u_{n-1}. \end{aligned} \tag{15.19}$$

Case 4: Cubic Runout Spline

This method requires the cubic over $[x_0, x_1]$ to be an extension of that over $[x_1, x_2]$, thus imposing the same cubic function over the entire interval $[x_0, x_2]$. This is also known as the *not-a-knot* condition. A similar condition is imposed at the other end.

This can be achieved by requiring that the third derivatives at the endpoints also agree at x_1 and x_{n-1}; that is,

$$f_0'''(x_1) = f_1'''(x_1)$$

and

$$f_{n-2}'''(x_{n-1}) = f_{n-1}'''(x_{n-1}).$$

Because the third derivative is then constant throughout $[x_0, x_2]$ and also throughout $[x_{n-2}, x_n]$, the second derivative will be a linear function throughout the same two intervals. Hence, the slope of the second derivative will be the same in any subintervals within $[x_0, x_2]$ and within $[x_{n-2}, x_n]$. Thus, we can write

$$\begin{aligned} \frac{m_1 - m_0}{h_0} &= \frac{m_2 - m_1}{h_1}, \\ \frac{m_n - m_{n-1}}{h_{n-1}} &= \frac{m_{n-1} - m_{n-2}}{h_{n-2}}, \end{aligned}$$

or equivalently

$$\begin{aligned} m_0 &= m_1 - \frac{h_0(m_2 - m_1)}{h_1}, \\ m_n &= m_{n-1} + \frac{h_{n-1}(m_{n-1} - m_{n-2})}{h_{n-2}}. \end{aligned} \tag{15.20}$$

Then the first and last equations of (15.14) are replaced by

$$\begin{aligned} \left(3h_0 + 2h_1 + \frac{h_0^2}{h_1}\right)m_1 + \left(h_1 - \frac{h_0^2}{h_1}\right)m_2 &= u_1, \\ \left(h_{n-2} - \frac{h_{n-1}^2}{h_{n-2}}\right)m_{n-2} + \left(2h_{n-2} + 3h_{n-1} + \frac{h_{n-1}^2}{h_{n-2}}\right)m_{n-1} &= u_{n-1}. \end{aligned} \tag{15.21}$$

Case 5: Clamped Cubic Spline

This procedure fixes the slope $f_0'(x_0)$ and $f_{n-1}'(x_n)$ of the spline at each endpoint. In this case, from (15.11) and (15.12), the second derivatives are

$$\begin{aligned} m_0 &= \frac{3}{h_0}\left(\frac{y_1-y_0}{h_0} - f_0'(x_0)\right) - \frac{m_1}{2}, \\ m_n &= \frac{3}{h_{n-1}}\left(f_{n-1}'(x_n) - \frac{y_n-y_{n-1}}{h_{n-1}}\right) - \frac{m_{n-1}}{2}. \end{aligned} \tag{15.22}$$

As a result the first and last equations of (15.14) are replaced by

$$\left(\tfrac{3}{2}h_0 + 2h_1\right)m_1 + h_1 m_2 = u_1 - 3\left(\frac{y_1-y_0}{h_0} - f_0'(x_0)\right),$$

and

$$h_{n-2}m_{n-2} + \left(2h_{n-2} + \tfrac{3}{2}h_{n-1}\right)m_{n-1} = u_{n-1} - 3\left(f_{n-1}'(x_n) - \frac{y_n-y_{n-1}}{h_{n-1}}\right)$$

respectively.

Example 15.3 *From first principles, using conditions I–V, obtain the cubic spline through the points* $(2,50), (4,25)$, *and* $(5,20)$ *with the clamped boundary conditions* $f'(2) = -25$ *and* $f'(5) = -4$.

Let the cubic spline in the interval from $x_0 = 2$ to $x_1 = 4$ be the polynomial

$$f_0(x) = 50 + b_0(x-2) + c_0(x-2)^2 + d_0(x-2)^3$$

and the spline in the interval from $x_1 = 4$ to $x_2 = 5$ be the polynomial

$$f_1(x) = 25 + b_1(x-4) + c_1(x-4)^2 + d_1(x-4)^3.$$

The six coefficients $b_0, c_0, d_0, b_1, c_1, d_1$ are the unknowns that we need to determine. From the interpolation conditions

$$\begin{aligned} f_0(4) &= 50 + 2b_0 + 4c_0 + 8d_0 = 25, \\ f_1(5) &= 25 + b_1 + c_1 + d_1 = 20. \end{aligned}$$

From the smoothness conditions at $x = 4$

$$\begin{aligned} f_0'(4) &= b_0 + 2c_0(4-2) + 3d_0(4-2)^2 = f_1'(4) = b_1, \\ f_0''(4) &= 2c_0 + 6d_0(4-2) = f_1''(4) = 2c_1. \end{aligned}$$

Finally, from the boundary conditions, we get

$$\begin{aligned} f_0'(2) &= b_0 = -25, \\ f_1'(5) &= b_1 + 2c_1 + 3d_1 = -4. \end{aligned}$$

Thus, we have six linear equations to determine the six unknowns. In matrix form, the equations are

$$\begin{bmatrix} 2 & 4 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 4 & 12 & -1 & 0 & 0 \\ 0 & 2 & 12 & 0 & -2 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} b_0 \\ c_0 \\ d_0 \\ b_1 \\ c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} -25 \\ -5 \\ 0 \\ 0 \\ -25 \\ -4 \end{bmatrix}.$$

The equations can be solved by successive elimination of unknowns. We get $b_0 = -25$, then

$$\begin{bmatrix} 4 & 8 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 4 & 12 & -1 & 0 & 0 \\ 2 & 12 & 0 & -2 & 0 \\ 0 & 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} c_0 \\ d_0 \\ b_1 \\ c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} 25 \\ -5 \\ 25 \\ 0 \\ -4 \end{bmatrix}.$$

Take $c_0 = 6.25 - 2d_0$, then

$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 4 & -1 & 0 & 0 \\ 8 & 0 & -2 & 0 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} d_0 \\ b_1 \\ c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \\ -12.5 \\ -4 \end{bmatrix}.$$

Take $d_0 = 0.25b_1$, then

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & -2 & 0 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} b_1 \\ c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} -5 \\ -12.5 \\ -4 \end{bmatrix}.$$

Take $b_1 = -6.25 + c_1$, then

$$\begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} 1.25 \\ 2.25 \end{bmatrix}.$$

Finally, take $c_1 = 0.625 - 0.5d_1$ and get $d_1 = 0.25$. The final answer is

$$\begin{aligned} b_0 &= -25 \\ c_0 &= 9.125 \\ d_0 &= -1.4375 \\ b_1 &= -5.75 \\ c_1 &= 0.5 \\ d_1 &= 0.25. \end{aligned}$$

Thus the final interpolating cubic spline is

$$f(x) = \begin{cases} 50 - 25(x-2) + 9.125(x-2)^2 - 1.4375(x-2)^3, & 2 \le x \le 4, \\ 25 - 5.75(x-4) + 0.5(x-4)^2 + 0.25(x-4)^3, & 4 \le x \le 5. \end{cases}$$

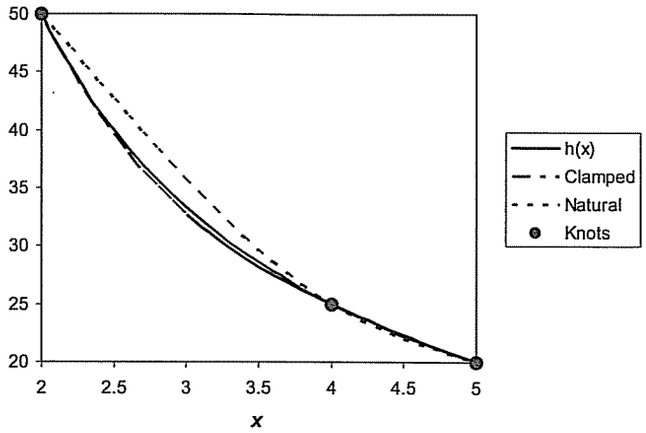


Fig. 15.7 Clamped and natural splines for Example 15.3.

Figure 15.7 shows the interpolating cubic spline and the corresponding natural cubic spline that is the solution of Exercise 15.3. It also shows the function $h(x) = 100/x$, which also passes through the same three knots. The slope of the clamped spline at the endpoints is the same as the slope of the function $h(x)$. These endpoint conditions force the clamped spline to be much closer to the function $h(x)$ than the natural spline. The natural spline has endpoint conditions that force the spline to look more like a straight line near the ends due to requiring the second derivative to be zero at the endpoints. □

The cubic splines in this section all pass through the knots. If smoothing is desired, that restriction may be lifted. Smoothing splines are introduced in Section 15.6.

Example 15.4 *The data in the last column of Table* 15.1 *are one-year mortality rates for the* 15 *five-year age intervals shown in the first column. The last interval is treated as* 95–99. *We have used a natural cubic spline to interpolate between these values as follows. The listed mortality rate is treated as the one-year mortality rate for the middle age within the five-year interval. The resulting values are treated as knots for a natural cubic spline. The fitted interpolating cubic spline is shown in Figure* 15.8 *on a logarithmic scale. The formula for the spline is given in Property I of Definition* 15.2. *The coefficients of the* 14 *cubic segments of the spline are given in Table* 15.2.

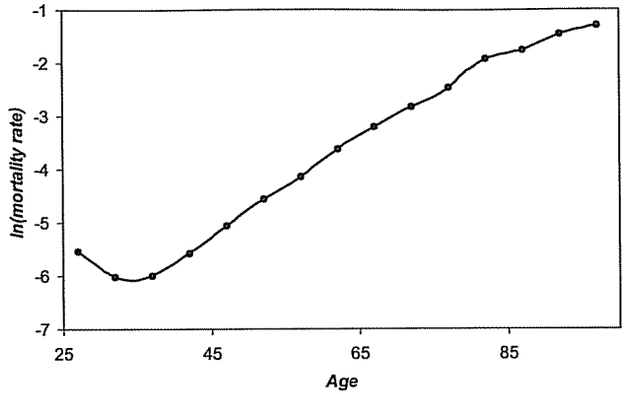


Fig. 15.8 Cubic spline fit to mortality data for Example 15.4.

Table 15.2 Spline coefficients for Example 15.4

j	x_j	a_j	b_j	c_j	d_j
0	27	3.8936×10^{-3}	-3.5093×10^{-4}	0	2.5230×10^{-6}
1	32	2.4543×10^{-3}	-1.6171×10^{-4}	3.7844×10^{-5}	-8.4886×10^{-7}
2	37	2.4857×10^{-3}	1.5307×10^{-4}	2.5112×10^{-5}	-5.1079×10^{-7}
3	42	3.8150×10^{-3}	3.6587×10^{-4}	1.7450×10^{-5}	2.0794×10^{-6}
4	47	6.3405×10^{-3}	6.9632×10^{-4}	4.8640×10^{-5}	-4.3460×10^{-6}
5	52	1.0495×10^{-2}	8.5678×10^{-4}	-1.6550×10^{-5}	1.2566×10^{-5}
6	57	1.5936×10^{-2}	1.6337×10^{-3}	1.7194×10^{-4}	-1.1922×10^{-5}
7	62	2.6913×10^{-2}	2.4590×10^{-3}	-6.8828×10^{-6}	1.3664×10^{-5}
8	67	4.0744×10^{-2}	3.4150×10^{-3}	1.9808×10^{-4}	-2.5761×10^{-5}
9	72	5.9551×10^{-2}	3.4638×10^{-3}	-1.8833×10^{-4}	1.1085×10^{-4}
10	77	8.6018×10^{-2}	9.8939×10^{-3}	1.4744×10^{-3}	-2.1542×10^{-4}
11	82	1.4542×10^{-1}	8.4813×10^{-3}	-1.7569×10^{-3}	2.2597×10^{-4}
12	87	1.7215×10^{-1}	7.8602×10^{-3}	1.6327×10^{-3}	-1.7174×10^{-4}
13	92	2.3080×10^{-1}	1.1306×10^{-2}	-9.4349×10^{-4}	6.2899×10^{-5}

15.3.2 Exercises

15.3 Repeat Example 15.3 for the natural cubic spline by removing the clamped spline boundary conditions.

15.4 Construct a natural cubic spline through the points $(-2,0)$, $(-1,1)$, $(0,0)$, $(1,1)$, and $(2,0)$ by setting up the system of equations (15.16).

15.5 Determine if the following functions can be cubic splines:

(a)

$$f(x) = \begin{cases} x, & -4 \le x \le 0, \\ x^3 + x, & 0 \le x \le 1, \\ 3x^2 - 2x + 1, & 1 \le x \le 9. \end{cases}$$

(b)

$$f(x) = \begin{cases} x^3, & 0 \le x \le 1, \\ 3x^2 - 3x + 1, & 1 \le x \le 2, \\ x^3 - 4x^2 + 13x - 11, & 2 \le x \le 4. \end{cases}$$

(c)

$$f(x) = \begin{cases} x^3 + 2x, & -1 \le x \le 0, \\ 2x^2 + 2x, & 0 \le x \le 1, \\ x^3 - x^2 + 5x - 1, & 1 \le x \le 3. \end{cases}$$

15.6 Determine the coefficients a, b, and c so that

$$f(x) = \begin{cases} x^3 + 4, & 0 \le x \le 1, \\ a + b(x-1) + c(x-1)^2 + 4(x-1)^3, & 1 \le x \le 3. \end{cases}$$

is a cubic spline.

15.7 Determine the clamped cubic spline that agrees with $\sin(x\pi/2)$ at $x = -1, 0, 1$.

15.8 Consider the function

$$f(x) = \begin{cases} 28 + 25x + 9x^2 + x^3, & -3 \le x \le -1, \\ 26 + 19x + 3x^2 - x^3, & -1 \le x \le 0, \\ 26 + 19x + 3x^2 - 2x^3, & 0 \le x \le 3, \\ -163 + 208x - 60x^2 + 5x^3, & 3 \le x \le 4. \end{cases}$$

(a) Prove that $f(x)$ can be a cubic spline.

(b) Determine which of the five endpoint conditions could have been used in developing this spline.

15.4 APPROXIMATING FUNCTIONS WITH SPLINES

The natural and clamped cubic splines have a particularly desirable property when the spline is considered to be an approximation to some other continuous function. For example, consider the function

$$h(x) = \frac{100}{x}, \quad 2 \le x \le 5.$$

This function collocates with the knots at $x = 2, 4, 5$ in Example 15.3. Let us suppose the knots had indeed come from this function. Then, we could

consider the interpolating cubic spline to be an approximation to the function $h(x)$. In many applications, such as computer graphics, where smooth images are needed, those smooth images can be represented very efficiently using a limited number of selected knots and a cubic spline interpolation algorithm.

Smoothness can be measured by the total curvature of a function. The most popular of such measures is the squared norm

$$S = \int_{x_0}^{x_n} [f''(x)]^2 dx, \tag{15.23}$$

representing the total squared second derivative.

Now consider any continuous function $h(x)$ that also has continuous first and second derivatives over some interval $[x_0, x_n]$. Suppose that we select $n-1$ interior knots $\{x_j, h(x_j)\}_{j=1}^{n-1}$ with $x_0 < x_1 < x_2 < \cdots < x_n$.

Let $f(x)$ be a cubic spline that collocates with these knots and has endpoint conditions either

$$f'(x_0) = h'(x_0) \text{ and } f'(x_n) = h'(x_n) \quad \text{(clamped spline)}$$

or

$$f''(x_0) = 0 \text{ and } f''(x_n) = 0. \quad \text{(natural spline).}$$

The natural or clamped cubic spline $f(x)$ has less total curvature than any other function $h(x)$ passing through the $n+1$ knots, as shown in the following theorem.

Theorem 15.5 *Let $f(x)$ be the natural or clamped cubic spline passing through the $n+1$ given knots. Let $h(x)$ be any function with continuous first and second derivatives that passes through the same knots. Also, for the clamped cubic spline assume $h'(x_0) = f'(x_0)$ and $h'(x_n) = f'(x_n)$. Then*

$$\int_{x_0}^{x_n} [f''(x)]^2 dx \le \int_{x_0}^{x_n} [h''(x)]^2 dx. \tag{15.24}$$

Proof: Let us form the difference $D(x) = h(x) - f(x)$. Then, $D''(x) = h''(x) - f''(x)$ and therefore

$$[h''(x)]^2 = [f''(x)]^2 + [D''(x)]^2 + 2f''(x)D''(x).$$

Integrating both sides produces

$$\int_{x_0}^{x_n} [h''(x)]^2 dx = \int_{x_0}^{x_n} [f''(x)]^2 dx + \int_{x_0}^{x_n} [D''(x)]^2 dx + 2\int_{x_0}^{x_n} f''(x)D''(x)dx.$$

The result will be proven if we can show that

$$\int_{x_0}^{x_n} f''(x)D''(x)dx = 0$$

because the total curvature of the function $h(x)$,

$$\int_{x_0}^{x_n} [h''(x)]^2 dx,$$

will be equal to the total curvature of the spline,

$$\int_{x_0}^{x_n} [f''(x)]^2 dx$$

plus a nonnegative quantity

$$\int_{x_0}^{x_n} [D''(x)]^2 dx.$$

Applying integration by parts, we get

$$\int_{x_0}^{x_n} f''(x)D''(x)dx = f''(x)D'(x)\Big|_{x_0}^{x_n} - \int_{x_0}^{x_n} f'''(x)D'(x)dx.$$

For the clamped cubic spline, the first term is zero because the clamped boundary conditions imply that

$$\begin{aligned} D'(x_0) &= h'(x_0) - f'(x_0) = 0, \\ D'(x_n) &= h'(x_n) - f'(x_n) = 0. \end{aligned}$$

For the natural cubic spline $f''(x_0) = f''(x_n) = 0$, which also makes the first term zero.

The integral in the second term can be divided into subintervals as follows:

$$\int_{x_0}^{x_n} f'''(x)D'(x)dx = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} f'''(x)D'(x)dx.$$

Integration by parts in each subinterval yields

$$\int_{x_j}^{x_{j+1}} f'''(x)D'(x)dx = f'''(x)D(x)|_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} f^{(4)}(x)D(x)dx.$$

The first term is zero because of the interpolation condition

$$D(x_j) = h(x_j) - f(x_j) = 0, \quad j = 0, 1, \ldots, n.$$

That is, we are only considering functions $h(x)$ that pass through the knots.

The second term is zero because the spline $f(x)$ in each subinterval is a cubic polynomial and has zero fourth derivative. Thus, for the clamped or natural cubic spline,

$$\int_{x_0}^{x_n} f''(x)D''(x)dx = 0,$$

which proves the result. □

Thus the clamped cubic spline has great appeal if we want to produce a smooth set of successive values and if we have some knowledge of the slope of the function at each end of the interval. This is often the case in mortality table construction. At very early ages in the first few days and weeks of life, the force of mortality or hazard rate decreases sharply as a result of deaths of newborn lives with congenital and other conditions that contribute to neonatal deaths. At the highest ages, the force of mortality tends to flatten out at a level of between 0.3 and 0.4 at ages well over 100. Using a clamped cubic spline to graduate observed rates will result in obtaining the smoothest possible function that incorporates the desired properties at each end of the age spectrum. If the mortality data are only over some more limited age range (as is usually the case with life insurance or annuity data), either natural or clamped cubic splines can be used. Including a clamping condition controls the slope at the endpoints.

Example 15.6 *For the clamped cubic spline obtained in Example* 15.3 *calculate the value of the squared norm measure of curvature. Calculate the same quantity for the function* $h(x) = 100/x$ *which also passes through the given knots.*

The spline function is

$$f(x) = \begin{cases} 50 - 25(x-2) + 9.125(x-2)^2 - 1.4375(x-2)^3, & 2 \le x \le 4, \\ 25 - 5.75(x-4) + 0.5(x-4)^2 + 0.25(x-4)^3, & 4 \le x \le 5, \end{cases}$$

and the second derivative is

$$f''(x) = \begin{cases} 18.25 - 8.625(x-2) = 35.5 - 8.625x, & 2 \le x \le 4, \\ 1 + 1.5(x-4) = 1.5x - 5 & 4 \le x \le 5. \end{cases}$$

The total curvature of the spline is

$$\begin{aligned} \int_2^5 [f''(x)]^2 dx &= \int_2^4 (35.5 - 8.625x)^2 dx + \int_4^5 (1.5x - 5)^2 dx \\ &= \int_1^{18.25} y^2 \frac{1}{8.625} dy + \int_1^{2.5} y^2 \frac{1}{1.5} dy \\ &= \frac{y^3}{25.875}\Big|_1^{18.25} + \frac{y^3}{4.5}\Big|_1^{2.5} = 238.125. \end{aligned}$$

For $h(x)$, the second derivative is $h''(x) = 200x^{-3}$ and the curvature is

$$\begin{aligned} \int_2^5 (200x^{-3})^2 dx &= \int_2^5 40{,}000x^{-6} dx \\ &= -8{,}000x^{-5}\Big|_2^5 \\ &= 247.44. \end{aligned}$$

Notice how close the total curvature of the function $h(x)$ and the clamped spline are. Now look at Figure 15.7, which plots both functions. They are very similar in shape. Hence we would expect them to have similar curvature. Of course, as a result of Theorem 15.5, the curvature of the spline should be less, though in this case it is only slightly less. In Exercise 15.9 you are asked to calculate the total curvature of the corresponding natural spline (which also appears in Figure 15.7). Because it is much "straighter," you would expect its total curvature to be significantly less, which is confirmed in Exercise 15.9. □

15.4.1 Exercise

15.9 For the natural cubic spline obtained in Exercise 15.3 calculate the value of the squared norm measure of curvature.

15.5 EXTRAPOLATING WITH SPLINES

In many applications we may want to produce a model that can be faithful to a set of historical data but that can also be used to forecast into the future. For example, in determining liabilities of an insurer when future claim payments are subject to inflationary growth, the actuary may need to project the rate of future claims inflation for some 5 to 10 years into the future. One way to do this is by fitting a function, in this case a cubic spline, to historic claims inflation data.

Simply projecting the cubic in the last interval beyond x_n may result in excessive oscillatory behavior in the region beyond x_n. This could result in projected values that are wildly unreasonable. It makes much more sense to require projected values to form a simple pattern. In particular, a linear projection is likely to be reasonable in most practical situations. This is easily handled by cubic splines.

The natural cubic spline has endpoint conditions that require the second derivatives to be zero at the endpoints. The natural extrapolation is linear with the slope coming from the endpoints. Of course, the linear extrapolation function can be done for any spline using the first derivative at the end points. However, unless the second derivative is zero, as with the natural spline, the second derivative condition will be violated at the endpoints. The extrapolated values at each end are then

$$\begin{aligned} f(x) &= f(x_n) + f'(x_n)(x - x_n), \ x > x_n, \\ f(x) &= f(x_0) - f'(x_0)(x_0 - x), \ x < x_0. \end{aligned}$$

Example 15.7 *Obtain formulas for the extrapolated values for the clamped spline in Example* 15.3 *and determine the extrapolated values at* $x = 0$ *and* $x = 7$.

In the first interval, $f(x) = 50 - 25(x-2) + 9.125(x-2)^2 - 1.4375(x-2)^3$ and so $f(2) = 50$ and $f'(2) = -25$. Then, for $x < 2$, the extrapolation is $f(x) = 50 - (-25)(2 - x) = 100 - 25x$. In the final interval, $f(x) = 25 - 5.75(x-4) + 0.5(x-4)^2 + 0.25(x-4)^3$ and so $f(5) = 20$ and $f'(5) = -4$. Then, for $x > 5$, the extrapolation is $f(x) = 20 - 4(x-5) = 40 - 4x$. At $x = 0$ the extrapolated value is $100 - 25(0) = 100$ and at $x = 7$ it is $40 - 4(7) = 12$. □

15.5.1 Exercise

15.10 Obtain formulas for the extrapolated values for the natural spline in Exercise 15.3 and determine the extrapolated values at $x = 0$ and $x = 7$.

15.6 SMOOTHING SPLINES

In many actuarial applications, it may be desirable to do more than interpolate between observed data. If data include a random (or "noise") element, it is often best to allow the cubic spline or other smooth function to lie near the data points, rather than requiring the function to pass through each data point.

In the terminology of graduation theory as developed by actuaries in the early 1900s, this is called **modified osculatory interpolation**. The term *modified* is added to recognize that the points of intersection (or knots in the language of splines) are modified from the original data points.

The technical development of smoothing cubic splines is identical to interpolating cubic splines except that the original knots at each data point (x_i, y_i) are replaced by knots at (x_j, a_j) where the ordinate a_j is the constant term in the smoothing cubic spline

$$f_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3. \tag{15.25}$$

We first imagine that the ordinates of original data points are the outcomes of the model

$$y_j = g(x_j) + \epsilon_j,$$

where ϵ_j, $j = 0, 1, \ldots, n$, are independently distributed random variables with mean 0 and variance σ_j^2 and where $g(x)$ is a well-behaved function.[3]

Example 15.8 *Mortality rates* q_j *at each age* j *are estimated by the ratio of observed deaths to the number of life-years of exposure* D_j/n_j, *where* D_j *is a binomial* (n_j, q_j) *random variable. The estimator* $\hat{q}_j = d_j/n_j$, *where* d_j *is*

[3] Without specifying what "well-behaved" means in technical terms, we are simply trying to say that $g(x)$ is smooth in a general way. Typically we will require at least the first two derivatives to be continuous.

the observed number of deaths, has variance $\sigma_j^2 = q_j(1-q_j)/n_j$, *which can be estimated by* $\hat{q}_j(1-\hat{q}_j)/n_j$.

We attempt to find a smooth function $f(x)$, in this case a cubic spline, that will serve as an approximation to the "true" function $g(x)$. Because $g(x)$ is assumed to be well behaved, we will require the smoothing cubic spline $f(x)$ itself to be as smooth as possible. On the other hand, we want it to be faithful to the given data as much as possible. These are conflicting objectives. Therefore, a compromise will be necessary between fit and smoothness.

The degree of fit can be measured using the chi-square criterion

$$F = \sum_{j=0}^{n} \left(\frac{y_j - a_j}{\sigma_j} \right)^2. \tag{15.26}$$

This is a standard statistical criterion for measuring the degree of fit and was discussed in that context in Section 13.4.3. It has a chi-square distribution with $n+1$ degrees of freedom.[4]

The degree of smoothness can be measured by the overall smoothness of the cubic spline. The smoothness, or equivalently the total curvature, can be measured by the squared norm smoothness criterion

$$S = \int_{x_0}^{x_n} [f''(x)]^2 dx.$$

It was shown in Theorem 15.5 that within the broad class of functions with continuous first and second derivatives, the natural or clamped cubic spline minimizes the squared norm. This supports the choice of the cubic spline as the smoothing function.

In order to recognize the conflicting objectives of fit and smoothness, we construct a criterion which is a weighted average of the measures of fit and smoothness. Let

$$\begin{aligned} L &= pF + (1-p)S \\ &= p \sum_{j=0}^{n} \left(\frac{y_j - a_j}{\sigma_j} \right)^2 + (1-p) \int_{x_0}^{x_n} [f''(x)]^2 dx. \end{aligned}$$

The parameter p reflects the relative importance which we give to the conflicting objectives of remaining close to the data, on the one hand, and of obtaining a smooth curve, on the other hand. Notice that a linear function satisfies the equation

$$S = \int_{x_0}^{x_n} [f''(x)]^2 dx = 0,$$

[4]No degree of freedom is lost, because unlike with the goodness-of-fit test, if you know all but one of the terms of the sum, it is not possible to infer the remaining value.

which suggests that, in the limiting case, where $p = 0$ and thus smoothness is all that matters, the spline function $f(x)$ will become a straight line. At the other extreme, where $p = 1$ and thus the closeness of the spline to the data is all that matters, we will obtain an interpolating spline which passes exactly through the data points.

The spline is piecewise cubic and thus the smoothness criterion can be written

$$S = \int_{x_0}^{x_n} [f''(x)]^2 dx = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} [f_j''(x)]^2 dx.$$

From (15.5),

$$f_j''(x) = \frac{m_j}{h_j}(x_{j+1} - x) + \frac{m_{j+1}}{h_j}(x - x_j),$$

and then

$$\begin{aligned} \int_{x_j}^{x_{j+1}} [f_j''(x)]^2 dx &= \int_{x_j}^{x_{j+1}} \left[\frac{m_j}{h_j}(x_{j+1} - x) + \frac{m_{j+1}}{h_j}(x - x_j) \right]^2 dx \\ &= \int_0^1 [m_j(1-y) + m_{j+1}y]^2 \, h_j \, dy \\ &= h_j \int_0^1 [m_j + (m_{j+1} - m_j)y]^2 \, dy \\ &= h_j \left. \frac{[m_j + (m_{j+1} - m_j)y]^3}{3(m_{j+1} - m_j)} \right|_0^1 \\ &= \frac{h_j}{3}(m_j^2 + m_j m_{j+1} + m_{j+1}^2), \end{aligned}$$

where the substitution $y = (x - x_j)/h_j$ is used in the second line. The criterion function then becomes

$$L = p \sum_{j=0}^{n} \left(\frac{y_j - a_j}{\sigma_j} \right)^2 + (1-p) \sum_{j=0}^{n-1} \frac{h_j}{3}(m_j^2 + m_j m_{j+1} + m_{j+1}^2).$$

We need to minimize this function with respect to the $2n+2$ unknown quantities $\{a_j, m_j; j = 0, \ldots, n\}$. Note that when we have solved for these variables we will have four pieces of information $\{a_j, a_{j+1}, m_j, m_{j+1}\}$ for each interval $[x_j, x_{j+1}]$. This allows us to fully specify the interpolating cubic spline in each interval. We now address the issue of solving for these quantities.

We now consider the natural smoothing spline. The equations developed for interpolating splines apply to smoothing cubic splines except that the y_j's are replaced by a_j's to recognize that the abscissas $\{a_j; j = 0, \ldots, n\}$ of the smoothing splines do not pass through the abscissas of the data points $\{y_j; j = 0, \ldots, n\}$. From (15.16), we can write

$$\mathbf{Hm} = \mathbf{u},$$

where $\mathbf{m} = (m_1, m_2, \ldots, m_{n-1})^T$ and $\mathbf{u} = (u_1, u_2, \ldots, u_{n-1})^T$ because $m_0 = m_n = 0$ from the natural spline condition. From (15.14), the vector $\mathbf{u}$ can be rewritten as

$$\mathbf{u} = \mathbf{R}\mathbf{a},$$

where $\mathbf{R}$ is the $(n-1) \times (n+1)$ matrix

$$\mathbf{R} = \begin{bmatrix} r_0 & -(r_0+r_1) & r_1 & 0 & \ldots & \ldots & 0 \\ 0 & r_1 & -(r_1+r_2) & r_2 & 0 & \ldots & 0 \\ \ddots & & \ddots & & \ddots & & \ddots \\ 0 & \ldots & \ldots & 0 & r_{n-2} & -(r_{n-2}+r_{n-1}) & r_{n-1} \end{bmatrix}$$

and

$$\mathbf{a} = (a_0, a_1, \ldots, a_n)^T, \quad r_j = 6h_j^{-1}.$$

Then we have

$$\mathbf{H}\mathbf{m} = \mathbf{R}\mathbf{a}. \tag{15.27}$$

We can now rewrite the criterion L as

$$L = p(\mathbf{y}-\mathbf{a})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{a}) + \tfrac{1}{6}(1-p)\mathbf{m}^T\mathbf{H}\mathbf{m}$$

where $\boldsymbol{\Sigma} = \text{diag}\{\sigma_0^2, \sigma_1^2, \ldots, \sigma_n^2\}$. Because $\mathbf{m} = \mathbf{H}^{-1}\mathbf{R}\mathbf{a}$, we can rewrite the criterion as

$$L = p(\mathbf{y}-\mathbf{a})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{a}) + \tfrac{1}{6}(1-p)\mathbf{a}^T\mathbf{R}^T\mathbf{H}^{-1}\mathbf{R}\mathbf{a}.$$

We can differentiate the criterion L with respect to each of $a_0, a_1, \ldots, a_n$ successively to obtain the optimal values of the ordinates. In matrix notation, the result is (after dividing the derivative by 2)

$$-p(\mathbf{y}-\mathbf{a})^T\boldsymbol{\Sigma}^{-1} + \tfrac{1}{6}(1-p)\mathbf{a}^T\mathbf{R}^T\mathbf{H}^{-1}\mathbf{R} = \mathbf{0},$$

where $\mathbf{0}$ is the $(n+1) \times 1$ vector of zeros, $(0, \ldots, 0)^T$. This yields, after transposition,

$$6p\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{a}) = (1-p)\mathbf{R}^T\mathbf{H}^{-1}\mathbf{R}\mathbf{a}$$

or

$$6p\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{a}) = (1-p)\mathbf{R}^T\mathbf{m}. \tag{15.28}$$

We now premultiply by $\mathbf{R}\boldsymbol{\Sigma}$, yielding

$$6p\mathbf{R}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{a}) = (1-p)\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^T\mathbf{m}$$

or

$$6p(\mathbf{R}\mathbf{y}-\mathbf{R}\mathbf{a}) = (1-p)\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^T\mathbf{m}. \tag{15.29}$$

Because $\mathbf{H}\mathbf{m} = \mathbf{R}\mathbf{a}$, this reduces to

$$p\mathbf{R}\mathbf{y} - p\mathbf{H}\mathbf{m} = \tfrac{1}{6}(1-p)\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^T\mathbf{m}$$

or

$$\left(p\mathbf{H} + \tfrac{1}{6}(1-p)\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^T\right)\mathbf{m} = p\mathbf{R}\mathbf{y}. \tag{15.30}$$

This is a system of $n-1$ equations in $n-1$ unknowns. The system of equations can be solved for $m_1, m_2, \ldots, m_{n-1}$. Using matrix methods, the solution can be obtained from (15.30) as

$$\mathbf{m} = \left(\mathbf{H} + \frac{1-p}{6p}\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^T\right)^{-1}\mathbf{R}\mathbf{y}. \tag{15.31}$$

Now, the values of $a_0, a_1, \ldots, a_n$ can be obtained by rewriting (15.28) as

$$\mathbf{a} = \mathbf{y} - \frac{1-p}{6p}\boldsymbol{\Sigma}\mathbf{R}^T\mathbf{m}. \tag{15.32}$$

Finally, substitution of (15.31) into (15.32) results in

$$\mathbf{a} = \mathbf{y} - \frac{1-p}{6p}\boldsymbol{\Sigma}\mathbf{R}^T\left(\mathbf{H} + \frac{1-p}{6p}\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^T\right)^{-1}\mathbf{R}\mathbf{y}. \tag{15.33}$$

Thus we have obtained the values of the intercepts of the n cubic spline segments of the smoothing spline. The values of the other coefficients of the spline segments can now be calculated in the same way as for the natural interpolating spline, as discussed in Section 15.3 using the knots $\{(x_j, a_j), j = 0, \ldots, n\}$ and setting $m_0 = m_n = 0$. It should be noted that the only additional calculation for the natural smoothing spline as compared with the natural interpolation spline is given by (15.33).

The magnitude of the values of the criteria for fit F and smoothness S may be very different. Therefore one should not place any significance on the specific choice of the value of p (unless it is 0 or 1). Smaller values of p result in more smoothing; larger values result in less. In some applications it may be necessary to make the value of p very small, for example, 0.001, to begin to get visual images with any significant amount of smoothing. This is, in part, due to the role of the variances which appear in the denominator of the fit criterion. Small variances can result in the fit term being much larger than the smoothness term. Therefore, it may be necessary to have a very small value for p to get any visible smoothing.

Example 15.9 *Construct natural cubic smoothing splines for the data in Table* 15.1. *The natural cubic interpolating spline through the mortality rates was shown in Figure* 15.8.

Natural cubic smoothing splines with $p = 0.5$ and $p = 0.1$ are shown in Figures 15.9 and 15.10. The coefficients for the smoothing spline with $p = 0.1$ are given in Table 15.3. Note that the resulting splines look much like the one in Figure 15.8 except near the upper end of the data where there are relatively fewer actual deaths and less smoothness in the successive observed

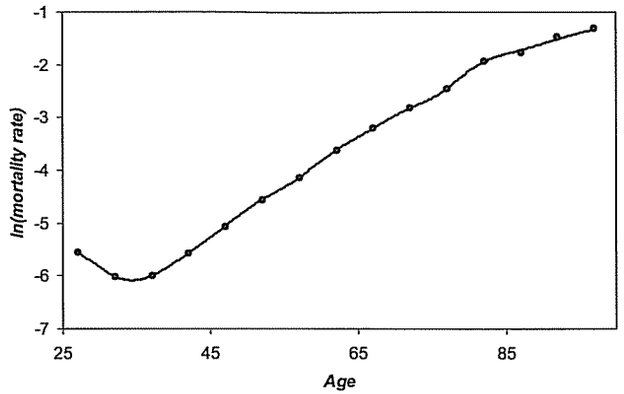


Fig. 15.9 Smoothing spline with $p = 0.5$ for Example 15.9.

values. Also observe the increased smoothness in the spline in Figure 15.10 resulting from the smaller emphasis on fit. The standard deviations were calculated as in Example 15.8 with the resulting values multiplied by 1,000 to make the numbers more reasonable.[5] □

Example 15.9 illustrates how the smoothing splines can be used to carry out both interpolation and smoothing automatically. The knots at quinquennial ages were smoothed using (15.32). The modified knots were then used as knots for an interpolating spline. The interpolated values are the revised mortality rates at the intermediate ages. The smoothing effect was not visually dramatic in Example 15.9 because the original data series was already quite smooth. The next example illustrates how successive values in a very noisy series can be smoothed dramatically using a smoothing spline.

Example 15.10 *Table* 15.4 *gives successive observed mortality rates for a* 15*-year period. The data can be found in Miller* [94], *p.* 11 *and are shown in Figure* 15.11. *Fit smoothing splines changing p until smoothing appears reasonable and provide values of the revised mortality rates at each age.*

Unlike Example 15.9, the numbers represent numbers of persons, not dollar amounts, and can be used directly in the estimates of the variances of the

[5] Had the values not been multiplied by 1,000 the same answers could have been obtained by altering the value of p. This method of calculating the standard deviations does not consider the possible variation in sizes of the insurance policies. See Klugman [75] for a more detailed treatment. The method used here implicitly treats all policies as being of the same size. That size is not important because as with the factor of 1,000, a constant of proportionality can be absorbed into p.

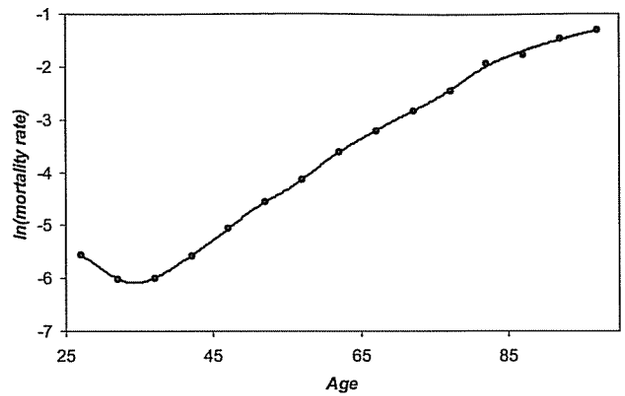


Fig. 15.10 Smoothing spline with $p = 0.1$ for Example 15.9.

Table 15.3 Spline coefficients for Example 15.9 with $p = 0.1$

j	x_j	a_j	b_j	c_j	d_j
0	27	3.8790×10^{-3}	-3.4670×10^{-4}	0	2.4846×10^{-6}
1	32	2.4560×10^{-3}	-1.6036×10^{-4}	3.7269×10^{-5}	-8.0257×10^{-7}
2	37	2.4856×10^{-3}	1.5214×10^{-4}	2.5230×10^{-5}	-5.0019×10^{-7}
3	42	3.8146×10^{-3}	3.6693×10^{-4}	1.7728×10^{-5}	1.9945×10^{-6}
4	47	6.3417×10^{-3}	6.9379×10^{-4}	4.7644×10^{-5}	-4.0893×10^{-6}
5	52	1.0491×10^{-2}	8.6353×10^{-4}	-1.3695×10^{-5}	1.1833×10^{-5}
6	57	1.5945×10^{-2}	1.6141×10^{-3}	1.6380×10^{-4}	-9.7024×10^{-6}
7	62	2.6898×10^{-2}	2.5244×10^{-3}	1.8268×10^{-5}	6.2633×10^{-6}
8	67	4.0759×10^{-2}	3.1769×10^{-3}	1.1222×10^{-4}	-9.4435×10^{-7}
9	72	5.9331×10^{-2}	4.2282×10^{-3}	9.8052×10^{-5}	3.9737×10^{-5}
10	77	8.7891×10^{-2}	8.1890×10^{-3}	6.9411×10^{-4}	-6.6804×10^{-5}
11	82	1.3784×10^{-1}	1.0120×10^{-2}	-3.0794×10^{-4}	2.1572×10^{-5}
12	87	1.8344×10^{-1}	8.6583×10^{-3}	1.5633×10^{-5}	-1.0282×10^{-6}
13	92	2.2699×10^{-1}	8.7376×10^{-3}	2.1021×10^{-7}	-1.4014×10^{-8}

mortality rates (see Example 15.8). The standard deviations are multiplied by a factor of 10 for convenience. For insurance purposes, we are more interested in the spline values at the knots, that is, the a_j. The interpolated values are given in Table 15.4 and the spline values are plotted in Figures 15.12–15.14 for $p = 0.5, 0.1, 0.05$. Note that for $p = 0.5$ there is significant smoothing but that some points still have a lot of influence on the result. For example the large number of actual deaths at age 76 causes the curve to be pulled upward.

Table 15.4 Mortality rates and interpolated values for Example 15.10

j	Age x_j	Exposed to risk	Observed deaths	Estimated mort. rate	Smoothed $p = 0.5$	Smoothed $p = 0.1$	Smoothed $p = 0.05$
0	70	135	6	0.044	0.046	0.050	0.052
1	71	143	12	0.084	0.078	0.069	0.065
2	72	140	10	0.071	0.077	0.071	0.069
3	73	144	11	0.076	0.066	0.064	0.065
4	74	149	6	0.040	0.049	0.062	0.066
5	75	154	16	0.104	0.100	0.084	0.080
6	76	150	24	0.160	0.126	0.096	0.089
7	77	139	8	0.058	0.076	0.087	0.088
8	78	145	16	0.110	0.091	0.091	0.093
9	79	140	13	0.093	0.102	0.105	0.107
10	80	137	19	0.139	0.131	0.128	0.128
11	81	136	21	0.154	0.157	0.155	0.154
12	82	126	23	0.183	0.182	0.181	0.181
13	83	126	26	0.206	0.208	0.209	0.208
14	84	109	26	0.239	0.238	0.237	0.236
Total		2,073	237				

More smoothing can be obtained by reducing p, as can be observed from the three figures. □

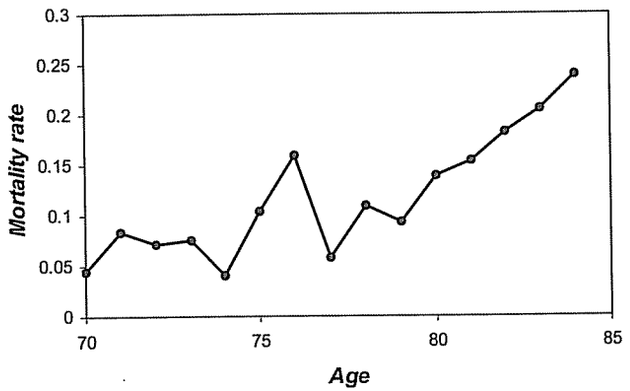


Fig. 15.11 Mortality data for Example 15.10.

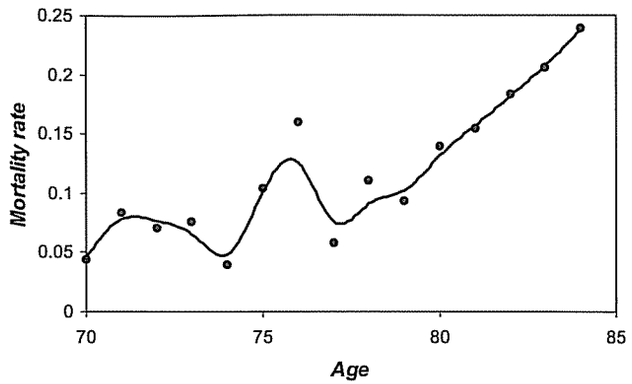


Fig. 15.12 Smoothing spline for mortality data with $p = 0.5$.

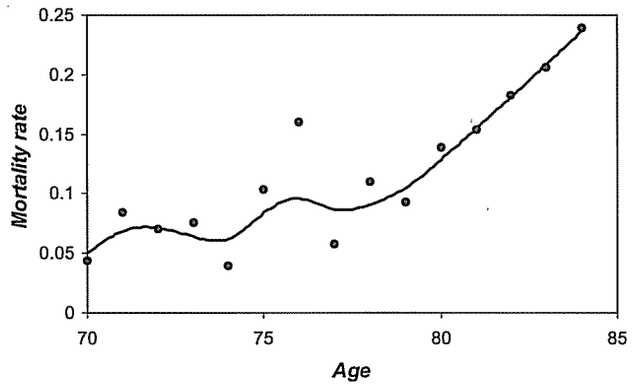


Fig. 15.13 Smoothing spline for mortality data with $p = 0.1$.

Example 15.10 demonstrated the smoothing capability of splines. However, one still needs to choose a value of p. In practice, this is done using professional judgment and visual inspection. If, as with Example 15.9, data sets are large and there is already some degree of smoothness in the observed data, then a fitted curve which closely follows the data is likely highly desirable. If the available data set is more limited, as with Example 15.10, considerable smoothing is needed and judgment plays a large role. For data sets of any size, formal tests of fit can be conducted. The fit criterion F has a chi-square distribution with $n + 1$ degrees of freedom. This can provide some guidance.

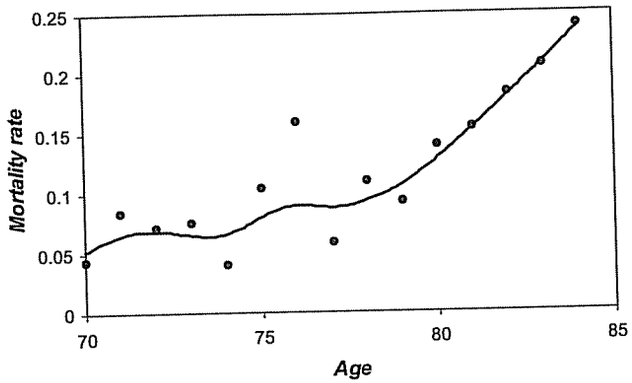


Fig. 15.14 Smoothing spline for mortality data with $p = 0.05$.

The choice of other tests of fit such as the runs test can be employed to identify specific anomalies of the fitted spline.

15.6.1 Exercise

15.11 Consider the natural cubic smoothing spline that smooths the points $(0,0)$, $(1,2)$, $(2,1)$, $(3,3)$ using $p = 0.9$ and standard deviations of 0.5. (Use a spreadsheet for the calculations.)

(a) Obtain the values of the intercepts of the nodes by using (15.33).

(b) Obtain the natural cubic smoothing spline as the natural interpolating spline through the nodes using (15.16) and (15.18).

(c) Graph the resulting spline from $x = -0.5$ to $x = 2.5$.

16

Credibility

16.1 INTRODUCTION

Credibility theory is a set of quantitative tools which allows an insurer to perform prospective experience rating (adjust future premiums based on past experience) on a risk or group of risks. If the experience of a policyholder is consistently better than that assumed in the underlying manual rate (sometimes called the **pure premium**), then the policyholder may demand a rate reduction.

The policyholder's argument is as follows: The manual rate is designed to reflect the expected experience of the entire rating class and implicitly assumes that the risks are homogeneous. However, no rating system is perfect, and there always remains some heterogeneity in the risk levels after all the underwriting criteria are accounted for. Consequently, some policyholders will be better risks than that assumed in the underlying manual rate. Of course, the same logic dictates that a rate increase should be applied to a poor risk, but the policyholder in this situation is certainly not going to ask for a rate increase! Nevertheless, an increase may be necessary, due to considerations of equity and the economics of the situation.

The insurer is then forced to answer the following question: How much of the difference in experience of a given policyholder is due to random variation

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

in the underlying claims experience and how much is due to the fact that the policyholder really is a better or worse risk than average for the given rating class? In other words, how credible is the policyholder's own experience? Two facts must be considered in this regard:

1. The more past information the insurer has on a given policyholder, the more *credible* the policyholder's own experience, all else being equal. In the same vein, in group insurance the experience of larger groups is more credible than that of smaller groups.

2. Competitive considerations may force the insurer to give full (using the past experience of the policyholder only and not the manual rate) or nearly full credibility to a given policyholder in order to retain the business.

Another use for credibility is in the setting of rates for classification systems. For example, in workers compensation insurance there may be hundreds of occupational classes, some of which may provide very little data. In order to accurately estimate the expected cost for insuring these classes, it may be appropriate to combine the limited actual experience with some other information, such as past rates, or the experience of occupations that are closely related.

From a statistical perspective, credibility theory leads to a result that would appear to be counterintuitive. If experience from an insured or group of insureds is available, our statistical training may convince us to use the sample mean or some other unbiased estimator. But credibility theory tells us that it is optimal to give only partial weight to this experience and give the remaining weight to an estimator produced from other information. We will discover that what we sacrifice in terms of bias we gain in terms of reducing the average (squared) error.

Credibility theory allows an insurer to quantitatively formulate the above problem, and this chapter provides an introduction to this theory. A few relevant statistical concepts are reviewed in the next section. Some topics were covered in Sections 9.2 and 12.4 are repeated and there are some new formulas and concepts as well.

Section 16.3 deals with **limited fluctuation credibility theory**, a subject developed in the early part of the twentieth century. This provides a mechanism for assigning full (Section 16.3.1) or partial (Section 16.3.2) credibility to a policyholder's experience. The difficulty with this approach is the lack of a sound underlying mathematical theory justifying the use of these methods. Nevertheless, this approach provided the original treatment of the subject and is still in use today.

A classic paper by Bühlmann in 1967 [18] provided a statistical framework within which credibility theory has developed and flourished. While this approach, termed **greatest accuracy credibility theory**,[1] was formalized by Bühlmann, the basic ideas were around for some time. This approach is introduced in Section 16.4. The simplest model, that of Bühlmann [18], is discussed in Section 16.4.4. Practical improvements were made by Bühlmann and Straub in 1970 [20]. Their model is discussed in Section 16.4.5. The concept of exact credibility is presented in Section 16.4.6.

Practical use of the theory requires that unknown model parameters be estimated from data. Nonparametric estimation (where the problem is somewhat model free and the parameters are generic, such as the mean and variance) is considered in Section 16.5.1, semiparametric estimation (where some of the parameters are based on assuming particular distributions) in Section 16.5.2, and finally the fully parametric situation (where all parameters come from assumed distributions) in Section 16.5.3.

We close with a quote from Arthur Bailey in 1950 [8], p. 8, that aptly summarizes much of the history of credibility. We, too, must tip our hats to the early actuaries, who, with unsophisticated mathematical tools at their disposal, were able to come up with formulas that not only worked but also were very similar to those we carefully develop in this chapter.

> It is at this point in the discussion that the ordinary individual has to admit that, while there seems to be some hazy logic behind the actuaries' contentions, it is too obscure for him to understand. The trained statistician cries "Absurd! Directly contrary to any of the accepted theories of statistical estimation." The actuaries themselves have to admit that they have gone beyond anything that has been proven mathematically, that all of the values involved are still selected on the basis of judgment, and that the only demonstration they can make is that, in actual practice, it works. Let us not forget, however, that they have made this demonstration many times. It does work!

16.2 STATISTICAL CONCEPTS

In this section various statistical concepts relevant to credibility theory are presented. Much of the material is of a review nature and hence may be quickly glossed over by a reader with a good background in statistics. Nevertheless, there may be some material which may not have been seen before, and so this section should not be completely ignored. Subsequent sections will refer back to this material.

[1] The terms *limited fluctuation* and *greatest accuracy* go back at least as far as a 1943 paper by Arthur Bailey [7].

16.2.1 Conditional distributions

Suppose that X and Y are two random variables with joint probability function (pf) or probability density function (pdf)[2] $f_{X,Y}(x,y)$ and marginal pfs $f_X(x)$ and $f_Y(y)$, respectively. The conditional pf of X given that $Y=y$ is

$$f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}. \tag{16.1}$$

If X and Y are discrete random variables, then (16.1) is the conditional probability of the event $X=x$ under the hypothesis that $Y=y$. If X and Y are continuous, then (16.1) may be interpreted as a definition. When X and Y are independent random variables,

$$f_{X,Y}(x,y)=f_X(x)f_Y(y),$$

and in this case, (16.1) yields

$$f_{X|Y}(x|y)=f_X(x),.$$

We observe that the conditional and marginal distributions of X are identical.

Example 16.1 *Suppose X and Z are independent Poisson random variables with means λ_1 and λ_2, respectively. Let $Y=X+Z$. Demonstrate that $X|Y=y$ is binomial with parameters $m=y$ and $q=\lambda_1/(\lambda_1+\lambda_2)$ (see, for example,* [58], *p.* 131).

The conditional distribution of X given that $Y=y$ is

[2] When it is unclear, or when the random variable may be continuous, discrete, or a mixture of the two, the term **probability function** and abbreviation pf will be used. The term **probability density function** and the abbreviation pdf will be used only when the random variable is known to be continuous.

$$\begin{aligned}
f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)}\\
&= \frac{\Pr(X=x,Y=y)}{\Pr(Y=y)}\\
&= \frac{\Pr(X=x,Z=y-x)}{\Pr(Y=y)}\\
&= \frac{\Pr(X=x)\Pr(Z=y-x)}{\Pr(Y=y)}\\
&= \frac{\dfrac{\lambda_1^x e^{-\lambda_1}}{x!}\dfrac{\lambda_2^{y-x}e^{-\lambda_2}}{(y-x)!}}{\dfrac{(\lambda_1+\lambda_2)^y e^{-\lambda_1-\lambda_2}}{y!}}\\
&= \frac{y!}{x!(y-x)!}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{y-x}
\end{aligned}$$

for $x=0,1,2,\cdots,y$. This is a binomial distribution with parameters $m=y$ and $q=\lambda_1/(\lambda_1+\lambda_2)$. □

Note that (16.1) may be rewritten as

$$f_{X,Y}(x,y)=f_{X|Y}(x|y)f_Y(y), \tag{16.2}$$

demonstrating that joint distributions may be constructed from products of conditional and marginal distributions. Because the marginal distribution of X may be obtained by integrating (or summing) y out of the joint distribution,

$$f_X(x)=\int f_{X,Y}(x,y)\,dy,$$

we find using (16.2) that

$$f_X(x)=\int f_{X|Y}(x|y)f_Y(y)\,dy. \tag{16.3}$$

Formula (16.3) has an interesting interpretation as a mixed distribution (see Section 4.4.5). To see this, assume that the conditional distribution $f_{X|Y}(x|y)$ is one of the usual parametric distributions where y is the realization of a random parameter Y with distribution $f_Y(y)$. In Section 4.6.3 it was shown that if, given $\Theta=\theta$, X has a Poisson distribution with mean θ and Θ has a gamma distribution, then the marginal distribution of X will be negative binomial. Also, Example 4.30 showed that, if $X|\Theta$ has a normal distribution with mean Θ and variance v and Θ has a normal distribution with mean μ

and variance a, then the marginal distribution of X is normal with mean μ and variance $a + v$.

Note that the roles of X and Y in (16.2) can be interchanged, yielding

$$f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x),$$

because both sides of this equation equal the joint distribution of X and Y. Division by $f_Y(y)$ yields Bayes' theorem, namely,

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}.$$

16.2.2 Conditional expectation

As in the previous subsection, assume that X and Y are two random variables and the conditional pf of X given that $Y = y$ is $f_{X|Y}(x|y)$. Clearly, this is a valid probability distribution, and its mean is denoted by

$$\mathrm{E}(X|Y = y) = \int x\, f_{X|Y}(x|y)\, dx \tag{16.4}$$

with the integral replaced by a sum in the discrete case. Clearly, (16.4) is a function of y, and it is often of interest to view this conditional expectation as a random variable obtained by replacing y by Y in the right-hand side of (16.4). Thus we can write $\mathrm{E}(X|Y)$ instead of the left-hand side of (16.4), and so $\mathrm{E}(X|Y)$ is itself a random variable because it is a function of the random variable Y. The expectation of $\mathrm{E}(X|Y)$ is given by

$$\mathrm{E}[\mathrm{E}(X|Y)] = \mathrm{E}(X). \tag{16.5}$$

To see this, note that from (16.3) and (16.4)

$$\begin{aligned}
\mathrm{E}[\mathrm{E}(X|Y)] &= \int \mathrm{E}(\mathrm{E}|Y = y) f_Y(y)\, dy \\
&= \int\int x f_{X|Y}(x|y) dx f_Y(y)\, dy \\
&= \int x \int f_{X|Y}(x|y) f_Y(y)\, dy\, dx \\
&= \int x f_X(x)\, dx \\
&= \mathrm{E}(X)
\end{aligned}$$

with a similar proof in the discrete case.

Example 16.2 *Derive the mean of the negative binomial distribution by conditional expectation, recalling that, if* $X|\Theta \sim Poisson(\Theta)$ *and* $\Theta \sim gamma(\alpha, \beta)$, *then* $X \sim$ *negative binomial with* $r = \alpha$ *and* $\beta = \beta$.

We have

$$\mathrm{E}(X|\Theta) = \Theta$$

and so

$$\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X|\Theta)] = \mathrm{E}(\Theta).$$

From Appendix A the mean of the gamma distribution of Θ is $\alpha\beta$, and so $\mathrm{E}(X) = \alpha\beta$. □

It is often convenient to replace X by an arbitrary function $h(X, Y)$ in (16.4), yielding the more general definition

$$\mathrm{E}[h(X,Y)|Y = y] = \int h(x,y) f_{X|Y}(x|y)\, dx.$$

Similarly, $\mathrm{E}[h(X,Y)|Y]$ is the conditional expectation viewed as a random variable which is a function of Y. Then, (16.5) generalizes to

$$\mathrm{E}\{\mathrm{E}[h(X,Y)|Y]\} = \mathrm{E}[h(X,Y)]. \tag{16.6}$$

To see (16.6), note that

$$\begin{aligned}
\mathrm{E}\{\mathrm{E}[h(X,Y)|Y]\} &= \int \mathrm{E}[h(X,Y)|Y = y] f_Y(y)\, dy \\
&= \int\int h(x,y) f_{X|Y}(x|y)\, dx f_Y(y)\, dy \\
&= \int\int h(x,y)[f_{X|Y}(x|y) f_Y(y)]\, dx\, dy \\
&= \int\int h(x,y) f_{X,Y}(x,y)\, dx\, dy \\
&= \mathrm{E}[h(X,Y)]
\end{aligned}$$

from (16.2).

If we choose $h(X,Y) = [X - \mathrm{E}(X|Y)]^2$, then its expected value, based on the conditional distribution of X given Y, is the variance of this conditional distribution,

$$\mathrm{Var}(X|Y) = \mathrm{E}\{[X - \mathrm{E}(X|Y)]^2|Y\}. \tag{16.7}$$

Clearly, (16.7) is still a function of the random variable Y.

It is instructive now to analyze the variance of X where X and Y are two random variables. To begin, note that (16.7) may be written as

$$\mathrm{Var}(X|Y) = \mathrm{E}(X^2|Y) - [\mathrm{E}(X|Y)]^2.$$

Thus,

$$\begin{aligned}
\mathrm{E}[\mathrm{Var}(X|Y)] &= \mathrm{E}\{\mathrm{E}(X^2|Y) - [\mathrm{E}(X|Y)]^2\} \\
&= \mathrm{E}[\mathrm{E}(X^2|Y)] - \mathrm{E}\{[E(X|Y)]^2\} \\
&= \mathrm{E}(X^2) - \mathrm{E}\{[\mathrm{E}(X|Y)]^2\}.
\end{aligned}$$

Also, because $\text{Var}[h(Y)] = \text{E}\{[h(Y)]^2\} - \{\text{E}[h(Y)]\}^2$, we may use $h(Y) = \text{E}(X|Y)$ to obtain

$$\begin{aligned}\text{Var}[\text{E}(X|Y)] &= \text{E}\{[\text{E}(X|Y)]^2\} - \{\text{E}[\text{E}(X|Y)]\}^2 \\ &= \text{E}\{[\text{E}(X|Y)]^2\} - [\text{E}(X)]^2.\end{aligned}$$

Thus,

$$\begin{aligned}\text{E}[\text{Var}(X|Y)] + \text{Var}[\text{E}(X|Y)] &= \text{E}(X^2) - \text{E}\{[\text{E}(X|Y)]^2\} \\ &\quad +\text{E}\{[\text{E}(X|Y)]^2\} - [\text{E}(X)]^2 \\ &= \text{E}(X^2) - [\text{E}(X)]^2 \\ &= \text{Var}(X).\end{aligned}$$

Thus, we have established the important formula

$$\text{Var}(X) = \text{E}[\text{Var}(X|Y)] + \text{Var}[\text{E}(X|Y)]. \tag{16.8}$$

Formula (16.8) states that the variance of X is composed of the sum of two parts: the mean of the conditional variance plus the variance of the conditional mean.

Example 16.3 *Derive the variance of the negative binomial distribution.*

The Poisson distribution has equal mean and variance, that is,

$$\text{E}(X|\Theta) = \text{Var}(X|\Theta) = \Theta,$$

and so, from (16.8),

$$\begin{aligned}\text{Var}(X) &= \text{E}[Var(X|\Theta)] + \text{Var}[\text{E}(X|\Theta)] \\ &= \text{E}(\Theta) + \text{Var}(\Theta).\end{aligned}$$

Because Θ itself has a gamma distribution with parameters α and β, $\text{E}(\Theta) = \alpha\beta$ and $\text{Var}(\Theta) = \alpha\beta^2$. Thus the variance of the negative binomial distribution is

$$\begin{aligned}\text{Var}(X) &= \text{E}(\Theta) + \text{Var}(\Theta) \\ &= \alpha\beta + \alpha\beta^2 \\ &= \alpha\beta(1+\beta).\end{aligned}$$

□

Example 16.4 *It was shown in Example* 4.30 *that, if $X|\Theta$ is normally distributed with mean Θ and variance v where Θ is itself normally distributed with mean μ and variance a, then X (unconditionally) is normally distributed with mean μ and variance $a + v$. Use* (16.5) *and* (16.8) *to obtain the mean and variance of X directly.*

For the mean we have

$$\text{E}(X) = \text{E}[\text{E}(X|\Theta)] = \text{E}(\Theta) = \mu$$

and for the variance we obtain

$$\begin{aligned}\text{Var}(X) &= \text{E}[\text{Var}(X|\Theta)] + \text{Var}[\text{E}(X|\Theta)] \\ &= \text{E}(v) + \text{Var}(\Theta) \\ &= v + a\end{aligned}$$

because v is a constant. □

Example 16.5 *Consider a compound Poisson distribution with Poisson mean λ, where $X = Y_1 + \cdots + Y_N$ with $\text{E}(Y_i) = \mu_Y$ and $\text{Var}(Y_i) = \sigma_Y^2$. Determine the mean and variance of X.*

Formula (16.8) was used in Chapter 6 to obtain the answers:

$$\text{E}(X) = \lambda\mu_Y \text{ and } \text{Var}(X) = \lambda(\mu_Y^2 + \sigma_Y^2).$$

□

16.2.3 Nonparametric unbiased estimators

Unbiased estimation was covered in Section 9.2.2. It plays an important role in the development of credibility formulas. We begin by showing that two commonly used estimators are unbiased.

Theorem 16.6 *If $X_1, \ldots, X_n$ are independent but not necessarily identically distributed with common mean $\mu = \text{E}(X_j)$ and common variance $v = \text{Var}(X_j)$, then*

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

is an unbiased estimator of μ and

$$\hat{v} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2 \tag{16.9}$$

is an unbiased estimator of v.

Proof: For $\bar{X}$, we have

$$\text{E}(\bar{X}) = \text{E}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n}\sum_{j=1}^{n}\text{E}(X_j) = \mu.$$

For the variance estimator, begin with the following result, which will be used later:

$$\begin{aligned}
\sum_{j=1}^{n}(X_j-\bar{X})^2 &= \sum_{j=1}^{n}(X_j-\mu+\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n}(X_j-\mu)^2+2\sum_{j=1}^{n}(X_j-\mu)(\mu-\bar{X})+\sum_{j=1}^{n}(\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n}(X_j-\mu)^2+2(\mu-\bar{X})\sum_{j=1}^{n}(X_j-\mu)+n(\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n}(X_j-\mu)^2+2(\mu-\bar{X})n(\bar{X}-\mu)+n(\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n}(X_j-\mu)^2-n(\bar{X}-\mu)^2. \qquad (16.10)
\end{aligned}$$

We also have (from the independence of $X_1,\ldots,X_n$),

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

Take expectations in (16.10) to obtain

$$\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{n}(X_j-\bar{X})^2\right] &= \mathrm{E}\left[\sum_{j=1}^{n}(X_j-\mu)^2\right]-n\mathrm{E}[(\bar{X}-\mu)^2] \\
&= \sum_{j=1}^{n}\mathrm{E}[(X_j-\mu)^2]-n\ \mathrm{Var}(\bar{X}) \\
&= \sum_{j=1}^{n}\mathrm{Var}(X_j)-n\frac{v}{n} \\
&= \left(\sum_{j=1}^{n}v\right)-v \\
&= (n-1)v.
\end{aligned}$$

Dividing both sides by $n-1$ demonstrates that $\hat{v}$ is an unbiased estimator of v. □

The following example generalizing these results may appear somewhat artificial at this point, but is important in connection with the Bühlmann–Straub model of Section 16.4.5.

Example 16.7 *Suppose $X_1,\ldots,X_n$ are independent with common mean $\mu=\mathrm{E}(X_j)$ and variance $\mathrm{Var}(X_j)=\beta+\alpha/m_j$, $\alpha,\beta>0$ and all $m_j\geq 1$. Let $m=\sum_{j=1}^{n}m_j$ and consider the three estimators*

$$\bar{X}=\frac{1}{m}\sum_{j=1}^{n}m_jX_j, \quad \hat{\mu}_1=\frac{1}{n}\sum_{j=1}^{n}X_j,$$

and

$$\hat{\mu}_2=\frac{\sum_{j=1}^{n}\dfrac{m_jX_j}{m_j\beta+\alpha}}{\sum_{j=1}^{n}\dfrac{m_j}{m_j\beta+\alpha}}.$$

Show that all three estimators are unbiased for μ and then rank them in order by mean-squared error. Also obtain the expected value of a sum of squares that may be useful for estimating α and β.

First consider $\bar{X}$.

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= m^{-1}\sum_{j=1}^{n}m_j\mathrm{E}(X_j)=m^{-1}\sum_{j=1}^{n}m_j\mu=\mu, \\
\mathrm{Var}(\bar{X}) &= m^{-2}\sum_{j=1}^{n}m_j^2\,\mathrm{Var}(X_j) \\
&= m^{-2}\sum_{j=1}^{n}m_j^2\left(\beta+\frac{\alpha}{m_j}\right) \\
&= \alpha m^{-1}+\beta m^{-2}\sum_{j=1}^{n}m_j^2.
\end{aligned}$$

The estimator $\hat{\mu}_1$ is the one defined in Theorem 16.6 and has already been shown to be unbiased. We also have

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_1) &= n^{-2}\sum_{j=1}^{n}\mathrm{Var}(X_j)\\ &= n^{-2}\sum_{j=1}^{n}\left(\beta+\frac{\alpha}{m_j}\right)\\ &= \beta n^{-1}+n^{-2}\alpha\sum_{j=1}^{n}m_j^{-1}.\end{aligned}$$

With regard to $\hat{\mu}_2$,

$$\mathrm{E}(\hat{\mu}_2)=\frac{\sum_{j=1}^{n}\frac{m_j}{m_j\beta+\alpha}\mathrm{E}(X_j)}{\sum_{j=1}^{n}\frac{m_j}{m_j\beta+\alpha}}=\frac{\sum_{j=1}^{n}\frac{m_j}{m_j\beta+\alpha}\mu}{\sum_{j=1}^{n}\frac{m_j}{m_j\beta+\alpha}}=\mu$$

and

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_2) &= \frac{\sum_{j=1}^{n}\left(\frac{m_j}{m_j\beta+\alpha}\right)^2\left(\beta+\frac{\alpha}{m_j}\right)}{\left(\sum_{j=1}^{n}\frac{m_j}{m_j\beta+\alpha}\right)^2}\\ &= \frac{\sum_{j=1}^{n}\left(\frac{m_j}{m_j\beta+\alpha}\right)}{\left(\sum_{j=1}^{n}\frac{m_j}{m_j\beta+\alpha}\right)^2}\\ &= \left(\sum_{j=1}^{n}\frac{m_j}{m_j\beta+\alpha}\right)^{-1}.\end{aligned}$$

We now consider the relative ranking of these variances (because all three estimators are unbiased, their mean-squared errors equal their variances, so it is sufficient to rank the variances). To show that it is not possible to order $\mathrm{Var}(\hat{\mu}_1)$ and $\mathrm{Var}(\bar{X})$, examine their difference:

$$\mathrm{Var}(\bar{X})-\mathrm{Var}(\hat{\mu}_1)=\alpha\left(m^{-1}-n^{-2}\sum_{j=1}^{n}m_j^{-1}\right)+\beta\left(m^{-2}\sum_{j=1}^{n}m_j^2-n^{-1}\right).$$

The coefficient of β must be nonnegative. To see this, note that

$$\frac{1}{n}\sum_{j=1}^{n}m_j^2\ge\left(\frac{1}{n}\sum_{j=1}^{n}m_j\right)^2=\frac{m^2}{n^2}$$

(the left-hand side is like a sample second moment and the right-hand side is like the square of the sample mean) and then multiply both sides by nm^{-2}. To show that the coefficient of α must be nonpositive, note that

$$\frac{n}{\sum_{j=1}^{n}m_j^{-1}}\le\frac{1}{n}\sum_{j=1}^{n}m_j=\frac{m}{n}$$

(the harmonic mean is always less than or equal to the arithmetic mean) and then multiply both sides by n and then invert both sides. Therefore, by suitable choice of α and β, the difference in the variances can be made positive or negative.

We can do more than just show that $\hat{\mu}_2$ has the smallest variance of the three. Consider an arbitrary estimator of the form $\hat{\mu}=\sum_{j=1}^{n}a_jX_j$, where $\sum_{j=1}^{n}a_j=1$ (needed to ensure that $\hat{\mu}$ is unbiased). All three estimators are of this type. Incorporating the constraint by using Lagrange multipliers, the smallest variance is found by minimizing

$$\sum_{j=1}^{n}a_j^2\,\mathrm{Var}(X_j)+\lambda\left(\sum_{j=1}^{n}a_j-1\right).$$

The derivative with regard to a_i is

$$2a_i\,\mathrm{Var}(X_i)+\lambda,$$

and setting it equal to zero gives $a_i=-\lambda[2\,\mathrm{Var}(X_i)]^{-1}$. In other words, the weights should be proportional to the reciprocal of the variance. These are precisely the weights used in $\hat{\mu}_2$, and therefore it must have the smallest variance of all linear estimators.

With regard to a sum of squares, consider

$$\begin{aligned}
\sum_{j=1}^{n} m_j(X_j-\bar{X})^2 &= \sum_{j=1}^{n} m_j(X_j-\mu+\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n} m_j(X_j-\mu)^2 + 2\sum_{j=1}^{n} m_j(X_j-\mu)(\mu-\bar{X}) \\
&\quad + \sum_{j=1}^{n} m_j(\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n} m_j(X_j-\mu)^2 + 2(\mu-\bar{X})\sum_{j=1}^{n} m_j(X_j-\mu) \\
&\quad + m(\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n} m_j(X_j-\mu)^2 + 2(\mu-\bar{X})m(\bar{X}-\mu) + m(\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n} m_j(X_j-\mu)^2 - m(\bar{X}-\mu)^2. \qquad (16.11)
\end{aligned}$$

Taking expectations yields

$$\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{n} m_j(X_j-\bar{X})^2\right] &= \sum_{j=1}^{n} m_j \mathrm{E}[(X_j-\mu)^2] - m\mathrm{E}[(\bar{X}-\mu)^2] \\
&= \sum_{j=1}^{n} m_j \,\mathrm{Var}(X_j) - m\,\mathrm{Var}(\bar{X}) \\
&= \sum_{j=1}^{n} m_j\left(\beta+\frac{\alpha}{m_j}\right) - \beta\left(m^{-1}\sum_{j=1}^{n} m_j^2\right) - \alpha
\end{aligned}$$

and thus

$$\mathrm{E}\left[\sum_{j=1}^{n} m_j(X_j-\bar{X})^2\right] = \beta\left(m - m^{-1}\sum_{j=1}^{n} m_j^2\right) + \alpha(n-1). \qquad (16.12)$$

In addition to being of interest in its own right, (16.12) provides an unbiased estimator in situations more general than (16.9). The latter is recovered with the choice $\alpha = 0$ and $m_j = 1$ for $j = 1, 2, \cdots, n$, implying that $m = n$. Also, if $\beta = 0$, (16.12) allows us to derive an estimator of α when each X_j is the average of m_j independent observations each with mean μ and variance α. In any event, it is usually the case that the m_js (and hence m) are known. □

16.2.4 Exercises

16.1 Suppose X is binomially distributed with parameters n_1 and p, that is,

$$f_X(x) = \binom{n_1}{x} p^x (1-p)^{n_1-x}, \quad x = 0, 1, 2, \ldots, n_1.$$

Suppose also that Z is binomially distributed with parameters n_2 and p independently of X. Then $Y = X + Z$ is binomially distributed with parameters $n_1 + n_2$ and p. Find the conditional distribution of X given that $Y = y$.

16.2 Let X and Y have joint probability distribution as follows:

	y		
x	0	1	2
0	0.20	0	0.10
1	0	0.15	0.25
2	0.05	0.15	0.10

(a) Compute the marginal distributions of X and Y.

(b) Compute the conditional distribution of X given $Y = y$ for $y = 0, 1, 2$.

(c) Compute $\mathrm{E}(X|y)$,$\mathrm{E}(X^2|y)$, and $\mathrm{Var}(X|y)$ for $y = 0, 1, 2$.

(d) Compute $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ using (16.5), (16.8), and (c).

16.3 Suppose that X and Y are two random variables with bivariate normal joint density function

$$\begin{aligned}
f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\
&\quad \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)\right.\right. \\
&\qquad \left.\left. + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}.
\end{aligned}$$

Show that:

(a) The conditional density function is

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\left[\frac{x-\mu_1-\rho\frac{\sigma_1}{\sigma_2}(y-\mu_2)}{\sigma_1\sqrt{1-\rho^2}}\right]^2\right\}.$$

Hence,

$$\mathrm{E}(X|Y=y) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2).$$

(b) The marginal pdf is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right].$$

(c) The variables X and Y are independent if and only if $\rho = 0$.

16.4 Suppose that the random variables $Y_1, \cdots, Y_n$ are independent with

$$\mathrm{E}(Y_j) = \gamma \quad \text{and} \quad \mathrm{Var}(Y_j) = a_j + \sigma^2/b_j, \quad j = 1, 2, \ldots, n.$$

Define $b = b_1 + b_2 + \cdots + b_n$ and $\bar{Y} = \sum_{j=1}^n \frac{b_j}{b} Y_j$. Prove that

$$\mathrm{E}\left[\sum_{j=1}^n b_j(Y_j - \overline{Y})^2\right] = (n-1)\sigma^2 + \sum_{j=1}^n a_j\left(b_j - \frac{b_j^2}{b}\right).$$

16.5 Suppose that given $\boldsymbol{\Theta} = (\Theta_1, \Theta_2)$ the random variable X is normally distributed with mean Θ_1 and variance Θ_2.

(a) Show that $\mathrm{E}(X) = \mathrm{E}(\Theta_1)$ and $\mathrm{Var}(X) = \mathrm{E}(\Theta_2) + \mathrm{Var}(\Theta_1)$.

(b) If Θ_1 and Θ_2 are independent, show that X has the same distribution as $\Theta_1 + Y$, where Θ_1 and Y are independent and Y conditional on Θ_2 is normally distributed with mean 0 and variance Θ_2.

16.6 Suppose that Θ has pdf $\pi(\theta)$, $\theta > 0$, and Θ_1 has pdf $\pi_1(\theta) = \pi(\theta - \alpha)$, $\theta > \alpha > 0$. If, given Θ_1, X is Poisson distributed with mean Θ_1, show that X has the same distribution as $Y + Z$, where Y and Z are independent, Y is Poisson distributed with mean α, and $Z|\Theta$ is Poisson distributed with mean Θ.

16.3 LIMITED FLUCTUATION CREDIBILITY THEORY

This branch of credibility theory represents the first attempt to quantify the credibility problem. This approach was suggested in the early part of the century in connection with workers compensation insurance. The original paper on the subject was by Mowbray in 1914 [96]. The problem may be formulated as follows. Suppose that a policyholder has experienced X_j claims or losses[3] in past experience period j, where $j \in \{1, 2, 3, \ldots, n\}$. Another

[3] "Claims" will refer to the number of claims and "losses" will refer to payment amounts. In many cases, such as in this introductory paragraph, the ideas apply equally whether we are counting claims or losses.

view is that X_j is the experience from the jth policy in a group or from the jth member of a particular class in a rating scheme. Suppose that $\mathrm{E}(X_j) = \xi$, that is, the mean is stable over time or across the members of a group or class.[4] This quantity would be the premium to charge (net of expenses, profits, and a provision for adverse experience) if only we knew its value. Also suppose $\mathrm{Var}(X_j) = \sigma^2$, again, the same for all j. The past experience may be summarized by the average $\bar{X} = n^{-1}(X_1 + \cdots + X_n)$. We know that $\mathrm{E}(\bar{X}) = \xi$, and if the X_j are independent, $\mathrm{Var}(\bar{X}) = \sigma^2/n$. The insurer's goal is to decide on the value of ξ. One possibility is to ignore the past data (no credibility) and simply charge M a value obtained from experience on other similar but not identical policyholders. This quantity is often called the **manual premium** because it would come from a book (manual) of premiums. Another possibility is to ignore M and charge $\bar{X}$ (full credibility). A third possibility is to choose some combination of M and $\bar{X}$ (partial credibility).

From the insurer's standpoint, it seems sensible to "lean toward" the choice $\bar{X}$ if the experience is more "stable" (less variable, σ^2 small). This implies that $\bar{X}$ is of more use as a predictor of next year's results. Conversely, if the experience is more volatile (variable), then $\bar{X}$ is of less use as a predictor of next year's results and the choice M makes more sense.

Also, if we have an a priori reason to believe that the chances are great that this policyholder is unlike those who produced the manual premium M, then more weight should be given to $\bar{X}$. This is because as an unbiased estimator $\bar{X}$ tells us something useful about ξ while M is likely to be of little value. On the other hand, if all of our other policyholders have similar values of ξ there is no point in relying on the (perhaps limited) experience of any one of them when M is likely to provide an excellent description of the propensity for claims or losses.

While reference is made to policyholders, the entity contributing to each X_j could arise from a single policyholder, a class of policyholders possessing similar underwriting characteristics, or a group of insureds assembled for some other reason. For example, for a given year j, X_j could be the number of claims filed in respect of a single automobile policy in one year, the average number of claims filed by all policyholders in a certain ratings class (e.g., single, male, under age 25, living in an urban area, driving over 7,500 miles per year), or the average amount of losses per vehicle for a fleet of delivery trucks owned by a food wholesaler.

We first present one approach to decide whether to assign full credibility (charge $\bar{X}$), and then we present an approach to assign partial credibility if it is felt that full credibility is inappropriate.

[4] The customary symbol for the mean, μ, is not used here because that symbol is used for a different but related mean in the next section. We have chosen this particular symbol ("Xi") because it is the most difficult Greek letter to write and pronounce. It is an unwritten rule of textbook writing that it appear at least once.

16.3.1 Full credibility

One method of quantifying the stability of $\bar{X}$ is to infer that $\bar{X}$ is stable if the difference between $\bar{X}$ and ξ is small relative to ξ with high probability. In statistical terms, this means that we should select two numbers $r > 0$ and $0 < p < 1$ (with r close to 0 and p close to 1, common choices being $r = 0.05$ and $p = 0.9$) and assign full credibility if

$$\Pr(-r\xi \le \bar{X} - \xi \le r\xi) \ge p. \tag{16.13}$$

It is convenient to restate (16.13) as

$$\Pr\left(\left|\frac{\bar{X}-\xi}{\sigma/\sqrt{n}}\right| \le \frac{r\xi\sqrt{n}}{\sigma}\right) \ge p.$$

Now let y_p be defined by

$$y_p = \inf_y \left\{\Pr\left(\left|\frac{\bar{X}-\xi}{\sigma/\sqrt{n}}\right| \le y\right) \ge p\right\}. \tag{16.14}$$

That is, y_p is the smallest value of y which satisfies the probability statement in braces in (16.14). If $\bar{X}$ has a continuous distribution, the "$\ge$" sign in (16.14) may be replaced by an "=" sign and y_p satisfies

$$\Pr\left(\left|\frac{\bar{X}-\xi}{\sigma/\sqrt{n}}\right| \le y_p\right) = p. \tag{16.15}$$

Then the condition for full credibility is $r\xi\sqrt{n}/\sigma \ge y_p$,

$$\frac{\sigma}{\xi} \le \frac{r}{y_p}\sqrt{n} = \sqrt{\frac{n}{\lambda_0}}, \tag{16.16}$$

where $\lambda_0 = (y_p/r)^2$. Condition (16.16) states that full credibility is assigned if the coefficient of variation σ/ξ is no larger than $\sqrt{n/\lambda_0}$, an intuitively reasonable result.

Also of interest is that (16.16) can be rewritten to show that full credibility occurs when

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} \le \frac{\xi^2}{\lambda_0}. \tag{16.17}$$

Alternatively, solving (16.16) for n gives the number of exposure units required for full credibility, namely

$$n \ge \lambda_0 \left(\frac{\sigma}{\xi}\right)^2. \tag{16.18}$$

In many situations it is reasonable to approximate the distribution of $\bar{X}$ by a normal distribution with mean ξ and variance σ^2/n. For example, central limit theorem arguments may be applicable if n is large. In that case $(\bar{X} - \xi)/(\sigma/\sqrt{n})$ has a standard normal distribution. Then (16.15) becomes (where Z has a standard normal distribution and $\Phi(y)$ is its cdf)

$$\begin{aligned} p &= \Pr(|Z| \le y_p) \\ &= \Pr(-y_p \le Z \le y_p) \\ &= \Phi(y_p) - \Phi(-y_p) \\ &= \Phi(y_p) - 1 + \Phi(y_p) \\ &= 2\Phi(y_p) - 1. \end{aligned}$$

Therefore $\Phi(y_p) = (1+p)/2$ and therefore y_p is the $(1+p)/2$ percentile of the standard normal distribution.

For example, if $p = 0.9$, then standard normal tables give $y_{0.9} = 1.645$. If, in addition, $r = 0.05$, then $\lambda_0 = (32.9)^2 = 1{,}082.41$ and (16.18) yields $n \ge 1{,}082.41\sigma^2/\xi^2$. Note that this answer assumes we know the coefficient of variation of X_j. It is possible we have some idea of its value, even though we do not know the value of ξ (remember, that is the quantity we want to estimate).

The important thing to note when using (16.18) is that the coefficient of variation is for the estimator of the quantity to be estimated. The right-hand side gives the standard for full credibility when measuring it in terms of exposure units. If some other unit is desired, it is usually sufficient to multiply both sides by an appropriate quantity. Finally, any unknown quantities will have be to estimated from the data. This implies that the credibility question can be posed in a variety of ways. The following examples cover the most common cases.

Example 16.8 *Suppose past losses $X_1, \ldots, X_n$ are available for a particular policyholder. The sample mean is to be used to estimate $\xi = \mathrm{E}(X_j)$. Determine the standard for full credibility. Then suppose there were* 10 *observations with 6 being zero and the others being* 253, 398, 439, *and* 756. *Determine the full-credibility standard for this case with $r = 0.05$ and $p = 0.9$.*

The solution is available directly from (16.18) as

$$n \ge \lambda_0 \left(\frac{\sigma}{\xi}\right)^2.$$

For this specific case, the mean and standard deviation can be estimated from the data as 184.6 and 267.89 (where the variance estimate is the unbiased version using $n - 1$). With $\lambda_0 = 1082.41$, the standard is

$$n \ge 1082.41 \left(\frac{267.89}{184.6}\right)^2 = 2279.51$$

and the 10 observations do not deserve full credibility. □

In the next example it is further assumed that the observations are from a particular type of distribution.

Example 16.9 *Suppose that past losses $X_1, \ldots, X_n$ are available for a particular policyholder and it is reasonable to assume that the X_js are independent and compound Poisson distributed, that is, $X_j = Y_{j1} + \cdots + Y_{jN_j}$, where each N_j is Poisson with parameter λ and the claim size distribution Y has mean θ_Y and variance σ_Y^2. Determine the standard for full credibility when estimating the expected number of claims per policy and then when estimating the expected dollars of claims per policy. Then determine if these standards are met for the data in Example* 16.8*, where it is now known that the first three nonzero payments came from a single claim but the final one was from two claims, one for* 129 *and the other for* 627.

Case 1: Accuracy is to be measured with regard to the average number of claims. Then, using the N_js rather than the X_js, we have $\xi = \mathrm{E}(N_j) = \lambda$ and $\sigma^2 = \mathrm{Var}(N_j) = \lambda$, implying from (16.18) that

$$n \geq \lambda_0 \left(\frac{\lambda^{1/2}}{\lambda}\right)^2 = \frac{\lambda_0}{\lambda}.$$

Thus, if the standard is in terms of the number of policies, it will have to exceed λ_0/λ for full credibility and λ will have to be estimated from the data. If the standard is in terms of the number of expected claims, that is, $n\lambda$, we must multiply both sides by λ. This sets the standard as

$$n\lambda \geq \lambda_0.$$

While it appears that no estimation is needed for this standard, it is in terms of the expected number of claims needed. In practice, the standard is set in terms of the actual number of claims experienced, effectively replacing $n\lambda$ on the left by its estimate $N_1 + \cdots + N_n$.

For the given data, there were 5 claims, for an estimate of λ of 0.5 per policy. The standard is then

$$n \geq \frac{1{,}082.41}{0.5} = 2{,}164.82$$

and the 10 policies are far short of this standard. Or the 5 actual claims could be compared to $\lambda_0 = 1{,}082.41$, which leads to the same result.

Case 2: When accuracy is with regard to the average total payment, we have $\xi = \mathrm{E}(X_j) = \lambda\theta_Y$ and $\mathrm{Var}(Y_j) = \lambda(\theta_Y^2 + \sigma_Y^2)$, formulas developed in Chapter 6. In terms of the sample size, the standard is

$$n \geq \lambda_0 \frac{\lambda(\theta_Y^2 + \sigma_Y^2)}{\lambda^2\theta_Y^2} = \frac{\lambda_0}{\lambda}\left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right].$$

If the standard is in terms of the expected number of claims, multiply both sides by λ to obtain

$$n\lambda \geq \lambda_0\left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right].$$

Finally, if the standard is in terms of the expected total dollars of claims, multiply both sides by θ_Y to obtain

$$n\lambda\theta_Y \geq \lambda_0\left(\theta_Y + \frac{\sigma_Y^2}{\theta_Y}\right).$$

For the given data, the five claims have mean 369.2 and standard deviation 189.315 and thus

$$n \geq \frac{\lambda_0}{\lambda}\left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right] = \frac{1{,}082.41}{0.5}\left[1 + \left(\frac{189.315}{369.2}\right)^2\right] = 2{,}734.02$$

and again the 10 observations are far short of what is needed. If the standard is to be set in terms of claims (of which there are 5), multiply both sides by 0.5 to obtain a standard of 1,367.01. Finally, the standard could be set in terms of total dollars of claims. To do so, multiply both sides by 369.2 to obtain 504,701. Note that in all three cases the ratio of the observed quantity to the corresponding standard is unchanged:

$$\frac{10}{2{,}734.02} = \frac{5}{1{,}367.01} = \frac{1{,}846}{504{,}701} = 0.003658. \qquad \square$$

In these examples, the standard for full credibility is not met and so the sample means are not sufficiently accurate to be used as estimates of the expected value. We need a method for dealing with this situation.

16.3.2 Partial credibility

If it is decided that full credibility is inappropriate, then for competitive reasons (or otherwise) it may be desirable to reflect the past experience $\bar{X}$ in the net premium as well as the externally obtained mean, M. An intuitively appealing method for doing this is through a weighted average, that is, through the credibility premium

$$P_c = Z\bar{X} + (1 - Z)M, \tag{16.19}$$

where the credibility factor $Z \in [0, 1]$ needs to be chosen. There are many formulas for Z which have been suggested in the actuarial literature, usually justified on intuitive rather than theoretical grounds. (We remark that Mowbray [96] considered full, but not partial credibility.) One important choice is

$$Z = \frac{n}{n + k}, \tag{16.20}$$

where k needs to be determined. This particular choice will be shown to be theoretically justified on the basis of a statistical model to be presented in the next section. Another choice, based on the same idea as full credibility (and including the full-credibility case $Z = 1$), will now be discussed.

A variety of arguments have been used for developing the value of Z, many of which lead to the same answer. All of them are flawed in one way or another. The development we have chosen to present is also flawed but is at least simple. Recall that the goal of the full-credibility standard was to ensure that the difference between the net premium we are considering ($\bar{X}$) and what we should be using (ξ) is small with high probability. Because $\bar{X}$ is unbiased, this is essentially (and exactly if $\bar{X}$ has the normal distribution) equivalent to controlling the variance of the proposed net premium, $\bar{X}$, in this case. We see from (16.17) that there is no assurance that the variance of $\bar{X}$ will be small enough. However, it is possible to control the variance of the credibility premium, P_c, as follows:

$$\begin{aligned}\frac{\xi^2}{\lambda_0} &= \text{Var}(P_c)\\ &= \text{Var}[Z\bar{X} + (1-Z)M]\\ &= Z^2\,\text{Var}(\bar{X})\\ &= Z^2\frac{\sigma^2}{n}.\end{aligned}$$

Thus $Z = (\xi/\sigma)\sqrt{n/\lambda_0}$, provided it is less than 1. This can be written using the single formula

$$Z = \min\left\{\frac{\xi}{\sigma}\sqrt{\frac{n}{\lambda_0}}, 1\right\}. \tag{16.21}$$

One interpretation of (16.21) is that the credibility factor Z is the ratio of the coefficient of variation required for full credibility ($\sqrt{n/\lambda_0}$) to the actual coefficient of variation. For obvious reasons this is often called the square root rule for partial credibility.

While we could do the algebra with regard to (16.21), it is sufficient to note that it always turns out that Z is the square root of the ratio of the actual count to the count required for full credibility.

Example 16.10 *Suppose in Example* 16.8 *that the manual premium M is* 225. *Determine the credibility estimate.*

The average of the payments is 184.6. With the square root rule the credibility factor is

$$Z = \sqrt{\frac{10}{2{,}279.51}} = 0.06623.$$

Then the credibility premium is

$$P_c = 0.06623(184.6) + 0.93377(225) = 222.32.$$

□

Example 16.11 *Suppose in Example* 16.9 *that the manual premium M is* 225. *Determine the credibility estimate using both cases.*

For the first case, the credibility factor is

$$Z = \sqrt{\frac{5}{1{,}082.41}} = 0.06797$$

and applying it yields

$$P_c = 0.06797(184.6) + 0.93203(225) = 222.25.$$

At first glance this may appear inappropriate. The standard was set in terms of estimating the frequency but was applied to the aggregate claims. Often, individuals are distinguished more by differences in the frequency with which they have claims rather than by differences in the cost per claim. So this factor captures the most important feature.

For the second case, we can use any of the three calculations:

$$Z = \sqrt{\frac{10}{2{,}734.02}} = \sqrt{\frac{5}{1{,}367.01}} = \sqrt{\frac{1{,}846}{504{,}701}} = 0.06048.$$

Then,

$$P_c = 0.06048(184.6) + 0.93952(225) = 222.56.$$

□

Earlier we mentioned a flaw in the approach. Other than assuming that the variance captures the variability of $\bar{X}$ in the right way, all of the mathematics is correct. The flaw is in the goal. Unlike $\bar{X}$, P_c is not an unbiased estimator of ξ. In fact, one of the qualities that allows credibility to work is its use of biased estimators. But that means that the appropriate measure of the quality of P_c is not its variance, but its mean-squared error. However, the mean-squared error requires knowledge of the bias, and, in turn, that requires knowledge of the relationship of ξ and M. However, we know nothing about that relationship, and the data we have collected are of little help. As noted in the next subsection, this is not only a problem with our determination of Z, it is a problem that is characteristic of the limited fluctuation approach. A model for this relationship is introduced in the next section.

This section closes with a few additional examples. In each of the first two examples $\lambda_0 = 1{,}082.41$ is used.

Example 16.12 *For group dental insurance, historical experience on many groups has revealed that annual losses per life insured have a mean of* 175 *and a standard deviation of* 140. *A particular group has been covered for two years with* 100 *lives insured in year* 1 *and* 110 *in year* 2 *and has experienced average claims of* 150 *over that period. Determine if full or partial credibility is appropriate, and determine the credibility premium for next year's losses if there will be* 125 *lives insured.*

We will apply the credibility on a per-life-insured basis. We have observed 100+110 = 210 exposure units (assume experience is independent for different lives and years), and $\bar{X} = 150$. Now $M = 175$ and we assume that σ will be 140 for this group. Because we are trying to estimate the average cost per person, the calculations done in Example 16.11 for Case 2 apply. Thus, with $n = 210$ and $\lambda_0 = 1{,}082.41$ we estimate θ_Y with the sample mean of 150 to obtain the standard for full credibility as

$$n \geq 1{,}082.41 \left(\frac{140}{150}\right)^2 = 942.90$$

and then calculate

$$Z = \sqrt{\frac{210}{942.90}} = 0.472$$

(note that $\bar{X}$ is the average of 210 claims, so approximate normality is assumed by the central limit theorem). Thus, the net premium per life insured is

$$P_c = 0.472(150) + 0.528(175) = 163.2.$$

The net premium for the whole group is $125(163.2) = 20{,}400$. □

Example 16.13 *An insurance coverage involves credibility based on number of claims only. For a particular group,* 715 *claims have been observed. Determine an appropriate credibility factor, assuming that the number of claims is Poisson distributed.*

This is Case 1 from Example 16.11 and the standard for full credibility with regard to the number of claims is $n\lambda \geq \lambda_0 = 1082.41$. Then

$$Z = \sqrt{\frac{715}{1{,}082.41}} = 0.813.$$

□

Example 16.14 *Past data on a particular group are* $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$, *where the* X_j *are independent and identically distributed compound Poisson random variables with exponentially distributed claim sizes. If the credibility factor based on claim numbers is* 0.8, *determine the appropriate credibility factor based on total claims.*

When based on Poisson claim numbers, from Example 16.9, $Z = 0.8$ implies that $\lambda n/\lambda_0 = (0.8)^2 = 0.64$, where λn is the observed number of claims. For exponentially distributed claim sizes $\sigma_Y^2 = \theta_Y^2$. From Case 2 of Example 16.9, the standard for full credibility in terms of the number of claims is

$$n\lambda \geq \lambda_0 \left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right] = 2\lambda_0.$$

Then

$$Z = \sqrt{\frac{\lambda n}{2\lambda_0}} = \sqrt{0.32} = 0.566.$$

□

16.3.3 Problems with the approach

While the limited fluctuation approach yields simple solutions to the problem, there are theoretical difficulties. First, there is no underlying theoretical model for the distribution of the X_js and thus no reason why a premium of the form (16.19) is appropriate and preferable to M. Why not just estimate ξ from a collection of homogeneous policyholders and charge all policyholders the same rate? While there is a practical reason for using (16.19), no model has been presented to suggest that this may be appropriate. Consequently, the choice of Z (and hence P_c) is completely arbitrary.

Second, even if (16.19) were appropriate for a particular model, there is no guidance for the selection of r and p.

Finally, the limited fluctuation approach does not examine the difference between ξ and M. When (16.19) is employed, we are essentially stating that the value of M is accurate as a representation of the expected value given no information about this particular policyholder. However, it is usually the case that M is also an estimate and therefore unreliable in itself. The correct credibility question should be "how much more reliable is $\bar{X}$ compared to M?" and not "how reliable is $\bar{X}$?"

In the remainder of this chapter, a systematic modeling approach is presented for the claims experience of a particular policyholder which suggests that the past experience of the policyholder is relevant for prospective rate making. Furthermore, the intuitively appealing formula (16.19) is a consequence of this approach, and Z is often obtained from relations of the form (16.20).

16.3.4 Notes and References

The limited fluctuation approach is discussed by Herzog [52] and Longley-Cook [87]. See also Norberg [100].

16.3.5 Exercises

16.7 An insurance company has decided to establish its full-credibility requirements for an individual state rate filing. The full-credibility standard is to be set so that the observed total amount of claims underlying the rate filing would be within 5% of the true value with probability 0.95. The claim frequency follows a Poisson distribution and the severity distribution has pdf

$$f(x) = \frac{100 - x}{5{,}000}, \quad 0 \leq x \leq 100.$$

Determine the expected number of claims necessary to obtain full credibility using the normal approximation.

16.8 For a particular policyholder, the past total claims experience is given by $X_1, \ldots, X_n$, where the X_js are independent and identically distributed

Table 16.1 Data for Exercise 16.9

Year	1	2	3
Claims	475	550	400

compound random variables with Poisson parameter λ and gamma claim size distribution with pdf

$$f_Y(y) = \frac{y^{\alpha-1}e^{-y/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \quad y > 0.$$

You also know the following:

1. The credibility factor based on numbers of claims is 0.9.
2. The expected claim size $\alpha\beta = 100$.
3. The credibility factor based on total claims is 0.8.

Determine α and β.

16.9 For a particular policyholder, the manual premium is 600 per year. The past claims experience is given in Table 16.1. Assess whether full or partial credibility is appropriate and determine the net premium for next year's claims assuming the normal approximation. Use $r = 0.05$ and $p = 0.9$.

16.10 Redo Example 16.10 assuming that X_j is a compound negative binomial distribution rather than compound Poisson.

16.11 (*) The total number of claims for a group of insureds is Poisson with mean λ. Determine the value of λ such that the observed number of claims will be within 3% of λ with a probability of 0.975 using the normal approximation.

16.12 (*) An insurance company is revising rates based on old data. The expected number of claims for full credibility is selected so that observed total claims will be within 5% of the true value 90% of the time. Individual claim amounts have pdf $f(x) = 1/200,000$, $0 < x < 200{,}000$, and the number of claims has the Poisson distribution. The recent experience consists of 1,082 claims. Determine the credibility, Z, to be assigned to the recent experience. Use the normal approximation.

16.13 (*) The average claim size for a group of insureds is 1,500 with a standard deviation of 7,500. Assume that claim counts have the Poisson distribution. Determine the expected number of claims so that the total loss will be within 6% of the expected total loss with probability 0.90.

16.14 (*) A group of insureds had 6,000 claims and a total loss of 15,600,000. The prior estimate of the total loss was 16,500,000. Determine the limited fluctuation credibility estimate of the total loss for the group. Use the standard for full credibility determined in Exercise 16.13.

16.15 (*) The full-credibility standard is set so that the total number of claims is within 5% of the true value with probability p. This standard is 800 claims. The standard is then altered so that the total cost of claims is to be within 10% of the true value with probability p. The claim frequency has a Poisson distribution and the claim severity distribution has pdf $f(x) = 0.0002(100 - x)$, $0 < x < 100$. Determine the expected number of claims necessary to obtain full credibility under the new standard.

16.16 (*) A standard for full credibility of 1,000 claims has been selected so that the actual pure premium will be within 10% of the expected pure premium 95% of the time. The number of claims has the Poisson distribution. Determine the coefficient of variation of the severity distribution.

16.17 (*) For a group of insureds you are given the following information:

1. The prior estimate of expected total losses is 20,000,000.
2. The observed total losses are 25,000,000.
3. The observed number of claims is 10,000.
4. The number of claims required for full credibility is 17,500.

Determine the credibility estimate of the group's expected total losses based upon all the above information. Use the credibility factor that is appropriate if the goal is to estimate the expected number of losses.

16.18 (*) A full-credibility standard is determined so that the total number of claims is within 5% of the expected number with probability 98%. If the same expected number of claims for full credibility is applied to the total cost of claims, the actual total cost would be within $100K\%$ of the expected cost with 95% probability. Individual claims have severity pdf $f(x) = 2.5x^{-3.5}$, $x > 1$ and the number of claims has the Poisson distribution. Determine K.

16.19 (*) The number of claims has the Poisson distribution. The number of claims and the claim severity are independent. Individual claim amounts can be for 1, 2, or 10 with probabilities 0.5, 0.3, and 0.2, respectively. Determine the expected number of claims needed so that the total cost of claims is within 10% of the expected cost with 90% probability.

16.20 (*) The number of claims has the Poisson distribution. The coefficient of variation of the severity distribution is 2. The standard for full credibility in estimating total claims is 3,415. With this standard the observed pure

premium will be within $k\%$ of the expected pure premium 95% of the time. Determine k.

16.21 (*) You are given the following:

1. P = Prior estimate of pure premium for a particular class of business.
2. O = Observed pure premium during the latest experience period for the same class of business.
3. R = Revised estimate of pure premium for the same class following the observations.
4. F = Number of claims required for full credibility of the pure premium.

Express the observed number of claims as a function of these four items.

16.4 GREATEST ACCURACY CREDIBILITY THEORY

16.4.1 Introduction

In this and the following section, we consider a model-based approach to the solution of the credibility problem. This approach, referred to as greatest accuracy credibility theory, is the outgrowth of a classic 1967 paper by Bühlmann [18]. Many of the ideas are also found in Whitney [136] and Bailey [8].

We return to the basic problem. For a particular policyholder, we have observed n exposure units of past claims $\mathbf{X} = (X_1, \ldots, X_n)^T$. We have a manual rate μ (we no longer use M for the manual rate) which is applicable to this policyholder, but the past experience indicates that this may not be appropriate [$\bar{X} = n^{-1}(X_1 + \cdots + X_n)$, as well as $\mathrm{E}(X)$, could be quite different from μ]. This raises the question of whether next year's net premium (per exposure unit) should be based on μ, on $\bar{X}$, or on a combination of the two.

The insurer needs to consider the following question: Is the policyholder really different from what has been assumed in the calculation of μ or has it just been random chance which has been responsible for the differences between μ and $\bar{X}$?

While it is difficult to definitively answer the above question, it is clear that no underwriting system is perfect. The manual rate μ has presumably been obtained by (a) evaluation of the underwriting characteristics of the policyholder and (b) assignment of the rate on the basis of inclusion of the policyholder in a rating class. Such a class should include risks with similar underwriting characteristics. In other words, the rating class is viewed as homogeneous with respect to the underwriting characteristics used. Surely, not all risks in the class are truly homogeneous, however. No matter how

detailed the underwriting procedure, there still remains some heterogeneity with respect to risk characteristics within the rating class (good and bad risks, relatively speaking).

Thus, it is possible that the given policyholder may be different from what has been assumed. If this is the case, how should one choose an appropriate rate for the policyholder?

To proceed, let us assume that the risk level of each policyholder in the rating class may be characterized by a risk parameter θ (possibly vector valued), but the value of θ varies by policyholder. This allows us to quantify the differences between policyholders with respect to the risk characteristics. Because all observable underwriting characteristics have already been used, θ may be viewed as representative of the residual, unobserved factors which affect the risk level. Consequently, we shall assume the existence of θ, but we shall further assume that it is not observable and that we can never know its true value.

Because θ varies by policyholder, there is a probability distribution with pf $\pi(\theta)$ of these values across the rating class. Thus, if θ is a scalar parameter, the cumulative distribution function $\Pi(\theta)$ may be interpreted as the proportion of policyholders in the rating class with risk parameter Θ less than or equal to θ. [In statistical terms, Θ is a random variable with distribution function $\Pi(\theta) = \Pr(\Theta \le \theta)$.] Stated another way, $\Pi(\theta)$ represents the probability that a policyholder picked at random from the rating class has a risk parameter less than or equal to θ (to accommodate the possibility of new insureds, we slightly generalize the "rating class" interpretation to include the population of all potential risks, whether insured or not).

While the θ value associated with an individual policyholder is not (and cannot be) known, we assume (for this section) that $\pi(\theta)$ is known. That is, the structure of the risk characteristics within the population is known. This assumption can be relaxed, and we shall decide later how to estimate the relevant characteristics of $\pi(\theta)$ because this is needed in order to implement the theory.

Because risk levels vary within the population, it is clear that the experience of the policyholder varies in a systematic way with θ. Imagine that the experience of a policyholder picked (at random) from the population arises from a two-stage process. First, the risk parameter θ is selected from the distribution $\pi(\theta)$. Then the claims or losses X arise from the conditional distribution of X given θ, $f_{X|\Theta}(x|\theta)$. Thus the experience varies with θ via the distribution given the risk parameter θ. The distribution of claims thus differs from policyholder to policyholder to reflect the differences in the risk parameters.

Example 16.15 *Consider a rating class for automobile insurance, where θ represents the expected number of claims for a policyholder with risk parameter θ. To accommodate the variability in claims incidence, we assume that the values of θ vary across the rating class. Relatively speaking, the good*

Table 16.2 Probabilities for Example 16.16

x	$\Pr(X=x\|\Theta=G)$	$\Pr(X=x\|\Theta=B)$	θ	$\Pr(\Theta=\theta)$
0	0.7	0.5	G	0.75
1	0.2	0.3	B	0.25
2	0.1	0.2		

drivers are those with small values of θ, whereas the poor drivers are those with larger values of θ. It is convenient mathematically in this case to assume that the number of claims for a policyholder with risk parameter θ is Poisson distributed with mean θ. The random variable Θ may also be assumed to be gamma distributed with parameters α and β. Suppose it is known that the average number of expected claims for this rating class is 0.15 $[\mathrm{E}(\Theta)=0.15]$, *and* 95% *of the policyholders have expected claims between* 0.10 *and* 0.20. *Determine α and β.*

Assuming the normal approximation to the gamma, where it is known that 95% of the probability lies within about two standard deviations of the mean, it follows that Θ has standard deviation 0.025. Thus $\mathrm{E}(\Theta)=\alpha\beta=0.15$ and $\mathrm{Var}(\Theta)=\alpha\beta^2=(0.025)^2$. Solving for α and β yields $\beta=1/240$ and $\alpha=36$.□

Example 16.16 *There are two types of driver. Good drivers make up* 75% *of the population and in one year have zero claims with probability* 0.7, *one claim with probability* 0.2, *and two claims with probability* 0.1. *Bad drivers make up the other* 25% *of the population and have zero, one, or two claims with probabilities* 0.5, 0.3, *and* 0.2, *respectively. Describe this process and how it relates to an unknown risk parameter.*

When a driver buys our insurance, we do not know if the individual is a good or bad driver. So the risk parameter Θ can be one of two values. We can set $\Theta=G$ for good drivers and $\Theta=B$ for bad drivers. The probability model for the number of claims, X, and risk parameter Θ is given in Table 16.2. □

Example 16.17 *The amount of a claim has the exponential distribution with mean $1/\Theta$. Among the class of insureds and potential insureds, the parameter Θ varies according to the gamma distribution with $\alpha=4$ and scale parameter $\beta=0.001$. Provide a mathematical description of this model.*

For claims,

$$f_{X|\Theta}(x|\theta)=\theta e^{-\theta x},\ x,\theta>0,$$

and for the risk parameter,

$$\pi_\Theta(\theta)=\frac{\theta^3 e^{-1{,}000\theta}1{,}000^4}{6},\ \theta>0.$$

□

16.4.2 The Bayesian methodology

Continue to assume that the distribution of the risk characteristics in the population may be represented by $\pi(\theta)$, and the experience of a particular policyholder with risk parameter θ arises from the conditional distribution $f_{X|\Theta}(x|\theta)$ of claims or losses given θ.

We now return to the problem introduced in Section 16.3. That is, for a particular policyholder, we have observed $\mathbf{X}=\mathbf{x}$, where $\mathbf{X}=(X_1,\ldots,X_n)^T$ and $\mathbf{x}=(x_1,\ldots,x_n)^T$, and are interested in setting a rate to cover X_{n+1}. We assume that the risk parameter associated with the policyholder is θ (which is unknown). Furthermore, the experience of the policyholder corresponding to different exposure periods is assumed to be independent. In statistical terms, conditional on θ, the claims or losses $X_1,\ldots,X_n,X_{n+1}$ are independent (although not necessarily identically distributed).

Let X_j have conditional pf

$$f_{X_j|\Theta}(x_j|\theta),\quad j=1,\ldots,n,n+1.$$

Note that, if the X_j are identically distributed (conditional on $\Theta=\theta$), then $f_{X_j|\Theta}(x_j|\theta)$ does not depend on j. Ideally, we are interested in the conditional distribution of X_{n+1} given $\Theta=\theta$ in order to predict the claims experience X_{n+1} of the same policyholder (whose value of θ has been assumed not to have changed). If we knew θ, we could use $f_{X_{n+1}|\Theta}(x_{n+1}|\theta)$. Unfortunately, we do not know θ, but we do know $\mathbf{x}$ for the same policyholder. The obvious next step is to condition on $\mathbf{x}$ rather than θ. Consequently, we will calculate the conditional distribution of X_{n+1} given $\mathbf{X}=\mathbf{x}$, termed the **predictive distribution** as defined in Section 12.4.

The predictive distribution of X_{n+1} given $\mathbf{X}=\mathbf{x}$ is the relevant distribution for risk analysis, management, and decision making. It combines the uncertainty about the claims losses with that of the parameters associated with the risk process.

Here we repeat the development in Section 12.4, noting that if Θ has a discrete distribution the integrals are replaced by sums. Because the X_js are independent conditional on $\Theta=\theta$, we have

$$f_{\mathbf{X},\Theta}(\mathbf{x},\theta)=f(x_1,\ldots,x_n|\theta)\pi(\theta)=\left[\prod_{j=1}^{n}f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta).$$

The joint distribution of $\mathbf{X}$ is thus the marginal distribution obtained by integrating θ out, that is,

$$f_{\mathbf{X}}(\mathbf{x})=\int\left[\prod_{j=1}^{n}f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta)\,d\theta. \tag{16.22}$$

Similarly, the joint distribution of $X_1,\ldots,X_{n+1}$ is the right-hand side of (16.22) with n replaced by $n+1$ in the product. Finally, the conditional

density of X_{n+1} given $\mathbf{X} = \mathbf{x}$ is the joint density of $(X_1, \ldots, X_{n+1})$ divided by that of $\mathbf{X}$, namely,

$$f_{X_{n+1}|\mathbf{X}}(x_{n+1}|\mathbf{x}) = \frac{1}{f_{\mathbf{X}}(\mathbf{x})} \int \left[\prod_{j=1}^{n+1} f_{X_j|\Theta}(x_j|\theta) \right] \pi(\theta)\, d\theta. \tag{16.23}$$

There is a hidden mathematical structure underlying (16.23) which may often be exploited. The posterior density of Θ given $\mathbf{X}$ is

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X},\Theta}(\mathbf{x},\theta)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{1}{f_{\mathbf{X}}(\mathbf{x})} \left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta) \right] \pi(\theta). \tag{16.24}$$

In other words, $\left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta) = \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})$, and substitution in the numerator of (16.23) yields

$$f_{X_{n+1}|\mathbf{X}}(x_{n+1}|\mathbf{x}) = \int f_{X_{n+1}|\Theta}(x_{n+1}|\theta) \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta. \tag{16.25}$$

Equation (16.25) provides the additional insight that the conditional distribution of X_{n+1} given $\mathbf{X}$ may be viewed as a mixture distribution, with the mixing distribution the posterior distribution $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$.

The posterior distribution combines and summarizes the information about θ contained in the prior distribution and the likelihood and consequently (16.25) reflects this information. As noted in Theorem 12.49, the posterior distribution admits a convenient form when the likelihood is derived from the linear exponential family and $\pi(\theta)$ is the natural conjugate prior. This provides an easy method to evaluate the conditional distribution of X_{n+1} given $\mathbf{X}$ in these cases.

Example 16.18 (Example 16.16 continued) *For a particular policyholder suppose we have observed $x_1 = 0$ and $x_2 = 1$. Determine the predictive distribution of $X_3|X_1 = 0, X_2 = 1$ and the posterior distribution of $\Theta|X_1 = 0, X_2 = 1$.*

From (16.22), the marginal probability is

$$\begin{aligned} f_{\mathbf{X}}(0,1) &= \sum_{\theta} f_{X_1|\Theta}(0|\theta) f_{X_2|\Theta}(1|\theta) \pi(\theta) \\ &= 0.7(0.2)(0.75) + 0.5(0.3)(0.25) \\ &= 0.1425. \end{aligned}$$

Similarly, the joint probability of all three variables is

$$f_{\mathbf{X},X_3}(0,1,x_3) = \sum_{\theta} f_{X_1|\Theta}(0|\theta) f_{X_2|\Theta}(1|\theta) f_{X_3|\Theta}(x_3|\theta) \pi(\theta).$$

Thus,

$$\begin{aligned} f_{\mathbf{X},X_3}(0,1,0) &= 0.7(0.2)(0.7)(0.75) + 0.5(0.3)(0.5)(0.25) = 0.09225, \\ f_{\mathbf{X},X_3}(0,1,1) &= 0.7(0.2)(0.2)(0.75) + 0.5(0.3)(0.3)(0.25) = 0.03225, \\ f_{\mathbf{X},X_3}(0,1,2) &= 0.7(0.2)(0.1)(0.75) + 0.5(0.3)(0.2)(0.25) = 0.01800. \end{aligned}$$

The predictive distribution is then

$$\begin{aligned} f_{X_3|\mathbf{X}}(0|0,1) &= \frac{0.09225}{0.1425} = 0.647368, \\ f_{X_3|\mathbf{X}}(1|0,1) &= \frac{0.03225}{0.1425} = 0.226316, \\ f_{X_3|\mathbf{X}}(2|0,1) &= \frac{0.01800}{0.1425} = 0.126316. \end{aligned}$$

The posterior probabilities are, from (16.24),

$$\begin{aligned} \pi(G|0,1) &= \frac{f(0|G)f(1|G)\pi(G)}{f(0,1)} = \frac{0.7(0.2)(0.75)}{0.1425} = 0.736842, \\ \pi(B|0,1) &= \frac{f(0|B)f(1|B)\pi(B)}{f(0,1)} = \frac{0.5(0.3)(0.25)}{0.1425} = 0.263158. \end{aligned}$$

From this point forward the subscripts on f and π will be dropped unless needed for clarity. The predictive probabilities could also have been obtained using (16.25). This method is often simpler from a computation viewpoint.

$$\begin{aligned} f(0|0,1) &= \sum_{\theta} f(0|\theta)\pi(\theta|0,1) \\ &= 0.7(0.736842) + 0.5(0.263158) = 0.647368, \\ f(1|0,1) &= 0.2(0.736842) + 0.3(0.263158) = 0.226316, \\ f(2|0,1) &= 0.1(0.736842) + 0.2(0.263158) = 0.126316, \end{aligned}$$

which matches the previous calculations. □

Example 16.19 (Example 16.17 continued) *Suppose a person had claims of 100, 950, and 450. Determine the predictive distribution of the fourth claim and the posterior distribution of Θ.*

The marginal density at the observed values is

$$\begin{aligned} f(100,950,450) &= \int_0^{\infty} \theta e^{-100\theta} \theta e^{-950\theta} \theta e^{-450\theta} \frac{1{,}000^4}{6} \theta^3 e^{-1{,}000\theta} d\theta \\ &= \frac{1{,}000^4}{6} \int_0^{\infty} \theta^6 e^{-2{,}500\theta} d\theta = \frac{1{,}000^4}{6} \frac{720}{2{,}500^7}. \end{aligned}$$

Similarly,

$$\begin{aligned} f(100,950,450,x_4) &= \int_0^\infty \theta e^{-100\theta}\theta e^{-950\theta}\theta e^{-450\theta}\theta^{-\theta x_4}\frac{1{,}000^4}{6}\theta^3 e^{-1{,}000\theta}\,d\theta \\ &= \frac{1{,}000^4}{6}\int_0^\infty \theta^7 e^{-(2{,}500+x_4)\theta}\,d\theta \\ &= \frac{1{,}000^4}{6}\frac{5{,}040}{(2{,}500+x_4)^8}. \end{aligned}$$

Then the predictive density is

$$f(x_4|100,950,450) = \frac{\frac{1{,}000^4}{6}\frac{5{,}040}{(2{,}500+x_4)^8}}{\frac{1{,}000^4}{6}\frac{720}{2{,}500^7}} = \frac{7(2{,}500)^7}{(2{,}500+x_4)^8}$$

which is a Pareto density with parameters 7 and 2,500.

For the posterior distribution we take a shortcut. The denominator is an integral that produces a number and can be ignored for now. The numerator can be written

$$\pi(\theta|100,950,450) \propto \theta e^{-100\theta}\theta e^{-950\theta}\theta e^{-450\theta}\frac{1{,}000^4}{6}\theta^3 e^{-1{,}000\theta},$$

which was the term to be integrated in the calculation of the marginal density. Because there are constants in the denominator that have been ignored, we might as well ignore constants in the numerator. Only multiplicative terms involving the variable (θ in this case) need to be retained. Then

$$\pi(\theta|100,950,450) \propto \theta^6 e^{-2{,}500\theta}.$$

We could integrate this expression in order to determine the constant needed to make this a density function (that is, make the integral equal 1). But we recognize this function as that of a gamma distribution with parameters 7 and 1/2,500. Therefore,

$$\pi(\theta|100,950,450) = \frac{\theta^6 e^{-2{,}500\theta}2{,}500^7}{\Gamma(7)}.$$

Then the predictive density can be alternatively calculated from

$$\begin{aligned} f(x_4|100,950,450) &= \int_0^\infty \theta e^{-\theta x_4}\frac{\theta^6 e^{-2{,}500\theta}2{,}500^7}{\Gamma(7)}\,d\theta \\ &= \frac{2{,}500^7}{6!}\int_0^\infty \theta^7 e^{-(2{,}500+x_4)\theta}\,d\theta \\ &= \frac{2{,}500^7}{6!}\frac{7!}{(2{,}500+x_4)^8}, \end{aligned}$$

matching the answer previously obtained. □

Note that the posterior distribution is of the same type (gamma) as the prior distribution. The concept of a conjugate prior distribution was introduced in Section 12.4.3. This also implies that $X_{n+1}|\mathbf{x}$ is a mixture distribution with a simple mixing distribution, facilitating evaluation of the density of $X_{n+1}|\mathbf{x}$. Further examples of this idea are found in the exercises.

To return to the original problem, we have observed $\mathbf{X} = \mathbf{x}$ for a particular policyholder and we wish to predict X_{n+1} (or its mean). An obvious choice would be the hypothetical mean (or individual premium)

$$\mu_{n+1}(\theta) = \mathrm{E}(X_{n+1}|\Theta = \theta) = \int x_{n+1} f_{X_{n+1}|\Theta}(x_{n+1}|\theta)\,dx_{n+1} \tag{16.26}$$

if we knew θ. Note that replacement of θ by Θ in (16.26) yields, upon taking the expectation,

$$\mu_{n+1} = \mathrm{E}(X_{n+1}) = \mathrm{E}[\mathrm{E}(X_{n+1}|\Theta)] = \mathrm{E}[\mu_{n+1}(\Theta)]$$

so that the pure, or collective, premium is the mean of the hypothetical means. This is the premium we would use if we knew nothing about the individual. It does not depend on the individual's risk parameter, θ, nor does it use $\mathbf{x}$, the data collected from the individual. Because θ is unknown, the best we can do is try to use the data. This suggests the use of the Bayesian premium (the mean of the predictive distribution)

$$\mathrm{E}(X_{n+1}|\mathbf{X} = \mathbf{x}) = \int x_{n+1} f_{X_{n+1}|\mathbf{X}}(x_{n+1}|\mathbf{x})\,dx_{n+1}. \tag{16.27}$$

A computationally more convenient form is

$$\mathrm{E}\,(X_{n+1}|\mathbf{X} = \mathbf{x}) = \int \mu_{n+1}(\theta)\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\,d\theta. \tag{16.28}$$

In other words, the Bayesian premium is the expected value of the hypothetical means, with expectation taken over the posterior distribution $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$. We remind the reader that the integrals are replaced by sums in the discrete case. To prove (16.28), we see from (16.25) that

$$\begin{aligned} \mathrm{E}(X_{n+1}|\mathbf{X} = \mathbf{x}) &= \int x_{n+1} f_{X_{n+1}|\mathbf{X}}(x_{n+1}|\mathbf{x})\,dx_{n+1} \\ &= \int x_{n+1}\left[\int f_{X_{n+1}|\Theta}(x_{n+1}|\theta)\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})d\theta\right]dx_{n+1} \\ &= \int\left[\int x_{n+1} f_{X_{n+1}|\Theta}(x_{n+1}|\theta)dx_{n+1}\right]\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\,d\theta \\ &= \int \mu_{n+1}(\theta)\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\,d\theta. \end{aligned}$$

Example 16.20 (Example 16.18 continued) *Determine the Bayesian premium using both* (16.27) *and* (16.28).

The (unobservable) hypothetical means are

$$\begin{aligned}\mu_3(G) &= (0)(0.7)+1(0.2)+2(0.1)=0.4,\\ \mu_3(B) &= (0)(0.5)+1(0.3)+2(0.2)=0.7.\end{aligned}$$

If, as in Example 16.18, we have observed $x_1=0$ and $x_2=1$, we have the Bayesian premiums obtained directly from (16.27):

$$\mathrm{E}(X_3|0,1)=0(0.647368)+1(0.226316)+2(0.126316)=0.478948.$$

The (unconditional) pure premium is

$$\mu_3=\mathrm{E}(X_3)=\sum_\theta \mu_3(\theta)\pi(\theta)=(0.4)(0.75)+(0.7)(0.25)=0.475.$$

To verify (16.28) with $x_1=0$ and $x_2=1$, we have the posterior distribution $\pi(\theta|0,1)$ from Example 16.18.

Thus, (16.28) yields

$$\mathrm{E}(X_3|0,1)=0.4(0.736842)+0.7(0.263158)=0.478947$$

with the difference due to rounding. In general, the latter approach utilizing (16.28) is simpler than the direct approach using the conditional distribution of $X_{n+1}|\mathbf{X}=\mathbf{x}$. □

As expected, the revised value based on two observations is between the prior value (0.475) based on no data and the value based only on the data (0.5).

Example 16.21 (Example 16.19 continued) *Determine the Bayesian premium.*

From Example 16.19, we have $\mu_4(\theta)=\theta^{-1}$. Then, (16.28) yields

$$\begin{aligned}\mathrm{E}(X_4|100,950,450) &= \int_0^\infty \theta^{-1}\frac{\theta^6 e^{-2{,}500\theta}2{,}500^7}{720}d\theta\\ &= \frac{2{,}500^7}{720}\frac{120}{2{,}500^6}=416.67.\end{aligned}$$

This could also have been obtained from the formula for the moments of the gamma distribution in Appendix A. From the prior distribution,

$$\mu=\mathrm{E}(\Theta^{-1})=\frac{1{,}000}{3}=333.33$$

and once again the Bayesian estimate is between the prior estimate and one based solely on the data (the sample mean of 500).

From (16.27),

$$\mathrm{E}(X_4|100,950,450)=\frac{2{,}500}{6}=416.67,$$

the mean of the predictive Pareto distribution. □

Example 16.22 *Generalize the result of Example* 16.21 *for an arbitrary sample size of n and an arbitrary prior gamma distribution with parameters α and β, where β is the reciprocal of the usual scale parameter.*

The posterior distribution can be determined from

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto \left(\prod_{j=1}^n \theta e^{-\theta x_j}\right)\frac{\theta^{\alpha-1}e^{-\beta\theta}\beta^\alpha}{\Gamma(\alpha)}\\ &\propto \theta^{n+\alpha-1}e^{-(\Sigma x_j+\beta)\theta}.\end{aligned}$$

The second line follows because the posterior density is a function of θ and thus all multiplicative terms not involving θ may be dropped. Rather than perform the integral to determine the constant, we recognize that the posterior distribution is gamma with first parameter $n+\alpha$ and scale parameter $(\Sigma x_j+\beta)^{-1}$. The Bayes estimate of X_{n+1} is the expected value of Θ^{-1} using the posterior distribution. It is

$$\frac{\Sigma x_j+\beta}{n+\alpha-1}=\frac{n}{n+\alpha-1}\bar{x}+\frac{\alpha-1}{n+\alpha-1}\frac{\beta}{\alpha-1}.$$

Note that the estimate is a weighted average of the observed values and the unconditional mean. This formula is of the credibility weighted type (16.19). □

Here is an example where the random variables do not have identical distributions.

Example 16.23 *Suppose that the number of claims N_j in year j for a group policyholder with (unknown) risk parameter θ and m_j individuals in the group is Poisson distributed with mean $m_j\theta$, that is, for $j=1,\ldots,n$,*

$$\Pr(N_j=x|\Theta=\theta)=\frac{(m_j\theta)^x e^{-m_j\theta}}{x!},\quad x=0,1,2,\ldots.$$

This would be the case if, per individual, the number of claims were independently Poisson distributed with mean θ. Determine the Bayesian expected number of claims for the m_{n+1} individuals to be insured in year $n+1$.

With these assumptions, the average number of claims per individual in year j is

$$X_j = \frac{N_j}{m_j}, \quad j = 1, \ldots, n.$$

Therefore,

$$f_{X_j|\Theta}(x_j|\theta) = \Pr[N_j = m_j x_j | \Theta = \theta].$$

Assume Θ is gamma distributed with parameters α and β,

$$\pi(\theta) = \frac{\theta^{\alpha-1} e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \quad \theta > 0;$$

then the posterior distribution $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ is proportional (as a function of θ) to

$$\left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\right] \pi(\theta),$$

which is itself proportional to

$$\left[\prod_{j=1}^{n} \theta^{m_j x_j} e^{-m_j \theta}\right] \theta^{\alpha-1} e^{-\theta/\beta} = \theta^{\alpha+\sum_{j=1}^{n} m_j x_j - 1} e^{-\theta\left(\beta^{-1} + \sum_{j=1}^{n} m_j\right)}.$$

This is proportional to a gamma density with parameters $\alpha_* = \alpha + \sum_{j=1}^{n} m_j x_j$ and $\beta_* = (1/\beta + \sum_{j=1}^{n} m_j)^{-1}$, and so $\Theta|\mathbf{X}$ is also gamma, but with α and β replaced by α_* and β_*, respectively.

Now,

$$\mathrm{E}(X_j|\Theta = \theta) = \mathrm{E}\left(\frac{1}{m_j} N_j | \Theta = \theta\right) = \frac{1}{m_j} \mathrm{E}(N_j|\Theta = \theta) = \theta.$$

Thus $\mu_{n+1}(\theta) = \mathrm{E}(X_{n+1}|\Theta = \theta) = \theta$ and $\mu_{n+1} = \mathrm{E}(X_{n+1}) = \mathrm{E}[\mu_{n+1}(\Theta)] = \alpha\beta$ because Θ is gamma distributed with parameters α and β. From (16.28) and because $\Theta|\mathbf{X}$ is also gamma distributed with parameters α_* and β_*,

$$\begin{aligned} \mathrm{E}(X_{n+1}|\mathbf{X} = \mathbf{x}) &= \int_0^{\infty} \mu_{n+1}(\theta) \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \\ &= \mathrm{E}[\mu_{n+1}(\Theta)|\mathbf{X} = \mathbf{x}] \\ &= \mathrm{E}(\Theta|\mathbf{X} = \mathbf{x}) \\ &= \alpha_* \beta_*. \end{aligned}$$

Define the total number of lives observed to be $m = \sum_{j=1}^{n} m_j$.

Then,

$$\mathrm{E}(X_{n+1}|\mathbf{X} = \mathbf{x}) = Z\bar{x} + (1 - Z)\mu_{n+1},$$

where $Z = m/(m + \beta^{-1})$ and $\bar{x} = m^{-1} \sum_{j=1}^{n} m_j x_j$, and $\mu = \alpha\beta$, again an expression of the form (16.19).

The total Bayesian expected number of claims for m_{n+1} individuals in the group for the next year would be $m_{n+1}\mathrm{E}(X_{n+1}|\mathbf{X} = \mathbf{x})$.

The analysis based on independent and identically distributed Poisson claim counts is obtained with $m_j = 1$. Then $X_j \equiv N_j$ for $j = 1, 2, \ldots, n$ are independent (given θ) Poisson random variables with mean θ. In this case

$$\mathrm{E}(X_{n+1}|\mathbf{X} = \mathbf{x}) = Z\bar{x} + (1 - Z)\mu,$$

where $Z = n/(n + \beta^{-1})$, $\bar{x} = n^{-1} \sum_{j=1}^{n} x_j$, and $\mu = \alpha\beta$. □

In each of Examples 16.22 and 16.23 the Bayesian estimate was a weighted average of the sample mean $\bar{x}$ and the pure premium μ_{n+1}. This is appealing from a credibility standpoint. Furthermore, the credibility factor Z in each case is an increasing function of the number of exposure units. The greater the amount of past data observed, the closer Z is to 1, consistent with our intuition.

16.4.3 The credibility premium

In the previous section a systematic approach was suggested for treatment of the past data of a particular policyholder. Ideally, rather than the pure premium $\mu_{n+1} = \mathrm{E}(X_{n+1})$, one would like to charge the individual premium (or hypothetical mean) $\mu_{n+1}(\theta)$, where θ is the (hypothetical) parameter associated with the policyholder. Because θ is unknown, this is impossible, but we could instead condition on $\mathbf{x}$, the past data from the policyholder. This leads to the Bayesian premium $\mathrm{E}(X_{n+1}|\mathbf{x})$.

The major challenge with this approach is that it may be difficult to evaluate the Bayesian premium. Of course, in simple examples such as in the previous subsection, the Bayesian premium is not difficult to evaluate numerically. But these examples can hardly be expected to capture the essential features of a realistic insurance scenario. More realistic models may well introduce analytic difficulties with respect to evaluation of $\mathrm{E}(X_{n+1}|\mathbf{x})$, whether one uses (16.27) or (16.28). Often, numerical integration may be required. There are exceptions such as Examples 16.22 and 16.23.

We now present an alternative suggested by Bühlmann [18] in 1967. Recall the basic problem: We wish to use the conditional distribution $f_{X_{n+1}|\Theta}(x_{n+1}|\theta)$ or the hypothetical mean $\mu_{n+1}(\theta)$ for estimation of next year's claims. Because we have observed $\mathbf{x}$, one suggestion is to approximate $\mu_{n+1}(\theta)$ by a linear function of the past data. [After all, the formula $Z\overline{X} + (1 - Z)\mu$ is of this form.] Thus, let us restrict ourselves to estimators of the form $\alpha_0 + \sum_{j=1}^{n} \alpha_j X_j$, where $\alpha_0, \alpha_1, \ldots, \alpha_n$ need to be chosen. To this end, we will choose the αs to

minimize squared error loss, that is,

$$Q = \mathrm{E}\left\{\left[\mu_{n+1}(\Theta) - \alpha_0 - \sum_{j=1}^{n}\alpha_j X_j\right]^2\right\} \tag{16.29}$$

and the expectation is over the joint distribution of $X_1, \ldots, X_n$ and Θ. That is, the squared error is averaged over all possible values of Θ and all possible observations. To minimize Q, we take derivatives. Thus,

$$\frac{\partial Q}{\partial \alpha_0} = \mathrm{E}\left\{2\left[\mu_{n+1}(\Theta) - \alpha_0 - \sum_{j=1}^{n}\alpha_j X_j\right](-1)\right\}.$$

We shall denote by $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ the values of $\alpha_0, \alpha_1, \ldots, \alpha_n$ which minimize (16.29). Then equating $\partial Q/\partial \alpha_0$ to 0 yields

$$\mathrm{E}[\mu_{n+1}(\Theta)] = \tilde{\alpha}_0 + \sum_{j=1}^{n}\tilde{\alpha}_j E(X_j).$$

But $\mathrm{E}(X_{n+1}) = \mathrm{E}[\mathrm{E}(X_{n+1}|\Theta)] = \mathrm{E}[\mu_{n+1}(\Theta)]$, and so $\partial Q/\partial \alpha_0 = 0$ implies that

$$\mathrm{E}(X_{n+1}) = \tilde{\alpha}_0 + \sum_{j=1}^{n}\tilde{\alpha}_j \mathrm{E}(X_j). \tag{16.30}$$

Equation (16.30) may be termed the **unbiasedness equation** because it requires that the estimate $\tilde{\alpha}_0 + \sum_{j=1}^{n}\tilde{\alpha}_j X_j$ be unbiased for $\mathrm{E}(X_{n+1})$. However, the credibility estimate may be biased as an estimator of $\mu_{n+1}(\theta) = \mathrm{E}(X_{n+1}|\theta)$, the quantity we are trying to estimate. This bias will average out over the members of Θ. By accepting this bias we are able to reduce the overall mean-squared error. For $i = 1, \ldots, n$, we have

$$\frac{\partial Q}{\partial \alpha_i} = \mathrm{E}\left\{2\left[\mu_{n+1}(\Theta) - \alpha_0 - \sum_{j=1}^{n}\alpha_j X_j\right](-X_i)\right\}$$

and setting this equal to 0 yields

$$\mathrm{E}[\mu_{n+1}(\Theta)X_i] = \tilde{\alpha}_0 \mathrm{E}(X_i) + \sum_{j=1}^{n}\tilde{\alpha}_j \mathrm{E}(X_i X_j).$$

The left-hand side of this equation may be reexpressed as

$$\begin{aligned}
\mathrm{E}[\mu_{n+1}(\Theta)X_i] &= \mathrm{E}\{\mathrm{E}[X_i\mu_{n+1}(\Theta)|\Theta]\} \\
&= \mathrm{E}\{\mu_{n+1}(\Theta)\mathrm{E}[X_i|\Theta]\} \\
&= \mathrm{E}[\mathrm{E}(X_{n+1}|\Theta)\mathrm{E}(X_i|\Theta)] \\
&= \mathrm{E}[\mathrm{E}(X_{n+1}X_i|\Theta)] \\
&= \mathrm{E}(X_i X_{n+1}),
\end{aligned}$$

where the second from the last step follows by independence of X_i and X_{n+1} conditional on Θ. Thus $\partial Q/\partial \alpha_i = 0$ implies

$$\mathrm{E}(X_i X_{n+1}) = \tilde{\alpha}_0 \mathrm{E}(X_i) + \sum_{j=1}^{n}\tilde{\alpha}_j \mathrm{E}(X_i X_j). \tag{16.31}$$

Next multiply (16.30) by $\mathrm{E}(X_i)$ and subtract from (16.31) to obtain

$$\mathrm{Cov}(X_i, X_{n+1}) = \sum_{j=1}^{n}\tilde{\alpha}_j \ \mathrm{Cov}(X_i, X_j), \quad i = 1, \ldots, n. \tag{16.32}$$

Equation (16.30) and the n equations (16.32) together are called the **normal equations**. These equations may be solved for $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ to yield the credibility premium

$$\tilde{\alpha}_0 + \sum_{j=1}^{n}\tilde{\alpha}_j X_j. \tag{16.33}$$

While it is straightforward to express the solution $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ to the normal equations in matrix notation (if the covariance matrix of the X_js is nonsingular), we shall be content with solutions for some special cases.

Note that exactly one of the terms on the right-hand side of (16.32) is a variance term, that is, $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$. The other $n-1$ terms are true covariance terms.

As an added bonus, the values $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ also minimize

$$Q_1 = \mathrm{E}\left\{\left[\mathrm{E}(X_{n+1}|\mathbf{X}) - \alpha_0 - \sum_{j=1}^{n}\alpha_j X_j\right]^2\right\} \tag{16.34}$$

and

$$Q_2 = \mathrm{E}\left[\left(X_{n+1} - \alpha_0 - \sum_{j=1}^{n}\alpha_j X_j\right)^2\right]. \tag{16.35}$$

To see this, differentiate (16.34) or (16.35) with respect to $\alpha_0, \alpha_1, \ldots, \alpha_n$ and observe that the solutions still satisfy the normal equations (16.30) and (16.32). Thus the credibility premium (16.33) is the best linear estimator of each of the hypothetical mean $\mathrm{E}(X_{n+1}|\Theta)$, the Bayesian premium $\mathrm{E}(X_{n+1}|\mathbf{X})$, and X_{n+1}.

Example 16.24 *If* $\mathrm{E}(X_j) = \mu$, $\mathrm{Var}(X_j) = \sigma^2$, *and, for* $i \neq j$, $\mathrm{Cov}(X_i, X_j) = \rho\sigma^2$, *where the correlation coefficient* ρ *satisfies* $-1 < \rho < 1$, *determine the credibility premium* $\tilde{\alpha}_0 + \sum_{j=1}^{n}\tilde{\alpha}_j X_j$.

The unbiasedness equation (16.30) yields

$$\mu = \tilde{\alpha}_0 + \mu\sum_{j=1}^{n}\tilde{\alpha}_j$$

or

$$\sum_{j=1}^{n} \tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu}.$$

The n equations (16.32) become, for $i = 1, \ldots, n$,

$$\rho = \sum_{\substack{j=1 \\ j \neq i}}^{n} \tilde{\alpha}_j \rho + \tilde{\alpha}_i$$

or, stated another way,

$$\rho = \sum_{j=1}^{n} \tilde{\alpha}_j \rho + \tilde{\alpha}_i (1 - \rho), \quad i = 1, \ldots, n.$$

Thus

$$\tilde{\alpha}_i = \frac{\rho \left(1 - \sum_{j=1}^{n} \tilde{\alpha}_j\right)}{1 - \rho} = \frac{\rho \tilde{\alpha}_0}{\mu(1 - \rho)}$$

using the unbiasedness equation. Summation over i from 1 to n yields

$$\sum_{i=1}^{n} \tilde{\alpha}_i = \sum_{j=1}^{n} \tilde{\alpha}_j = \frac{n\rho \tilde{\alpha}_0}{\mu(1 - \rho)},$$

which combined with the unbiasedness equation gives an equation for $\tilde{\alpha}_0$, namely

$$1 - \frac{\tilde{\alpha}_0}{\mu} = \frac{n\rho \tilde{\alpha}_0}{\mu(1 - \rho)}.$$

Solving for $\tilde{\alpha}_0$ yields

$$\tilde{\alpha}_0 = \frac{(1 - \rho)\mu}{1 - \rho + n\rho}.$$

Thus,

$$\tilde{\alpha}_j = \frac{\rho \tilde{\alpha}_0}{\mu(1 - \rho)} = \frac{\rho}{1 - \rho + n\rho}.$$

The credibility premium is then

$$\begin{aligned} \tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j &= \frac{(1 - \rho)\mu}{1 - \rho + n\rho} + \sum_{j=1}^{n} \frac{\rho X_j}{1 - \rho + n\rho} \\ &= (1 - Z)\mu + Z\bar{X}, \end{aligned}$$

where $Z = n\rho/(1 - \rho + n\rho)$ and $\bar{X} = n^{-1} \sum_{j=1}^{n} X_j$. Thus, if $0 < \rho < 1$, then $0 < Z < 1$ and the credibility premium is a weighted average of $\mu = \mathrm{E}(X_{n+1})$ and $\bar{X}$, that is, is of the form (16.19). □

We now turn to some models which specify the conditional means and variances of $X_j|\Theta$ and hence the means $\mathrm{E}(X_j)$, variances $\mathrm{Var}(X_j)$, and covariances $\mathrm{Cov}(X_i, X_j)$.

16.4.4 The Buhlmann model

This, the first and simplest credibility model, specifies that for each policyholder (conditional on Θ) past losses $X_1, \ldots, X_n$ have the same mean and variance and are independent and identically distributed conditional on Θ.

Thus, define

$$\mu(\theta) = \mathrm{E}(X_j|\Theta = \theta)$$

and

$$v(\theta) = \mathrm{Var}(X_j|\Theta = \theta).$$

As discussed previously, $\mu(\theta)$ is referred to as the **hypothetical mean** whereas $v(\theta)$ is called the **process variance**. Define

$$\mu = \mathrm{E}[\mu(\Theta)], \tag{16.36}$$

$$v = \mathrm{E}[v(\Theta)], \tag{16.37}$$

and

$$a = \mathrm{Var}[\mu(\Theta)]. \tag{16.38}$$

The quantity μ in (16.36) is the **expected value of the hypothetical means**, v in (16.37) is the **expected value of the process variance**, and a in (16.38) is the **variance of the hypothetical means**. Note that μ is the estimate to use if we have no information about θ [and thus no information about $\mu(\theta)$]. It will also be referred to as the **collective premium**.

The mean, variance, and covariance of the X_js may now be obtained. First,

$$\mathrm{E}(X_j) = \mathrm{E}[\mathrm{E}(X_j|\Theta)] = \mathrm{E}[\mu(\Theta)] = \mu. \tag{16.39}$$

Second,

$$\begin{aligned} \mathrm{Var}(X_j) &= \mathrm{E}[\mathrm{Var}(X_j|\Theta)] + \mathrm{Var}[\mathrm{E}(X_j|\Theta)] \\ &= \mathrm{E}[v(\Theta)] + \mathrm{Var}[\mu(\Theta)] \\ &= v + a. \end{aligned} \tag{16.40}$$

Finally, for $i \neq j$,

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j) \\ &= \mathrm{E}[\mathrm{E}(X_i X_j|\Theta)] - \mu^2 \\ &= \mathrm{E}[\mathrm{E}(X_i|\Theta)\mathrm{E}(X_j|\Theta)] - \{\mathrm{E}[\mu(\Theta)]\}^2 \\ &= \mathrm{E}\left\{[\mu(\Theta)]^2\right\} - \{\mathrm{E}[\mu(\Theta)]\}^2 \\ &= \mathrm{Var}[\mu(\Theta)] \\ &= a. \end{aligned} \tag{16.41}$$

This is exactly of the form of Example 16.24 with parameters $\mu, \sigma^2 = v + a$, and $\rho = a/(v + a)$. Thus the credibility premium is

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = Z\bar{X} + (1 - Z)\mu, \tag{16.42}$$

where

$$Z = \frac{n}{n+k} \tag{16.43}$$

and

$$k = \frac{v}{a} = \frac{\mathrm{E}[\mathrm{Var}(X_j|\Theta)]}{\mathrm{Var}[\mathrm{E}(X_j|\Theta)]}. \tag{16.44}$$

The credibility factor Z in (16.43) with k given by (16.44) is referred to as the **Bühlmann credibility factor**. Note that (16.42) is of the form (16.19), and (16.43) is exactly (16.20). Now, however, we know how to obtain k, namely from (16.44).

Formula (16.42) has many appealing features. First, the credibility premium (16.42) is a weighted average of the sample mean $\bar{X}$ and the collective premium μ, a formula which we find desirable. Furthermore, Z approaches 1 as n increases, giving more credit to $\bar{X}$ rather than μ as more past data accumulates, a feature which agrees with intuition. Also, if the population is fairly homogeneous with respect to the risk parameter Θ, then (relatively speaking) the hypothetical means $\mu(\Theta) = \mathrm{E}(X_j|\Theta)$ do not vary greatly with Θ (i.e., they are close in value) and hence have small variability. Thus a is small relative to v, that is, k is large and Z is closer to 0. But this agrees with intuition because for a homogeneous population the overall mean μ is of more value in helping to predict next year's claims for a particular policyholder. Conversely, for a heterogeneous population, the hypothetical means $\mathrm{E}(X_j|\Theta)$ are more variable, that is, a is large and k is small, and so Z is closer to 1. Again this makes sense because in a heterogeneous population the experience of other policyholders is of less value in predicting the future experience of a particular policyholder than is the past experience of that policyholder.

We now present some examples.

Example 16.25 (Example 16.20 continued) *Determine the Bühlmann estimate of* $\mathrm{E}(X_3|0,1)$.

From earlier work,

$$\begin{array}{ll} \mu(G) = \mathrm{E}(X_j|G) = 0.4, & \mu(B) = \mathrm{E}(X_j|B) = 0.7, \\ \pi(G) = 0.75, & \pi(B) = 0.25, \end{array}$$

and therefore,

$$\begin{aligned} \mu &= \sum_\theta \mu(\theta)\pi(\theta) = 0.4(0.75) + 0.7(0.25) = 0.475, \\ a &= \sum_\theta \mu(\theta)^2\pi(\theta) - \mu^2 = 0.16(0.75) + 0.49(0.25) - 0.475^2 = 0.016875. \end{aligned}$$

For the process variance,

$$\begin{aligned} v(G) &= \mathrm{Var}(X_j|G) = 0^2(0.7) + 1^2(0.2) + 2^2(0.1) - 0.4^2 = 0.44, \\ v(B) &= \mathrm{Var}(X_j|B) = 0^2(0.5) + 1^2(0.3) + 2^2(0.2) - 0.7^2 = 0.61, \\ v &= \sum_\theta v(\theta)\pi(\theta) = 0.44(0.75) + 0.61(0.25) = 0.4825. \end{aligned}$$

Then (16.44) gives

$$k = \frac{v}{a} = \frac{0.4825}{0.016875} = 28.5926$$

and (16.43) gives

$$Z = \frac{2}{2 + 28.5926} = 0.0654.$$

The expected next value is then $0.0654(0.5) + 0.9346(0.475) = 0.4766$. This is the best linear approximation to the Bayesian premium (given in Example 16.20). □

Example 16.26 *Suppose as in Example* 16.23 *(with* $m_j = 1$*) that* $X_j|\Theta$*,* $j = 1, \ldots, n$*, are independently and identically Poisson distributed with (given) mean* Θ *and* Θ *is gamma distributed with parameters* α *and* β*. Determine the Bühlmann premium.*

We have

$$\mu(\theta) = \mathrm{E}(X_j|\Theta = \theta) = \theta, \quad v(\theta) = \mathrm{Var}(X_j|\Theta = \theta) = \theta,$$

and so

$$\mu = \mathrm{E}[\mu(\Theta)] = \mathrm{E}(\Theta) = \alpha\beta, \quad v = \mathrm{E}[v(\Theta)] = \mathrm{E}(\Theta) = \alpha\beta,$$

and

$$a = \mathrm{Var}[\mu(\Theta)] = \mathrm{Var}(\Theta) = \alpha\beta^2.$$

Then

$$k = \frac{v}{a} = \frac{\alpha\beta}{\alpha\beta^2} = \frac{1}{\beta}, \quad Z = \frac{n}{n+k} = \frac{n}{n+1/\beta} = \frac{n\beta}{n\beta+1},$$

and the credibility premium is

$$Z\bar{X} + (1-Z)\,\mu = \frac{n\beta}{n\beta+1}\bar{X} + \frac{1}{n\beta+1}\alpha\beta.$$

But, as shown at the end of Example 16.23, this is also the Bayesian estimate $\mathrm{E}(X_{n+1}|\mathbf{X})$. Thus, the credibility premium equals the Bayesian estimate in this case. □

Example 16.27 *Determine the Bühlmann estimate for the setting in Example* 16.22.

For this model,

$$\begin{aligned}
\mu(\Theta) &= \Theta^{-1},\ \mu = \mathrm{E}(\Theta^{-1}) = \frac{\beta}{\alpha-1},\\
v(\Theta) &= \Theta^{-2},\ v = \mathrm{E}(\Theta^{-2}) = \frac{\beta^2}{(\alpha-1)(\alpha-2)},\\
a &= \mathrm{Var}(\Theta^{-1}) = \frac{\beta^2}{(\alpha-1)(\alpha-2)} - \left(\frac{\beta}{\alpha-1}\right)^2 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)},\\
k &= \frac{v}{a} = \alpha - 1,\\
Z &= \frac{n}{n+k} = \frac{n}{n+\alpha-1},\\
P_c &= \frac{n}{n+\alpha-1}\bar{X} + \frac{\alpha-1}{n+\alpha-1}\frac{\beta}{\alpha-1},
\end{aligned}$$

which again matches the Bayesian estimate.

An alternative analysis for this problem could have started with a single observation of $S = X_1 + \cdots + X_n$. From the assumptions of the problem, S has a mean of $n\Theta^{-1}$ and a variance of $n\Theta^{-2}$. While it is true that S has a gamma distribution, that information is not needed because the Bühlmann approximation requires only moments. Following the above calculations,

$$\begin{aligned}
\mu &= \frac{n\beta}{\alpha-1},\ v = \frac{n\beta^2}{(\alpha-1)(\alpha-2)},\ a = \frac{n^2\beta^2}{(\alpha-1)^2(\alpha-2)},\\
k &= \frac{\alpha-1}{n},\ Z = \frac{1}{1+k} = \frac{n}{n+\alpha-1}.
\end{aligned}$$

The key is to note that in calculating Z the sample size is now 1, reflecting the single observation of S. Because $S = n\bar{X}$, the Bühlmann estimate is

$$P_c = \frac{n}{n+\alpha-1}n\bar{X} + \frac{\alpha-1}{n+\alpha-1}\frac{n\beta}{\alpha-1},$$

which is n times the previous answer. That is because we are now estimating the next value of S rather than the next value of X. However, the credibility factor itself (that is, Z) is the same whether we are predicting X_{n+1} or the next value of S. □

16.4.5 The Buhlmann–Straub model

The Bühlmann model of the previous section is the simplest of the credibility models because it effectively requires that the past claims experience of a policyholder comprise independent and identically distributed components with respect to each past year. An important practical difficulty with this assumption is that it does not allow for variations in exposure or size.

For example, what if the first year's claims experience of a policyholder reflected only a portion of a year due to an unusual policyholder anniversary?

What if a benefit change occurred part way through a policy year? For group insurance, what if the size of the group changed over time?

To handle these variations, we consider the following generalization of the Bühlmann model. Assume that $X_1, \ldots, X_n$ are independent, conditional on Θ, with common mean (as before)

$$\mu(\theta) = \mathrm{E}(X_j|\Theta = \theta)$$

but with conditional variances

$$\mathrm{Var}(X_j|\Theta = \theta) = \frac{v(\theta)}{m_j},$$

where m_j is a known constant measuring exposure. Note that m_j need only be proportional to the size of the risk. This model would be appropriate if each X_j were the average of m_j independent (conditional on Θ) random variables each with mean $\mu(\theta)$ and variance $v(\theta)$. In the above situations, m_j could be the number of months the policy was in force in past year j, or the number of individuals in the group in past year j, or the amount of premium income for the policy in past year j.

As in the Bühlmann model, let

$$\mu = \mathrm{E}[\mu(\Theta)], \quad v = \mathrm{E}[v(\Theta)],$$

and

$$a = \mathrm{Var}[\mu(\Theta)].$$

Then, for the unconditional moments, from (16.39) $\mathrm{E}(X_j) = \mu$, and from (16.41) $\mathrm{Cov}(X_i, X_j) = a$, but

$$\begin{aligned}
\mathrm{Var}(X_j) &= \mathrm{E}[\mathrm{Var}(X_j|\Theta)] + \mathrm{Var}[\mathrm{E}(X_j|\Theta)]\\
&= \mathrm{E}\left[\frac{v(\Theta)}{m_j}\right] + \mathrm{Var}[\mu(\Theta)]\\
&= \frac{v}{m_j} + a.
\end{aligned}$$

To obtain the credibility premium (16.33), we will solve the normal equations (16.30) and (16.32) to obtain $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$. For notational convenience, define

$$m = m_1 + m_2 + \cdots + m_n$$

to be the total exposure. Then using (16.39) the unbiasedness equation (16.30) becomes

$$\mu = \tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j \mu$$

which implies

$$\sum_{j=1}^{n} \tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu}. \tag{16.45}$$

For $i = 1, \ldots, n$, (16.32) becomes

$$a = \sum_{\substack{j=1 \\ j \neq i}}^{n} \tilde{\alpha}_j a + \tilde{\alpha}_i \left(a + \frac{v}{m_i} \right) = \sum_{j=1}^{n} \tilde{\alpha}_j a + \frac{v \tilde{\alpha}_i}{m_i},$$

which may be rewritten as

$$\tilde{\alpha}_i = \frac{a}{v} m_i \left(1 - \sum_{j=1}^{n} \tilde{\alpha}_j \right) = \frac{a}{v} \frac{\tilde{\alpha}_0}{\mu} m_i, \quad i = 1, \ldots, n. \tag{16.46}$$

Then, using (16.45) and (16.46),

$$1 - \frac{\tilde{\alpha}_0}{\mu} = \sum_{j=1}^{n} \tilde{\alpha}_j = \sum_{i=1}^{n} \tilde{\alpha}_i = \frac{a}{v} \frac{\tilde{\alpha}_0}{\mu} \sum_{i=1}^{n} m_i = \frac{a \tilde{\alpha}_0 m}{\mu v},$$

and so

$$\tilde{\alpha}_0 = \frac{\mu}{1 + am/v} = \frac{v/a}{m + v/a} \mu.$$

But this means that

$$\tilde{\alpha}_j = \frac{a \tilde{\alpha}_0}{\mu v} \cdot m_j = \frac{m_j}{m + v/a}.$$

The credibility premium (16.33) becomes

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = Z \bar{X} + (1 - Z) \mu, \tag{16.47}$$

where with $k = v/a$ from (16.44)

$$Z = \frac{m}{m + k}$$

and

$$\bar{X} = \sum_{j=1}^{n} \frac{m_j}{m} X_j. \tag{16.48}$$

Clearly, the credibility premium (16.47) is still of the form (16.19). In this case, m is the total exposure associated with the policyholder, and the Bühlmann–Straub credibility factor Z depends on m. Furthermore, $\bar{X}$ is a weighted average of the X_j, with weights proportional to m_j. Following the group interpretation, X_j is the average loss of the m_j group members in year j and so $m_j X_j$ is the total loss of the group in year j. Then $\bar{X}$ is the overall average loss per group member over the n years. The credibility premium to be charged to the group in year $n+1$ would thus be $m_{n+1}[Z\bar{X} + (1 - Z)\mu]$ for m_{n+1} members in the next year.

Had we known that (16.48) would be the correct weighting of the X_j to receive the credibility weight Z, the rest would have been easy. For the single observation $\bar{X}$ the process variance is

$$\text{Var}(\bar{X}|\theta) = \sum_{j=1}^{n} \frac{m_j^2}{m^2} \frac{v(\theta)}{m_j} = \frac{v(\theta)}{m}$$

and so the expected process variance is v/m. The variance of the hypothetical means is still a and therefore $k = v/(am)$. There is only one observation of $\bar{X}$ and so the credibility factor is

$$Z = \frac{1}{1 + v/(am)} = \frac{m}{m + v/a} \tag{16.49}$$

as before. Equation (16.48) should not have been surprising because the weights are simply inversely proportional to the (conditional) variance of each X_j.

Example 16.28 *As in Example* 16.23, *assume that in year j there are N_j claims from m_j policies, $j = 1, \ldots, n$. An individual policy has the Poisson distribution with parameter Θ and the parameter itself has the gamma distribution with parameters α and β. Determine the Bühlmann–Straub estimate of the number of claims in year $n+1$ if there will be m_{n+1} policies.*

In order to meet the conditions of this model, let $X_j = N_j/m_j$. Because N_j has the Poisson distribution with mean $m_j\Theta$, $\text{E}(X_j|\Theta) = \Theta = \mu(\Theta)$ and $\text{Var}(X_j|\Theta) = \Theta/m_j = v(\Theta)/m_j$. Then,

$$\begin{aligned} \mu &= \text{E}(\Theta) = \alpha\beta, \quad a = \text{Var}(\Theta) = \alpha\beta^2, \quad v = \text{E}(\Theta) = \alpha\beta, \\ k &= \frac{1}{\beta}, \quad Z = \frac{m}{m + 1/\beta} = \frac{m\beta}{m\beta + 1}, \end{aligned}$$

and the estimate for one policyholder is

$$P_c = \frac{m\beta}{m\beta + 1} \bar{X} + \frac{1}{m\beta + 1} \alpha\beta,$$

where $\bar{X} = m^{-1} \sum_{j=1}^{n} m_j X_j$. For year $n+1$, the estimate is $m_{n+1} P_c$, matching the answer to Example 16.23. □

The assumptions underlying the Bühlmann–Straub model may be too restrictive to represent reality. In a 1967 paper, Hewitt [55] observed that large risks do not behave the same as an independent aggregation of small risks and, in fact, are more variable than would be indicated by independence. A model that reflects this observation is created in the following example.

Example 16.29 *Let the conditional mean be* $\mathrm{E}(X_j|\Theta) = \mu(\Theta)$ *and the conditional variance be* $\mathrm{Var}(X_j|\Theta) = w(\Theta) + v(\Theta)/m_j$. *Further assume that* $X_1, \ldots, X_n$ *are conditionally independent given* Θ. *Show that this model supports Hewitt's observation and determine the credibility premium.*

Consider independent risks i and j with exposures m_i and m_j and with a common value of Θ. When aggregated, the variance of the average loss is

$$\begin{aligned}\mathrm{Var}\left(\frac{m_iX_i + m_jX_j}{m_i+m_j}\middle|\Theta\right) &= \left(\frac{m_i}{m_i+m_j}\right)^2 \mathrm{Var}(X_i|\Theta) \\ &\quad + \left(\frac{m_j}{m_i+m_j}\right)^2 \mathrm{Var}(X_j|\Theta) \\ &= \frac{m_i^2+m_j^2}{(m_i+m_j)^2}w(\Theta) + \frac{1}{m_i+m_j}v(\Theta)\end{aligned}$$

while a single risk with exposure m_i+m_j has variance $w(\Theta)+v(\Theta)/(m_i+m_j)$, which is larger.

With regard to the credibility premium, we have

$$\begin{aligned}\mathrm{E}(X_j) &= \mathrm{E}[\mathrm{E}(X_j|\Theta)] = \mathrm{E}[\mu(\Theta)] = \mu \\ \mathrm{Var}(X_j) &= \mathrm{E}[\mathrm{Var}(X_j|\Theta)] + \mathrm{Var}[\mathrm{E}(X_j|\Theta)] \\ &= \mathrm{E}\left[w(\Theta) + \frac{v(\Theta)}{m_j}\right] + \mathrm{Var}[\mu(\Theta)] \\ &= w + \frac{v}{m_j} + a,\end{aligned}$$

and for $i \neq j$, $\mathrm{Cov}(X_i, X_j) = a$ as in (16.41). The unbiasedness equation is still

$$\mu = \tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j \mu$$

and so

$$\sum_{j=1}^{n} \tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu}.$$

Equation (16.32) becomes

$$\begin{aligned}a &= \sum_{j=1}^{n} \tilde{\alpha}_j a + \tilde{\alpha}_i\left(w + \frac{v}{m_i}\right) \\ &= a\left(1 - \frac{\tilde{\alpha}_0}{\mu}\right) + \tilde{\alpha}_i\left(w + \frac{v}{m_i}\right), \quad i = 1, \ldots, n.\end{aligned}$$

Therefore,

$$\tilde{\alpha}_i = \frac{a\tilde{\alpha}_0/\mu}{w + v/m_i}.$$

Summing both sides yields

$$\frac{a\tilde{\alpha}_0}{\mu}\sum_{j=1}^{n}\frac{m_j}{v+wm_j} = \sum_{j=1}^{n}\tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu}$$

and so

$$\tilde{\alpha}_0 = \frac{1}{(a/\mu)\sum_{j=1}^{n}\frac{m_j}{v+wm_j} + \frac{1}{\mu}} = \frac{\mu}{1+am^*}$$

where

$$m^* = \sum_{j=1}^{n}\frac{m_j}{v+wm_j}.$$

Then

$$\tilde{\alpha}_j = \frac{am_j}{v+wm_j}\frac{1}{1+am^*}.$$

The credibility premium is

$$\frac{\mu}{1+am^*} + \frac{a}{1+am^*}\sum_{j=1}^{n}\frac{m_jX_j}{v+wm_j}.$$

The sum can be made to define a weighted average of the observations by letting

$$\bar{X} = \frac{\sum_{j=1}^{n}\frac{m_j}{v+wm_j}X_j}{\sum_{j=1}^{n}\frac{m_j}{v+wm_j}} = \frac{1}{m^*}\sum_{j=1}^{n}\frac{m_j}{v+wm_j}X_j.$$

If we now set

$$Z = \frac{am^*}{1+am^*},$$

the credibility premium is

$$Z\bar{X} + (1-Z)\mu.$$

Observe what happens as the exposures m_j go to infinity. The credibility factor becomes

$$Z \to \frac{an/w}{1+an/w} < 1.$$

Contrast this to the Bühlmann–Straub model where the limit is 1. Thus, no matter how large the risk, there is a limit to its credibility. A further generalization of this result is provided in Exercise 16.26. □

Another generalization is provided by letting the variance of $\mu(\Theta)$ depend on the exposure. This may be reasonable if we believe that the extent to which

a given risk's propensity to produce claims that differ from the mean is related to its size. For example, larger risks may be underwritten more carefully. In this case, extreme variations from the mean are less likely because we ensure that the risk not only meets the underwriting requirements but also appears to be exactly what it claims to be.

Example 16.30 (Example 16.29 continued) *In addition to the specification presented in Example* 16.29, *let* $\mathrm{Var}[\mu(\Theta)] = a + b/m$, *where* $m = \sum_{j=1}^{n} m_j$ *is the total exposure for the group. Develop the credibility formula.*

We now have

$$\begin{aligned}
\mathrm{E}(X_j) &= \mathrm{E}[\mathrm{E}(X_j|\Theta)] = \mathrm{E}[\mu(\Theta)] = \mu \\
\mathrm{Var}(X_j) &= \mathrm{E}[\mathrm{Var}(X_j|\Theta)] + \mathrm{Var}[\mathrm{E}(X_j|\Theta)] \\
&= \mathrm{E}\left[w(\Theta) + \frac{v(\Theta)}{m_j}\right] + \mathrm{Var}[\mu(\Theta)] \\
&= w + \frac{v}{m_j} + a + \frac{b}{m}
\end{aligned}$$

and for $i \neq j$

$$\begin{aligned}
\mathrm{Cov}(X_i, X_j) &= \mathrm{E}[\mathrm{E}(X_i X_j|\Theta)] - \mu^2 \\
&= \mathrm{E}[\mu(\Theta)^2] - \mu^2 \\
&= a + \frac{b}{m}.
\end{aligned}$$

It can be seen that all the calculations used in Example 16.29 apply here with a replaced by $a + b/m$. The credibility factor is

$$Z = \frac{(a + b/m)m^*}{1 + (a + b/m)m^*}$$

and the credibility premium is

$$Z\bar{X} + (1 - Z)\mu$$

with $\bar{X}$ and m^* defined as in Example 16.29. This particular credibility formula has been used in workers compensation experience rating. One example of this is presented in detail in [45]. □

16.4.6 Exact credibility

In Examples 16.26–16.28 we found that the credibility premium and the Bayesian premium were equal. From (16.34), one may view the credibility premium as the best linear approximation to the Bayesian premium in the sense of squared error loss. In these examples the approximation is exact

because the two premiums are equal. The term **exact credibility** is used to describe the situation when the credibility premium equals the Bayesian premium.

In fact, it is not hard to see that one can ascertain whether credibility is exact without even calculating the credibility premium. If the Bayesian premium is a linear function of $X_1, \ldots, X_n$,

$$\mathrm{E}(X_{n+1}|\mathbf{X}) = a_0 + \sum_{j=1}^{n} a_j X_j,$$

then it is clear that in (16.34) the quantity Q_1 attains its minimum value of zero with $\tilde{\alpha}_j = a_j$ for $j = 0, 1, \ldots, n$. Thus the credibility premium is $\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = a_0 + \sum_{j=1}^{n} a_j X_j = \mathrm{E}(X_{n+1}|\mathbf{X})$ and credibility is exact.

This phenomenon occurs fairly generally in connection with linear exponential family members (Section 12.4.3) and their conjugate priors. We parameterize such that $X_j|\Theta = \theta$ is independently (conditional on $\Theta = \theta$) distributed with pf for $j = 1, \ldots, n+1$,

$$f_{X_j|\Theta}(x_j|\theta) = \frac{p(x_j)e^{-\theta x_j}}{q(\theta)},$$

and Θ has pdf

$$\pi(\theta) = \frac{[q(\theta)]^{-k}e^{-\mu k\theta}}{c(\mu, k)}, \quad \theta_0 < \theta < \theta_1, \tag{16.50}$$

where $-\infty \leq \theta_0 < \theta_1 \leq \infty$. It is also assumed that $\pi(\theta_0) = \pi(\theta_1) = 0$. For the moment, μ and k are simply parameters of $\pi(\theta)$. We will now demonstrate that the choice of symbols was no coincidence.

In Section 12.4.3 it was shown that

$$\mu(\theta) = \mathrm{E}(X_j|\Theta = \theta) = -\frac{q'(\theta)}{q(\theta)}.$$

We wish to find $\mathrm{E}[\mu(\Theta)]$. From (16.50),

$$\ln \pi(\theta) = -k \ln q(\theta) - \mu k\theta - \ln c(\mu, k)$$

and differentiating with respect to θ gives

$$\frac{\pi'(\theta)}{\pi(\theta)} = -\frac{kq'(\theta)}{q(\theta)} - \mu k.$$

In other words,

$$\pi'(\theta) = k[\mu(\theta) - \mu]\pi(\theta) \tag{16.51}$$

and integrating from θ_0 to θ_1 gives

$$\pi(\theta_1) - \pi(\theta_0) = k\int_{\theta_0}^{\theta_1} \mu(\theta)\pi(\theta)\,d\theta - k\mu\int_{\theta_0}^{\theta_1} \pi(\theta)\,d\theta.$$

This implies that $0 = k\mathrm{E}[\mu(\Theta)] - k\mu$, or equivalently,

$$\mathrm{E}[\mu(\Theta)] = \mu. \tag{16.52}$$

Now consider the posterior distribution $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$. It is proportional to

$$\left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta),$$

itself proportional to

$$\begin{aligned}&\left[\prod_{j=1}^{n}\frac{e^{-\theta x_j}}{q(\theta)}\right][q(\theta)]^{-k}e^{-\mu k\theta}\\ &\quad=[q(\theta)]^{-(n+k)}e^{-\theta(\mu k+n\bar{x})}\\ &\quad=[q(\theta)]^{-k_*}e^{-\mu_* k_*\theta},\end{aligned} \tag{16.53}$$

where

$$k_* = n + k$$

and

$$\mu_* = \frac{\mu k + n\bar{x}}{k+n} = \frac{n}{n+k}\bar{x} + \frac{k}{n+k}\mu.$$

Observe that (16.53) is proportional to a density of the form (16.50) with μ and k replaced by μ_* and k_*, respectively. Hence

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{[q(\theta)]^{-k_*}e^{-\mu_* k_*\theta}}{c(\mu_*, k_*)}, \quad \theta_0 < \theta < \theta_1.$$

From (16.28) and using the same development that led to (16.52), the Bayesian premium is

$$\begin{aligned}\mathrm{E}(X_{n+1}|\mathbf{x}) &= \int_{\theta_0}^{\theta_1}\mu(\theta)\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\,d\theta\\ &= \mu_*\\ &= Z\bar{x} + (1-Z)\mu,\end{aligned}$$

where $Z = n/(n+k)$. This is of the form (16.19), and because it is a linear function of the x_js, credibility must be exact, that is, the credibility premium is

$$\tilde{\alpha}_0 + \sum_{j=1}^{n}\tilde{\alpha}_j X_j = Z\bar{X} + (1-Z)\mu = \mathrm{E}(X_{n+1}|\mathbf{x}).$$

Because the $X_j|\Theta$ are also identically distributed for $j = 1, \ldots, n$, the Bühlmann model applies and so (16.42) also applies; that is, k must also satisfy (16.44). To see this directly, recall from Section 12.4.3 that

$$v(\theta) = \mathrm{Var}(X_j|\Theta = \theta) = -\mu'(\theta).$$

Differentiation of (16.51) yields

$$\begin{aligned}\pi''(\theta) &= k\mu'(\theta)\pi(\theta) + k^2[\mu(\theta)-\mu]^2\pi(\theta)\\ &= -kv(\theta)\pi(\theta) + k^2[\mu(\theta)-\mu]^2\pi(\theta).\end{aligned}$$

Integration with respect to θ from θ_0 to θ_1 yields

$$\begin{aligned}\pi'(\theta_1) - \pi'(\theta_0) &= -kE[v(\Theta)] + k^2\mathrm{E}\{[\mu(\Theta)-\mu]^2\}\\ &= -kv + k^2 a\end{aligned}$$

because $\mu(\Theta)$ has mean μ and $\mathrm{E}\{[\mu(\Theta)-\mu]^2\} = \mathrm{Var}[\mu(\Theta)] = a$. If $\pi'(\theta_1) = \pi'(\theta_0) = 0$, this implies that $k = v/a$, and so (16.44) is satisfied.

16.4.7 Linear versus Bayesian versus no credibility

In Section 16.4.3 it was demonstrated that the credibility premium is the best linear estimator in the sense of minimizing the expected squared error with respect to the next observation, X_{n+1}. In Exercise 16.59 you are asked to demonstrate that the Bayesian premium is the best estimator with no restrictions, in the same least squares sense. It was also demonstrated in Section 16.4.3 that the credibility premium is the linear estimator that is closest to the Bayesian estimator, again in the mean squared error sense. Finally, we have seen that in a number of cases the credibility and Bayesian premiums are the same. This leaves two questions. Is the additional error caused by using the credibility premium in place of the Bayesian premium worth worrying about? Is it worthwhile to go through the bother of using credibility in the first place? While the exact answer to these questions depends on the underlying distributions, we can obtain some feel for the answers by considering two examples.

We begin with the second question and use a common situation that has already been discussed. What makes credibility work is that we expect to perform numerous estimations. As a result, we are willing to be biased in any one estimation provided that the biases cancel out over the numerous estimations. This allows us to reduce variability and, therefore, squared error. The following example shows the power of credibility in this setting.

Example 16.31 *Suppose there are* 50 *occasions on which we obtain a random sample of size* 10 *from a Poisson distribution with unknown mean. The samples are from different Poisson populations and therefore may involve different means. Let the true means be* $\theta_1, \ldots, \theta_{50}$. *Further assume that the Poisson parameters are drawn from a gamma distribution with parameters* $\alpha = 50$ *and* $\beta = 0.1$. *Compare the maximum likelihood estimates* $\bar{X}_j, j = 1, \ldots, 50$, *to the credibility estimates* $C_j = (\bar{X}_j + 5)/2$. *Note that this is the Bühlmann credibility estimate.*

We will first analyze the two estimates by determining their respective mean-squared errors. Using the sample mean the total squared error is, where

$\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_{50})$,

$$S_1 = \sum_{j=1}^{50} (\bar{X}_j - \Theta_j)^2,$$

and the mean-squared error is

$$\mathrm{E}(S_1) = \mathrm{E}[\mathrm{E}(S_1|\boldsymbol{\Theta})] = \mathrm{E}\left[\sum_{j=1}^{50} \mathrm{Var}(\bar{X}_j|\Theta_j)\right] = \mathrm{E}\left(\sum_{j=1}^{50} \frac{\Theta_j}{10}\right) = 25.$$

Using the credibility estimator, the squared error is

$$S_2 = \sum_{j=1}^{50} (0.5\bar{X}_j + 2.5 - \Theta_j)^2$$

and the mean-squared error is

$$\begin{aligned}
\mathrm{E}(S_2) &= \mathrm{E}[\mathrm{E}(S_2|\boldsymbol{\Theta})] \\
&= \mathrm{E}\left[\sum_{j=1}^{50} E(0.25\bar{X}_j^2 + 6.25 + \Theta_j^2 + 2.5\bar{X}_j - 5\Theta_j - \bar{X}_j\Theta_j|\Theta_j)\right] \\
&= \mathrm{E}\left\{\sum_{j=1}^{50} \left[0.25\left(\frac{\Theta_j}{10} + \Theta_j^2\right) + 6.25 + \Theta_j^2 + 2.5\Theta_j - 5\Theta_j - \Theta_j^2\right]\right\} \\
&= \sum_{j=1}^{50} [0.25(0.5 + 25.5) + 6.25 + 25.5 + 2.5(5) - 5(5) - 25.5] \\
&= 12.5.
\end{aligned}$$

Of course, we "cheated" a bit. We used squared error as our criterion and so knew in advance that the Bühlmann estimate would have the smaller value given that it is competing against another linear estimator. The interesting part is the significant improvement that resulted. This means that, even if the components of the credibility formula Z and μ were not set at their optimal values, the credibility formula is still likely to result in an improvement.

To get a feel for how this improvement comes about, consider a specific set of 50 values of θ_j. The ones presented in Table 16.3 are a random sample from the prior gamma distribution sorted in increasing order. The next column provides the mean-squared error of the sample mean ($\theta_j/10$). The final three columns provide the bias, variance, and mean-squared error for the credibility estimator based on $Z = 0.5$ and $\mu = 5$. The sample mean is always unbiased and therefore the variance matches the mean-squared error and so these two

Table 16.3 A comparison of the sample mean and the credibility estimator

	$\bar{X}$	$0.5\bar{X} + 2.5$				$\bar{X}$	$0.5\bar{X} + 2.5$		
θ	MSE	Bias	Var.	MSE	θ	MSE	Bias	Var.	MSE
3.510	.351	.745	.088	.643	4.875	.488	.062	.122	.126
3.637	.364	.681	.091	.555	4.894	.489	.053	.122	.125
3.742	.374	.629	.094	.489	4.900	.490	.050	.123	.125
3.764	.376	.618	.094	.476	4.943	.494	.028	.124	.124
3.793	.379	.604	.095	.459	4.977	.498	.012	.124	.125
4.000	.400	.500	.100	.350	5.002	.500	−.001	.125	.125
4.151	.415	.424	.104	.284	5.013	.501	−.006	.125	.125
4.153	.415	.424	.104	.283	5.108	.511	−.054	.128	.131
4.291	.429	.354	.107	.233	5.172	.517	−.086	.129	.137
4.405	.440	.298	.110	.199	5.198	.520	−.099	.130	.140
4.410	.441	.295	.110	.197	5.231	.523	−.116	.131	.144
4.413	.441	.293	.110	.196	5.239	.524	−.120	.131	.145
4.430	.443	.285	.111	.192	5.263	.526	−.132	.132	.149
4.438	.444	.281	.111	.190	5.300	.530	−.150	.132	.155
4.471	.447	.264	.112	.182	5.338	.534	−.169	.133	.162
4.491	.449	.254	.112	.177	5.400	.540	−.200	.135	.175
4.495	.449	.253	.112	.176	5.407	.541	−.203	.135	.176
4.505	.451	.247	.113	.174	5.431	.543	−.215	.136	.182
4.547	.455	.227	.114	.165	5.459	.546	−.229	.136	.189
4.606	.461	.197	.115	.154	5.510	.551	−.255	.138	.203
4.654	.465	.173	.116	.146	5.538	.554	−.269	.138	.211
4.758	.476	.121	.119	.134	5.646	.565	−.323	.141	.246
4.763	.476	.118	.119	.133	5.837	.584	−.419	.146	.321
4.766	.477	.117	.119	.133	5.937	.594	−.468	.148	.368
4.796	.480	.102	.120	.130	6.263	.626	−.631	.157	.555
					Mean	.482	.091	.120	.222

quantities are not presented. For the credibility estimator,

$$\begin{aligned}
\text{Bias} &= \mathrm{E}(0.5\bar{X}_j + 2.5 - \theta_j) = 2.5 - 0.5\theta_j, \\
\text{Variance} &= \mathrm{Var}(0.5\bar{X}_j + 2.5) = \frac{0.25\theta_j}{10} = 0.025\theta_j, \\
\text{Mean-squared error} &= \text{bias}^2 + \text{variance} = 0.25\theta_j^2 - 2.475\theta_j + 6.25.
\end{aligned}$$

We see that, as expected, the average mean-squared error is much lower for the credibility estimator, and this is achieved by allowing for some bias in the individual estimators. Further note that the credibility estimator is at its best near the mean of the prior distribution (5). □

We have seen that there is real value in using credibility. Our next task is to compare the linear credibility estimator to the Bayesian estimator. In

most examples, this is difficult because the Bayesian estimates must be obtained by approximate integration. An alternative would be to explore the mean squared errors by simulation. This approach is taken in an illustration presented in *Foundations of Casualty Actuarial Science* [24], p. 467. In the following example we use the same illustration but employ an approximation that avoids approximate integration. It should also be noted that the linear credibility approach requires only assumptions or estimation of the first two moments while the Bayesian approach requires the distributions to be completely specified. This nonparametric feature makes the linear approach more robust, which may compensate for any loss of accuracy.

Example 16.32 *Individual observations are samples of size* 25 *from an inverse gamma distribution with* $\alpha = 4$ *and unknown scale parameter* Θ. *The prior distribution for* Θ *is gamma with mean* 50 *and variance* 5,000. *Compare the linear credibility and Bayesian estimators.*

For the Bühlmann linear credibility estimator we have

$$\begin{aligned}
\mu &= \mathrm{E}[\mu(\Theta)] = \mathrm{E}\left(\frac{\Theta}{3}\right) = \frac{50}{3},\\
a &= \mathrm{Var}[\mu(\Theta)] = \mathrm{Var}\left(\frac{\Theta}{3}\right) = \frac{5{,}000}{9},\\
v &= \mathrm{E}[v(\Theta)] = \mathrm{E}\left(\frac{\Theta^2}{18}\right) = \frac{5{,}000+50^2}{18} = \frac{7{,}500}{18},
\end{aligned}$$

and so

$$Z = \frac{25}{25+\dfrac{7{,}500/18}{5{,}000/9}} = \frac{100}{103}$$

and the credibility estimator is $\hat{\mu}_{\mathrm{cred}} = (100\bar{X}+50)/103$.

For the Bayesian estimator, the posterior density is

$$\begin{aligned}
\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) &\propto e^{-\theta\sum_{j=1}^{25}x_j^{-1}}\theta^{100}\theta^{-0.5}e^{-\theta/100}\\
&\propto \theta^{99.5}e^{-\theta(0.01+\sum_{j=1}^{25}x_j^{-1})},
\end{aligned}$$

which is a gamma density with parameters 100.5 and $\left(0.01+\sum_{j=1}^{25}x_j^{-1}\right)^{-1}$. The posterior mean is

$$\hat{\theta}_{\mathrm{Bayes}} = \frac{100.5}{0.01+\sum_{j=1}^{25}x_j^{-1}} \quad \text{and so} \quad \hat{\mu}_{\mathrm{Bayes}} = \frac{33.5}{0.01+\sum_{j=1}^{25}x_j^{-1}},$$

which is clearly a nonlinear estimator.

With regard to accuracy, we can also consider the sample mean. Given the value of θ, the sample mean is unbiased with variance and mean squared error

$\theta^2/(18\times 25) = \theta^2/450$. For the credibility estimator the bias is

$$\begin{aligned}
\mathrm{bias}_\theta(\hat{\mu}_{\mathrm{cred}}) &= \mathrm{E}\left(\frac{100\bar{X}}{103}+\frac{50}{103}-\frac{\theta}{3}\right)\\
&= \frac{100\theta}{309}+\frac{50}{103}-\frac{\theta}{3}\\
&= \frac{50}{103}-\frac{\theta}{103},
\end{aligned}$$

the variance is

$$\mathrm{Var}_\theta(\hat{\mu}_{\mathrm{cred}}) = \frac{(100/103)^2\theta^2}{450},$$

and the mean-squared error is

$$\mathrm{MSE}_\theta(\hat{\mu}_{\mathrm{cred}}) = \frac{1}{103^2}\left(2{,}500-100\theta+\frac{10{,}450\theta^2}{450}\right).$$

For the Bayes estimate we observe that, given θ, $1/X$ has a gamma distribution with parameters 4 and $1/\theta$. Therefore, $\sum_{j=1}^{25}X_j^{-1}$ has a gamma distribution with parameters 100 and $1/\theta$. We note that in the denominator of $\hat{\mu}_{\mathrm{Bayes}}$, the term 0.01 will usually be small relative to the sum. An approximation can be created by ignoring this term, in which case $\hat{\mu}_{\mathrm{Bayes}}$ has approximately an inverse gamma distribution with parameters 100 and 33.5θ. Then

$$\begin{aligned}
\mathrm{Bias}_\theta(\hat{\mu}_{\mathrm{Bayes}}) &= \frac{33.5\theta}{99}-\frac{\theta}{3} = \frac{0.5\theta}{99},\\
\mathrm{Var}_\theta(\hat{\mu}_{\mathrm{Bayes}}) &= \frac{33.5^2\theta^2}{99^2(98)},\\
\mathrm{MSE}_\theta(\hat{\mu}_{\mathrm{Bayes}}) &= \frac{33.5^2+49/2}{99^2(98)}\theta^2 = 0.00119391\theta^2.
\end{aligned}$$

If we compare the coefficients of θ^2 in the MSE for the three estimators, we see that they are 0.00222 for the sample mean, 0.00219 for the credibility estimator, and 0.00119 for the Bayesian estimator. Thus for large θ the credibility estimator is not much of an improvement over the sample mean, but the Bayesian estimator cuts the mean squared error about in half. Calculated values of these quantities for various percentiles from the gamma prior distribution appear in Table 16.4. □

The inferior behavior of the credibility estimator when compared with the Bayes estimator is due to the heavy tails of the two distributions. One way to lighten the tail is to work with the logarithm of the data. This idea was proposed in *Foundations of Casualty Actuarial Science* [24] and evaluated for the above example. The idea is to work with the logarithms of the data and

Table 16.4 A comparison of the sample mean, credibility, and Bayes estimators

		$\bar{X}$	$\hat{\mu}_{\text{cred}}$		$\hat{\mu}_{\text{Bayes}}$	
Percentile	θ	MSE	Bias	MSE	Bias	MSE
1	0.008	0.000	0.485	0.236	0.000	0.000
5	0.197	0.000	0.484	0.234	0.001	0.000
10	0.790	0.001	0.478	0.230	0.004	0.001
25	5.077	0.057	0.436	0.244	0.026	0.031
50	22.747	1.150	0.265	1.154	0.115	0.618
75	66.165	9.729	−0.157	9.195	0.334	5.227
90	135.277	40.667	−0.828	39.018	0.683	21.849
95	192.072	81.982	−1.379	79.178	0.970	44.046
99	331.746	244.568	−2.735	238.011	1.675	131.397

use linear credibility to estimate the mean of the distribution of logarithms. The result is then exponentiated. Because this procedure is sure to introduce bias,[5] a multiplicative adjustment is made. The results are presented in the following example with many of the details left for Exercise 16.57.

Example 16.33 (Example 16.32 continued) *Obtain the log-credibility estimator and evaluate its bias and mean-squared error.*

Let $W_j = \ln X_j$. Then for the credibility on the logarithms

$$\begin{aligned}\mu(\Theta) &= \mathrm{E}(W|\Theta)\\ &= \int_0^\infty (\ln x)\Theta^4 x^{-5} e^{-\Theta/x}\tfrac{1}{6}\,dx\\ &= \int_0^\infty (\ln\Theta - \ln y) y^3 e^{-y}\tfrac{1}{6}\,dy\\ &= \ln\Theta - \Psi(4),\end{aligned}$$

where the second integral was obtained using the substitution $y = \Theta/x$. The last line follows from observing that the term $y^3e^{-y}/6$ is a gamma density and thus integrates to 1 while the second term is the **digamma** function (see Exercise 16.57) and using tables in [3] we have $\Psi(4) = 1.25612$. The next

[5] By Jensen's inequality, $\mathrm{E}[\ln X] < \ln\mathrm{E}(X)$, and therefore this procedure will underestimate the true value.

required quantity is

$$\begin{aligned}v(\Theta) &= \mathrm{E}(W^2|\theta) - \mu(\Theta)^2\\ &= \int_0^\infty (\ln x)^2\Theta^4 x^{-5} e^{-\Theta/x}\tfrac{1}{6}\,dx - [\ln\Theta - \Psi(4)]^2\\ &= \int_0^\infty (\ln\Theta - \ln y)^2 y^3 e^{-y}\tfrac{1}{6}\,dy - [\ln\Theta - \Psi(4)]^2\\ &= \Psi'(4),\end{aligned}$$

where $\Psi'(4) = 0.283823$ is the **trigamma** function (see Exercise 16.57). Then

$$\begin{aligned}\mu &= \mathrm{E}[\ln\Theta - \Psi(4)]\\ &= \int_0^\infty (\ln\theta)\theta^{-0.5}e^{-\theta/100}100^{-0.5}\frac{1}{\Gamma(0.5)}\,d\theta - \Psi(4)\\ &= \int_0^\infty (\ln 100 + \ln\lambda)\lambda^{-0.5}e^{-\lambda}\frac{1}{\Gamma(0.5)}\,d\lambda - \Psi(4)\\ &= \ln 100 + \Psi(0.5) - \Psi(4) = 1.38554.\end{aligned}$$

Also,

$$\begin{aligned}v &= \mathrm{E}[\Psi'(4)] = \Psi'(4) = 0.283823\\ a &= \mathrm{Var}[\ln\Theta - \Psi(4)]\\ &= \Psi'(0.5) = 4.934802\\ Z &= \frac{25}{25 + \dfrac{0.283823}{4.934802}} = 0.997705.\end{aligned}$$

The log-credibility estimate is

$$\hat{\mu}_{\text{log-cred}} = c\exp(0.997705\bar{W} + 0.00318024).$$

The value of c is obtained by setting

$$\begin{aligned}\mathrm{E}(X) = \tfrac{50}{3} &= c\mathrm{E}[\exp(0.997705\bar{W} + 0.00318024)]\\ &= ce^{0.00318024}\mathrm{E}\left[\exp\left(\frac{0.997705}{25}\sum_{j=1}^{25}\ln X_j\right)\right]\\ &= ce^{0.00318024}\mathrm{E}\left[\mathrm{E}\left(\prod_{j=1}^{25} X_j^{0.997705/25}|\Theta\right)\right].\end{aligned}$$

Given Θ, the X_js are independent and so the expected product is the product of the expected values. From Appendix A, the kth moment of the inverse

gamma distribution produces

$$\begin{aligned}\frac{50}{3} &= ce^{0.00318024}\mathrm{E}\left\{\left[\tfrac{1}{6}\Theta^{0.997705/25}\Gamma\left(4-\frac{0.997705}{25}\right)\right]^{25}\right\}\\ &= ce^{0.00318024}\left[\tfrac{1}{6}\Gamma\left(4-\frac{0.997705}{25}\right)\right]^{25}\frac{100^{0.997705}\Gamma(0.5+0.997705)}{\Gamma(0.5)},\end{aligned}$$

which produces $c = 1.169318$ and

$$\hat{\mu}_{\text{log-cred}} = 1.173043(2.712051)^{\bar{W}}.$$

In order to evaluate the bias and mean-squared error for a given value of Θ, we must obtain

$$\begin{aligned}\mathrm{E}(\hat{\mu}_{\text{log-cred}}|\Theta=\theta) &= 1.173043\mathrm{E}\left(e^{\bar{W}\ln 2.712051}|\Theta=\theta\right)\\ &= 1.173043\mathrm{E}\left[\prod_{j=1}^{25}X_j^{(\ln 2.712051)/25}|\Theta=\theta\right]\\ &= 1.173043\left[\tfrac{1}{6}\theta^{(\ln 2.712051)/25}\Gamma\left(4-\frac{\ln 2.712051}{25}\right)\right]^{25}\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}(\hat{\mu}^2_{\text{log-cred}}|\Theta &= \theta) = 1.173043^2\mathrm{E}\left(e^{2\bar{W}\ln 2.712051}|\Theta=\theta\right)\\ &= 1.173043^2\left[\tfrac{1}{6}\theta^{(2\ln 2.712051)/25}\Gamma\left(4-\frac{2\ln 2.712051}{25}\right)\right]^{25}.\end{aligned}$$

The measures of quality are then

$$\begin{aligned}\mathrm{Bias}_\theta(\hat{\mu}_{\text{log-cred}}) &= \mathrm{E}(\hat{\mu}_{\text{log-cred}}|\Theta=\theta)-\tfrac{1}{3}\theta,\\ \mathrm{MSE}_\theta(\hat{\mu}_{\text{log-cred}}) &= \mathrm{E}(\hat{\mu}^2_{\text{log-cred}}|\Theta=\theta)-[\mathrm{E}(\hat{\mu}_{\text{log-cred}}|\Theta=\theta)]^2\\ &\quad+[\mathrm{bias}_\theta(\hat{\mu}_{\text{log-cred}})]^2.\end{aligned}$$

Values of these quantities are calculated for various values of θ in Table 16.5. A comparison with Table 16.4 indicates that the log-credibility estimator is almost as good as the Bayes estimator. □

In practice, log-credibility is as easy to use as ordinary credibility. In either case, one of the computational methods of the next section would be used. For log-credibility, the logarithms of the observations are substituted for the observed values and then the final estimate is exponentiated. The bias is corrected by multiplying all the estimates by a constant such that the sample mean of the estimates matches the sample mean of the original data.

Table 16.5 Bias and mean squared error for the log-credibility estimator

Percentile	θ	Bias	MSE
1	0.008	0.000	0.000
5	0.197	0.001	0.000
10	0.790	0.003	0.001
25	5.077	0.012	0.034
50	22.747	0.026	0.666
75	66.165	0.023	5.604
90	135.277	−0.028	23.346
95	192.072	−0.091	46.995
99	331.746	−0.295	139.908

16.4.8 Notes and References

In this section, one of the two major criticisms of limited fluctuation credibility has been addressed. Through the use of the variance of the hypothetical means, we now have a means of relating the mean of the group of interest, $\mu(\theta)$, to the manual, or collective, premium, μ. The development was also mathematically sound in that the results followed directly from a specific model and objective. We have also seen that the additional restriction of a linear solution was not as bad as it might be in that often we still obtain the exact Bayesian solution. There has subsequently been a great deal of effort expended to generalize the model. With a sound basis for obtaining a credibility premium, we have but one remaining obstacle: how to numerically estimate the quantities a and v in the Bühlmann formulation, or how to specify the prior distribution in the Bayesian formulation. Those matters are addressed in the final section of this chapter.

A historical review of credibility theory including a description of the limited fluctuation and greatest accuracy approaches is provided by Norberg [100]. Since the classic paper of Bühlmann [18], there has developed a vast literature on credibility theory in the actuarial literature. Other elementary introductions are given by Herzog [52] and Waters [135]. Other more advanced treatments are Goovaerts and Hoogstad [46] and Sundt [127]. An important generalization of the Bühlmann–Straub model is the Hachemeister [48] regression model, which was not discussed here. See also Klugman [76]. The material on exact credibility is taken from Jewell [66]. See also Ericson [34]. A special issue of *Insurance: Abstracts and Reviews* (Sundt [126]) contains an extensive list of papers on credibility.

16.4.9 Exercises

16.22 Consider a die–spinner model. The first die has one "marked" face and five "unmarked" faces whereas the second die has four "marked" faces and two "unmarked" faces. There are three spinners, each with five equally spaced sectors marked 3 or 8. The first spinner has one sector marked 3 and four marked 8, the second has two marked 3 and three marked 8, and the third has four marked 3 and one marked 8. One die and one spinner are selected at random. If rolling the die produces an unmarked face, no claim occurs. If a marked face occurs, there is a claim and then the spinner is spun once to determine the amount of the claim.

(a) Determine $\pi(\theta)$ for each of the six die–spinner combinations.

(b) Determine the conditional distributions $f_{X|\Theta}(x|\theta)$ for the claim sizes for each die–spinner combination.

(c) Determine the hypothetical means $\mu(\theta)$ and the process variances $v(\theta)$ for each θ.

(d) Determine the marginal probability that the claim X_1 on the first iteration equals 3.

(e) Determine the posterior distribution $\pi_{\Theta|X_1}(\theta|3)$ of Θ using Bayes' theorem.

(f) Use (16.25) to determine the conditional distribution $f_{X_2|X_1}(x_2|3)$ of the claims X_2 on the second iteration given that $X_1 = 3$ was observed on the first iteration.

(g) Use (16.28) to determine the Bayesian premium $\mathrm{E}(X_2|X_1 = 3)$.

(h) Determine the joint probability that $X_2 = x_2$ and $X_1 = 3$ for $x_2 = 0, 3, 8$.

(i) Determine the conditional distribution $f_{X_2|X_1}(x_2|3)$ directly using (16.23) and compare your answer to that of (f).

(j) Determine the Bayesian premium directly using (16.27) and compare your answer to that of (g).

(k) Determine the structural parameters μ, v, and a.

(l) Compute the Bühlmann credibility factor and the Bühlmann credibility premium to approximate the Bayesian premium $\mathrm{E}(X_2|X_1 = 3)$.

16.23 Three urns have balls marked 0, 1, and 2 in the proportions given in Table 16.6. An urn is selected at random, and two balls are drawn from that urn with replacement. A total of 2 on the two balls is observed. Two more balls are then drawn with replacement from the same urn, and it is of interest to predict the total on these next two balls.

(a) Determine $\pi(\theta)$.

Table 16.6 Data for Exercise 16.23

Urn	0s	1s	2s
1	0.40	0.35	0.25
2	0.25	0.10	0.65
3	0.50	0.15	0.35

(b) Determine the conditional distributions $f_{X|\Theta}(x|\theta)$ for the totals on the two balls for each urn.

(c) Determine the hypothetical means $\mu(\theta)$ and the process variances $v(\theta)$ for each θ.

(d) Determine the marginal probability that the total X_1 on the first two balls equals 2.

(e) Determine the posterior distribution $\pi_{\Theta|X_1}(\theta|2)$ using Bayes' theorem.

(f) Use (16.25) to determine the conditional distribution $f_{X_2|X_1}(x_2|2)$ of the total X_2 on the next two balls drawn given that $X_1 = 2$ was observed on the first two draws.

(g) Use (16.28) to determine the Bayesian premium $\mathrm{E}(X_2|X_1 = 2)$.

(h) Determine the joint probability that the total X_2 on the next two balls equals x_2 and the total X_1 on the first two balls equals 2 for $x_2 = 0, 1, 2, 3, 4$.

(i) Determine the conditional distribution $f_{X_2|X_1}(x_2|2)$ directly using (16.23) and compare your answer to that of (f).

(j) Determine the Bayesian premium directly using (16.27) and compare your answer to that of (g).

(k) Determine the structural parameters μ, v, and a.

(l) Determine the Bühlmann credibility factor and the Bühlmann credibility premium.

(m) Show that the Bühlmann credibility factor is the same if each "exposure unit" consists of one draw from the urn rather than two draws.

16.24 Suppose that there are two types of policyholder: type A and type B. Two-thirds of the total number of the policyholders are of type A and one-third are of type B. For each type, the information on annual claim numbers and severity are given as follows:

A policyholder has a total claim amount of 500 in the last four years. Determine the credibility factor Z and the credibility premium for next year for this policyholder.

Type	Number of claims		Severity	
	Mean	Variance	Mean	Variance
A	0.2	0.2	200	4,000
B	0.7	0.3	100	1,500

16.25 Let Θ_1 represent the risk factor for claim numbers and let Θ_2 represent the risk factor for the claim severity for a line of insurance. Suppose that Θ_1 and Θ_2 are independent. Suppose also that given $\Theta_1 = \theta_1$ the claim number N is Poisson distributed and given $\Theta_2 = \theta_2$ the severity Y is exponentially distributed. The expectations of the hypothetical means and process variances for the claim number and severity as well as the variance of the hypothetical means for frequency are respectively

$$\mu_N = 0.1, \quad v_N = 0.1, \quad a_N = 0.05,$$
$$\mu_Y = 100, \quad v_Y = 25{,}000.$$

Three observations are made on a particular policyholder and we observe total claims of 200. Determine the Bühlmann credibility factor and the Bühlmann premium for this policyholder.

16.26 Suppose that $X_1, \ldots, X_n$ are independent (conditional on Θ) and that

$$\mathrm{E}(X_j|\Theta) = \beta_j \mu(\Theta) \text{ and } \mathrm{Var}(X_j|\Theta) = \tau_j(\Theta) + \psi_j v(\Theta), \quad j = 1, \ldots, n.$$

Let

$$\mu = \mathrm{E}[\mu(\Theta)], \quad v = \mathrm{E}[v(\Theta)], \quad \tau_j = \mathrm{E}[\tau_j(\Theta)], \quad a = \mathrm{Var}[\mu(\Theta)].$$

(a) Show that

$$\mathrm{E}(X_j) = \beta_j \mu, \quad \mathrm{Var}(X_j) = \tau_j + \psi_j v + \beta_j^2 a,$$

and

$$\mathrm{Cov}(X_i, X_j) = \beta_i \beta_j a, \quad i \neq j.$$

(b) Solve the normal equations for $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ to show that the credibility premium satisfies

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = (1 - Z)\,\mathrm{E}(X_{n+1}) + Z\beta_{n+1}\bar{X},$$

where

$$\begin{aligned} m_j &= \beta_j^2(\tau_j + \psi_j v)^{-1}, \quad j = 1, \ldots, n, \\ m &= m_1 + \cdots + m_n, \\ Z &= am(1 + am)^{-1}, \\ \bar{X} &= \sum_{j=1}^{n} \frac{m_j}{m} \frac{X_j}{\beta_j}. \end{aligned}$$

16.27 For the situation described in Exercise 12.72 determine $\mu(\theta)$ and the Bayesian premium $\mathrm{E}(X_{n+1}|\mathbf{x})$. Why is the Bayesian premium equal to the credibility premium?

16.28 For the situation described in Exercise 12.73 determine $\mu(\theta)$ and the Bayesian premium $\mathrm{E}(X_{n+1}|\mathbf{x})$ and verify directly that the credibility premium equals the Bayesian premium.

16.29 For the situation described in Exercise 12.74 determine $\mu(\theta)$ and the Bayesian premium $\mathrm{E}(X_{n+1}|\mathbf{x})$ and verify directly that the credibility premium equals the Bayesian premium.

16.30 Consider the generalization of the linear exponential family given by

$$f(x; \theta, m) = \frac{p(m, x)e^{-m\theta x}}{[q(\theta)]^m}.$$

If m is a parameter, this is called the **exponential dispersion family**. In Exercise 12.79 it was shown that the mean of this random variable is $-q'(\theta)/q(\theta)$. For this exercise, assume that m is known.

(a) Consider the prior distribution

$$\pi(\theta) = \frac{[q(\theta)]^{-k} \exp(-\theta\mu k)}{c(\mu, k)}, \quad \theta_0 < \theta < \theta_1 \text{ with } \pi(\theta_0) = \pi(\theta_1).$$

Determine the Bayesian premium.

(b) Using the same prior, determine the Bühlmann premium.

(c) Show that the inverse Gaussian distribution is a member of the exponential dispersion family.

16.31 Suppose that $X_1 \ldots, X_n$ are independent (conditional on Θ) and

$$\mathrm{E}(X_j|\Theta) = \tau^j \mu(\Theta) \quad \text{and} \quad \mathrm{Var}(X_j|\Theta) = \frac{\tau^{2j} v(\Theta)}{m_j}, \quad j = 1, \ldots, n.$$

Let $\mu = \mathrm{E}[\mu(\Theta)]$, $v = \mathrm{E}[v(\Theta)]$, $a = \mathrm{Var}[\mu(\Theta)]$, $k = v/a$, and $m = m_1 + \cdots + m_n$.

(a) Discuss when these assumptions may be appropriate.

(b) Show that

$$\mathrm{E}(X_j) = \tau^j \mu, \quad \mathrm{Var}(X_j) = \tau^{2j}(a + v/m_j),$$

and

$$\mathrm{Cov}(X_i, X_j) = \tau^{i+j} a, \quad i \neq j.$$

(c) Solve the normal equations for $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ to show that the credibility premium satisfies

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = \frac{k}{k+m}\tau^{n+1}\mu + \frac{m}{k+m}\sum_{j=1}^{n} \frac{m_j}{m}\tau^{n+1-j} X_j.$$

(d) Give a verbal interpretation of the formula in (c).

(e) Suppose that

$$f_{X_j|\Theta}(x_j|\theta) = \frac{p(x_j, m_j, \tau)e^{-m_j\tau^{-j}x_j\theta}}{[q(\theta)]^{m_j}}.$$

Show that $\mathrm{E}(X_j|\Theta) = \tau^j\mu(\Theta)$ and that $\mathrm{Var}(X_j|\Theta) = \tau^{2j}v(\Theta)/m_j$, where $\mu(\theta) = -\frac{d}{d\theta}\ln q(\theta)$ and $v(\theta) = -\mu'(\theta)$.

(f) Prove that credibility is exact if Θ has pdf

$$\pi(\theta) = \frac{[q(\theta)]^{-k}e^{-\theta\mu k}}{c(\mu, k)}, \quad \theta_0 < \theta < \theta_1,$$

which satisfies $\pi(\theta_0) = \pi(\theta_1) = 0$.

16.32 Suppose that given $\Theta = \theta$ the random variables $X_1, \cdots, X_n$ are independent with Poisson pf

$$f_{X_j|\Theta}(x_j|\theta) = \frac{\theta^{x_j}e^{-\theta}}{x_j!}, \quad x_j = 0, 1, 2, \ldots.$$

(a) Let $S = X_1 + \cdots + X_n$. Show that S has pf

$$f_S(s) = \int_0^\infty \frac{(n\theta)^s e^{-n\theta}}{s!}\pi(\theta)\, d\theta, \quad s = 0, 1, 2, \ldots,$$

where Θ has pdf $\pi(\theta)$.

(b) Show that the Bayesian premium is

$$\mathrm{E}(X_{n+1}|X_1 + \cdots + X_n = s) = \frac{s+1}{n}\frac{f_S(s+1)}{f_S(s)},$$

where $s = \sum_{j=1}^{n} x_j$.

(c) Evaluate the distribution of S in (a) when $\pi(\theta)$ is a gamma distribution. What type of distribution is this?

16.33 Suppose $X_j|\Theta$ is normally distributed with mean Θ and variance v for $j = 1, 2, \ldots, n+1$. Further suppose Θ is normally distributed with mean μ and variance a. Thus,

$$f_{X_j|\Theta}(x_j|\theta) = (2\pi v)^{-1/2}\exp\left[-\frac{1}{2v}(x_j - \theta)^2\right], \quad -\infty < x_j < \infty,$$

and

$$\pi(\theta) = (2\pi a)^{-1/2}\exp\left[-\frac{1}{2a}(\theta - \mu)^2\right], \quad -\infty < \theta < \infty.$$

Determine the posterior distribution of $\Theta|\mathbf{X}$ and the predictive distribution of $X_{n+1}|\mathbf{X}$. Then determine the Bayesian estimate of $\mathrm{E}(X_{n+1}|\mathbf{X})$. Finally, show that the Bayesian and Bühlmann estimates are equal.

16.34 (*) Your friend selected at random one of two urns and then she pulled a ball with number 4 on it from the urn. Then she replaced the ball in the urn. One of the urns contains four balls, numbered 1–4. The other urn contains six balls, numbered 1–6. Your friend will make another random selection from the same urn.

(a) Estimate the expected value of the number on the next ball using the Bayesian method.

(b) Estimate the expected number on the next ball using Bühlmann credibility.

16.35 The number of claims for a randomly selected insured has the Poisson distribution with parameter θ. The parameter θ is distributed across the population with pdf $\pi(\theta) = 3\theta^{-4}$, $\theta > 1$. For an individual, the parameter does not change over time. A particular insured experienced a total of 20 claims in the previous two years.

(a) (*) Determine the Bühlmann credibility estimate for the future expected claim frequency for this particular insured.

(b) Determine the Bayesian credibility estimate for the future expected claim frequency for this particular insured.

16.36 (*) The distribution of payments to an insured is constant over time. If the Bühlmann credibility assigned for one-half year of observation is 0.5, determine the Bühlmann credibility to be assigned for three years.

16.37 (*) Three urns contain balls marked either 0 or 1. In urn A, 10% are marked 0; in urn B, 60% are marked 0; and in urn C, 80% are marked 0. An urn is selected at random and three balls selected with replacement. The total of the values is 1. Three more balls are selected with replacement from the same urn.

(a) Determine the expected total of the three balls using Bayes' theorem.

(b) Determine the expected total of the three balls using Bühlmann credibility.

16.38 (*) The number of claims follows the Poisson distribution with parameter λ. A particular insured had three claims in the past three years.

(a) The value of λ has pdf $f(\lambda) = 4\lambda^{-5}, \quad \lambda > 1$. Determine the value of K used in Bühlmann's credibility formula. Then use Bühlmann credibility to estimate the claim frequency for this insured.

(b) The value of λ has pdf $f(\lambda) = 1, \quad 0 < \lambda < 1$. Determine the value of K used in Bühlmann's credibility formula. Then use Bühlmann credibility to estimate the claim frequency for this insured.

16.39 (*) The number of claims follows the Poisson distribution with parameter h. The value of h has the gamma distribution with pdf $f(h) = he^{-h}$, $h > 0$. Determine the Bühlmann credibility to be assigned to a single observation. (The Bayes solution was obtained in Exercise 12.86.)

16.40 Consider the situation of Exercise 12.88.

(a) Determine the expected number of claims in the second year using Bayesian credibility.

(b) (*) Determine the expected number of claims in the second year using Bühlmann credibility.

16.41 (*) One spinner is selected at random from a group of three spinners. Each spinner is divided into six equally likely sectors. The number of sectors marked 0, 12, and 48, respectively, on each spinner is as follows: spinner A: 2,2,2; spinner B: 3,2,1; spinner C: 4,1,1. A spinner is selected at random and a zero is obtained on the first spin.

(a) Determine the Bühlmann credibility estimate of the expected value of the second spin using the same spinner.

(b) Determine the Bayesian credibility estimate of the expected value of the second spin using the same spinner.

16.42 The number of claims in a year has the Poisson distribution with mean λ. The parameter λ has the uniform distribution over the interval $(1, 3)$.

(a) (*) Determine the probability that a randomly selected individual will have no claims.

(b) (*) If an insured had one claim during the first year, estimate the expected number of claims for the second year using Bühlmann credibility.

(c) If an insured had one claim during the first year, estimate the expected number of claims for the second year using Bayesian credibility.

16.43 (*) Each of two classes, A and B, has the same number of risks. In class A the number of claims per risk per year has mean $\frac{1}{6}$ and variance $\frac{5}{36}$ while the amount of a single claim has mean 4 and variance 20. In class B

the number of claims per risk per year has mean $\frac{5}{6}$ and variance $\frac{5}{36}$ while the amount of a single claim has mean 2 and variance 5. A risk is selected at random from one of the two classes and is observed for four years.

(a) Determine the value of Z for Bühlmann credibility for the observed pure premium.

(b) Suppose the pure premium calculated from the four observations is 0.25. Determine the Bühlmann credibility estimate for the risk's pure premium.

16.44 (*) Let X_1 be the outcome of a single trial and let $\mathrm{E}(X_2|X_1)$ be the expected value of the outcome of a second trial. You are given the following information:

Outcome, T	$\Pr(X_1 = T)$	Bühlmann estimate of $\mathrm{E}(X_2\|X_1 = T)$	Bayesian estimate of $\mathrm{E}(X_2\|X_1 = T)$
1	1/3	2.72	2.6
8	1/3	7.71	7.8
12	1/3	10.57	–

Determine the Bayesian estimate for $\mathrm{E}(X_2|X_1 = 12)$.

16.45 Consider the situation of Exercise 12.90.

(a) Determine the expected number of claims in the second year using Bayesian credibility.

(b) (*) Determine the expected number of claims in the second year using Bühlmann credibility.

16.46 Consider the situation of Exercise 12.91.

(a) Determine the expected number of claims in the second year using Bayesian credibility.

(b) Determine the expected number of claims in the second year using Bühlmann credibility.

16.47 Two spinners, A_1 and A_2, are used to determine the number of claims. For spinner A_1 there is a 0.15 probability of one claim and 0.85 of no claim. For spinner A_2 there is a 0.05 probability of one claim and 0.95 of no claim. If there is a claim, one of two spinners, B_1 and B_2, is used to determine the amount. Spinner B_1 produces a claim of 20 with probability 0.8 and 40 with probability 0.2. Spinner B_2 produces a claim of 20 with probability 0.3 and 40 with probability 0.7. A spinner is selected at random from each of A_1, A_2 and

from B_1, B_2. Three observations from the selected pair yields claims amounts of 0, 20, and 0.

(a) (*) Use Bühlmann credibility to separately estimate the expected number of claims and the expected severity. Use these estimates to estimate the expected value of the next observation from the same pair of spinners.

(b) Use Bühlmann credibility once on the three observations to estimate the expected value of the next observation from the same pair of spinners.

(c) (*) Repeat parts (a) and (b) using Bayesian estimation.

(d) (*) For the same selected pair of spinners, determine

$$\lim_{n\to\infty} \mathrm{E}(X_n|X_1 = X_2 = \cdots = X_{n-1} = 0).$$

16.48 (*) A portfolio of risks is such that all risks are normally distributed. Those of type A have a mean of 0.1 and a standard deviation of 0.03. Those of type B have a mean of 0.5 and a standard deviation of 0.05. Those of type C have a mean of 0.9 and a standard deviation of 0.01. There are an equal number of each type of risk. The observed value for a single risk is 0.12. Determine the Bayesian estimate of the same risk's expected value.

16.49 (*) You are given the following:

1. The conditional distribution $f_{X|\Theta}(x|\theta)$ is a member of the linear exponential family.
2. The prior distribution $\pi(\theta)$ is a conjugate prior for $f_{X|\Theta}(x|\theta)$.
3. $\mathrm{E}(X) = 1$.
4. $\mathrm{E}(X|X_1 = 4) = 2$, where X_1 is the value of a single observation.
5. The expected value of the process variance $\mathrm{E}[\mathrm{Var}(X|\Theta)] = 3$

Determine the variance of the hypothetical means $\mathrm{Var}[\mathrm{E}(X|\Theta)]$.

16.50 (*) You are given the following:

1. X is a random variable with mean μ and variance v.
2. μ is a random variable with mean 2 and variance 4.
3. v is a random variable with mean 8 and variance 32.

Determine the value of the Bühlmann credibility factor Z after three observations of X.

16.51 The amount of an individual claim has the exponential distribution with pdf $f_{Y|\Lambda}(y|\lambda) = \lambda^{-1}e^{-y/\lambda}$, $y, \lambda > 0$. The parameter λ has the inverse gamma distribution with pdf $\pi(\lambda) = 400\lambda^{-3}e^{-20/\lambda}$.

(a) (*) Determine the unconditional expected value, $\mathrm{E}(X)$.

(b) Suppose two claims were observed with values 15 and 25. Determine the Bühlmann credibility estimate of the expected value of the next claim from the same insured.

(c) Repeat part (b), but determine the Bayesian credibility estimate.

16.52 The distribution of the number of claims is binomial with $n = 1$ and θ unknown. The parameter θ is distributed with mean 0.25 and variance 0.07. Determine the value of Z for a single observation using Bühlmann's credibility formula.

16.53 (*) Consider four marksmen. Each is firing at a target that is 100 feet away. The four targets are 2 feet apart (that is, they lie on a straight line at positions 0, 2, 4, and 6 in feet). The marksmen miss to the left or right, never high or low. Each marksman's shot follows a normal distribution with mean at his target and a standard deviation that is a constant times the distance to the target. At 100 feet the standard deviation is 3 feet. By observing where an unknown marksman's shot hits the straight line, you are to estimate the location of the next shot by the same marksman.

(a) Determine the Bühlmann credibility assigned to a single shot of a randomly selected marksman.

(b) Which of the following will increase Bühlmann credibility the most?

i. Revise the targets to 0, 4, 8, and 12.
ii. Move the marksmen to 60 feet from the targets.
iii. Revise targets to 2, 2, 10, 10.
iv. Increase the number of observations from the same marksman to three.
v. Move two of the marksmen to 50 feet from the targets and increase the number of observations from the same marksman to two.

16.54 (*) Risk 1 produces claims of amounts 100, 1,000, and 20,000 with probabilities 0.5, 0.3, and 0.2, respectively. For risk 2 the probabilities are 0.7, 0.2, and 0.1. Risk 1 is twice as likely as risk 2 of being observed. A claim of 100 is observed, but the observed risk is unknown.

(a) Determine the Bayesian credibility estimate of the expected value of the second claim amount from the same risk.

(b) Determine the Bühlmann credibility estimate of the expected value of the second claim amount from the same risk.

16.55 (*) You are given the following:

1. The number of claims for a single insured follows a Poisson distribution with mean M.
2. The amount of a single claim has an exponential distribution with pdf $f_{X|\Lambda}(x|\lambda) = \lambda^{-1}e^{-x/\lambda}$, $x, \lambda > 0$.
3. M and Λ are independent.
4. $\mathrm{E}(M) = 0.10$ and $\mathrm{Var}(M) = 0.0025$.
5. $\mathrm{E}(\Lambda) = 1{,}000$ and $\mathrm{Var}(\Lambda) = 640{,}000$.
6. The number of claims and the claim amounts are independent.

(a) Determine the expected value of the pure premium's process variance for a single risk.

(b) Determine the variance of the hypothetical means for the pure premium.

16.56 In Example 16.24, if $\rho = 0$, then $Z = 0$, and the estimator is μ. That is, the data should be ignored. However, as ρ increases toward 1, Z increases to 1, and the sample mean becomes the preferred predictor of X_{n+1}. Explain why this is a reasonable result.

16.57 In this exercise you are asked to derive a number of the items from Example 16.33.

(a) The **digamma function** is formally defined as $\Psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$. From this definition, show that

$$\Psi(\alpha) = \frac{1}{\Gamma(\alpha)} \int_0^\infty (\ln x) x^{\alpha-1} e^{-x}\, dx.$$

(b) The **trigamma function** is formally defined as $\Psi'(\alpha)$. Derive an expression for

$$\int_0^\infty (\ln x)^2 x^{\alpha-1} e^{-x}\, dx$$

in terms of trigamma, digamma, and gamma functions.

16.58 Consider the following situation, which is similar to Examples 16.32 and 16.33. Individual observations are samples of size 25 from a lognormal distribution with μ unknown and $\sigma = 2$. The prior distribution for Θ (using Θ to represent the unknown value of μ) is normal with mean 5 and standard deviation 1. Determine the Bayes, credibility, and log-credibility estimators and compare their mean-squared errors, evaluating them at the same percentiles as used in Examples 16.32 and 16.33.

16.59 In the following, let the random vector $\mathbf{X}$ represent all the past data and let X_{n+1} represent the next observation. Let $g(\mathbf{X})$ be any function of the past data.

(a) Prove that the following is true.

$$\begin{aligned}\mathrm{E}\left\{[X_{n+1} - g(\mathbf{X})]^2\right\} &= \mathrm{E}\{[X_{n+1} - \mathrm{E}(X_{n+1}|\mathbf{X})]^2\} \\ &\quad + \mathrm{E}\{[\mathrm{E}(X_{n+1}|\mathbf{X}) - g(\mathbf{X})]^2\},\end{aligned}$$

where the expectation is taken over $(X_{n+1}, \mathbf{X})$.

(b) Show that setting $g(\mathbf{X})$ equal to the Bayesian premium (the mean of the predictive distribution) minimizes the expected squared error, $\mathrm{E}\left\{[X_{n+1} - g(\mathbf{X})]^2\right\}$.

(c) Show that, if $g(\mathbf{X})$ is restricted to be a linear function of the past data, then the expected squared error is minimized by the credibility premium.

16.5 EMPIRICAL BAYES PARAMETER ESTIMATION

In the previous section a modeling methodology was proposed which suggested the use of either the Bayesian or credibility premium as a way to incorporate past data into the prospective rate. There is a practical problem associated with the use of these models which has not yet been addressed.

In the examples, we were able to obtain numerical values for the quantities of interest because the input distributions $f_{X_j|\Theta}(x_j|\theta)$ and $\pi(\theta)$ were assumed to be known. These examples, while useful for illustration of the methodology, can hardly be expected to accurately represent the business of an insurance portfolio. More practical models of necessity involve the use of parameters which must be chosen to ensure a close agreement between the model and reality. Examples of this include: the Poisson–gamma model (Example 16.15), where the gamma parameters α and β need to be selected or the Bühlmann or Bühlmann–Straub parameters μ, v, and a. Assignment of numerical values to the Bayesian or credibility premium requires that these parameters be replaced by numerical values.

In general, the unknown parameters are those associated with the structure density $\pi(\theta)$, and hence we refer to these as **structural parameters**. The terminology we use follows the Bayesian framework of the previous section. Strictly speaking, in the Bayesian context all structural parameters are assumed known and there is no need for estimation. An example of this is the Poisson–gamma where our prior information about the structural density was quantified by the choice of $\alpha = 36$ and $\beta = \frac{1}{240}$. For our purposes, this fully Bayesian approach is often unsatisfactory (e.g., when there is little or no prior information available, such as with a new line of insurance) and we may need

to use the data at hand to estimate the structural (prior) parameters. This approach is called **empirical Bayes estimation.**

We refer to the situation where $\pi(\theta)$ and $f_{X_j|\Theta}(x_j|\theta)$ are left largely unspecified (for example, in the Bühlmann or Bühlmann–Straub models where only the first two moments need be known) as the nonparametric case. This situation is dealt with in Section 16.5.1. If $f_{X_j|\Theta}(x_j|\theta)$ is assumed to be of parametric form (e.g., Poisson, normal, etc.) but not $\pi(\theta)$, then we refer to the problem as being of a semiparametric nature, and this is considered in Section 16.5.2. Finally, the (technically more difficult) fully parametric case where both $f_{X_j|\Theta}(x_j|\theta)$ and $\pi(\theta)$ are assumed to be of parametric form is briefly discussed in Section 16.5.3.

This decision as to whether to select a parametric model or not depends partially on the situation at hand and partially on the judgment and knowledge of the person doing the analysis. For example, an analysis based on claim counts might involve the assumption that $f_{X_j|\Theta}(x_j|\theta)$ is of Poisson form, whereas the choice of a parametric model for $\pi(\theta)$ may not be reasonable.

Any parametric assumptions should be reflected (as far as possible) in parametric estimation. For example, in the Poisson case, because the mean and variance are equal, the same estimate would normally be used for both. Nonparametric estimators would normally be no more efficient than estimators appropriate for the parametric model selected, assuming that the model selected is appropriate. This notion is relevant for the decision as to whether to select a parametric model.

Finally, nonparametric models have the advantage of being appropriate for a wide variety of situations, a fact which may well eliminate the extra burden of a parametric assumption (often a stronger assumption than is reasonable).

In this section the data are assumed to be of the following form. For each of $r \geq 1$ policyholders we have the observed losses per unit of exposure $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})^T$ for $i = 1, \ldots, r$. The random vectors $\{\mathbf{X}_i, i = 1, \ldots, r\}$ are assumed to be statistically independent (experience of different policyholders is assumed to be independent). The (unknown) risk parameter for the ith policyholder is θ_i, $i = 1, \ldots, r$, and it is assumed further that $\theta_1, \ldots, \theta_r$ are realizations of the independent and identically distributed random variables Θ_i with structural density $\pi(\theta_i)$. For fixed i, the (conditional) random variables $X_{ij}|\Theta_i$ are assumed to be independent with pf $f_{X_{ij}|\Theta}(x_{ij}|\theta_i)$, $j = 1, \ldots, n_i$.

Two particularly common cases produce this data format. The first is classification rate making or experience rating. In either, i indexes the classes or groups and j indexes the individual members. The second case is like the first where i continues to index the class or group, but now j is the year and the observation is the average loss for that year. An example of the second setting is Meyers [91], where $i = 1, \ldots, 319$ employment classifications are studied over $j = 1, 2, 3$ years. Regardless of the potential settings, we will refer to the r entities as policyholders.

There may also be a known exposure vector $\mathbf{m}_i = (m_{i1}, m_{i2}, \cdots, m_{in_i})^T$ for policyholder i, where $i = 1, \ldots, r$. If not (and if it is appropriate) one may set $m_{ij} = 1$ in what follows for all i and j. For notational convenience let

$$m_i = \sum_{j=1}^{n_i} m_{ij}, \quad i = 1, \ldots, r,$$

be the total past exposure for policyholder i, and let

$$\bar{X}_i = \frac{1}{m_i} \sum_{j=1}^{n_i} m_{ij} X_{ij}, \quad i = 1, \ldots, r,$$

be the past average loss experience. Furthermore, the total exposure is

$$m = \sum_{i=1}^{r} m_i = \sum_{i=1}^{r} \sum_{j=1}^{n_i} m_{ij}$$

and the overall average losses are

$$\bar{X} = \frac{1}{m} \sum_{i=1}^{r} m_i \bar{X}_i = \frac{1}{m} \sum_{i=1}^{r} \sum_{j=1}^{n_i} m_{ij} X_{ij}. \tag{16.54}$$

The parameters which need to be estimated depend on what is assumed about the distributions $f_{X_{ij}|\Theta}(x_{ij}|\theta_i)$ and $\pi(\theta)$.

For the Bühlmann–Straub formulation there are additional quantities of interest. The hypothetical mean (assumed not to depend on j) is

$$\mathrm{E}(X_{ij}|\Theta_i = \theta_i) = \mu(\theta_i)$$

and the process variance is

$$\mathrm{Var}(X_{ij}|\Theta_i = \theta_i) = \frac{v(\theta_i)}{m_{ij}}.$$

The structural parameters are

$$\mu = \mathrm{E}[\mu(\Theta_i)], \quad v = \mathrm{E}[v(\Theta_i)],$$

and

$$a = \mathrm{Var}[\mu(\Theta_i)].$$

The approach to be followed in this section is to estimate μ, v, and a (when unknown) from the data. The credibility premium for next year's losses (per exposure unit) for policyholder i is

$$Z_i \bar{X}_i + (1 - Z_i)\mu, \quad i = 1, \ldots, r, \tag{16.55}$$

where

$$Z_i = \frac{m_i}{m_i + k}, \quad k = \frac{v}{a}.$$

If estimators of μ, v, and a are denoted by $\hat{\mu}, \hat{v}$, and $\hat{a}$, respectively, then one would replace the credibility premium (16.55) by its estimator

$$\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i)\hat{\mu}, \tag{16.56}$$

where

$$\hat{Z}_i = \frac{m_i}{m_i + \hat{k}}, \quad \hat{k} = \frac{\hat{v}}{\hat{a}}.$$

Note that, even if $\hat{v}$ and $\hat{a}$ are unbiased estimators of v and a, the same cannot be said of $\hat{k}$ and $\hat{Z}_i$. Finally, the credibility premium to cover all m_{i,n_i+1} exposure units for policyholder i in the next year would be (16.56) multiplied by m_{i,n_i+1}.

16.5.1 Nonparametric estimation

In this section we consider unbiased estimation of μ, v, and a. To illustrate the ideas, let us begin with the following simple Bühlmann-type example.

Example 16.34 *Suppose that $n_i = n > 1$ for all i and $m_{ij} = 1$ for all i and j. That is, for policyholder i, we have the loss vector*

$$\mathbf{X}_i = (X_{i1}, \ldots, X_{in})^T, \quad i = 1, \ldots, r.$$

Furthermore, conditional on $\Theta_i = \theta_i$, X_{ij} has mean

$$\mu(\theta_i) = E(X_{ij}|\Theta_i = \theta_i)$$

and variance

$$v(\theta_i) = \mathrm{Var}(X_{ij}|\Theta_i = \theta_i),$$

and $X_{i1}, \ldots, X_{in}$ are independent (conditionally). Also, different policyholders' past data are independent, so that if $i \neq s$, then X_{ij} and X_{st} are independent. In this case

$$\bar{X}_i = n^{-1}\sum_{j=1}^{n} X_{ij} \text{ and } \bar{X} = r^{-1}\sum_{i=1}^{r}\bar{X}_i = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n} X_{ij}.$$

Determine unbiased estimators of the Bühlmann quantities.

An unbiased estimator of μ is

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

because

$$\begin{aligned}\mathrm{E}(\hat{\mu}) &= (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}\mathrm{E}(X_{ij}) = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}\mathrm{E}[\mathrm{E}(X_{ij}|\Theta_i)] \\ &= (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}\mathrm{E}[\mu(\Theta_i)] = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}\mu = \mu.\end{aligned}$$

To estimate v, consider

$$\hat{v}_i = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2.$$

Recall that for fixed i the random variables $X_{i1}, \ldots, X_{in}$ are independent, conditional on $\Theta_i = \theta_i$. Thus, $\hat{v}_i$ is an unbiased estimate of $\mathrm{Var}(X_{ij}|\Theta_i = \theta_i) = v(\theta_i)$. Unconditionally,

$$\mathrm{E}(\hat{v}_i) = \mathrm{E}[\mathrm{E}(\hat{v}_i|\Theta_i)] = \mathrm{E}[v(\Theta_i)] = v$$

and $\hat{v}_i$ is unbiased for v. Hence an unbiased estimator of v is

$$\hat{v} = \frac{1}{r}\sum_{i=1}^{r}\hat{v}_i = \frac{1}{r(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2. \tag{16.57}$$

We now turn to estimation of the parameter a. Begin with

$$\mathrm{E}(\bar{X}_i|\Theta_i = \theta_i) = n^{-1}\sum_{j=1}^{n}\mathrm{E}(X_{ij}|\Theta_i = \theta_i) = n^{-1}\sum_{j=1}^{n}\mu(\theta_i) = \mu(\theta_i).$$

Thus,

$$\mathrm{E}(\bar{X}_i) = \mathrm{E}[\mathrm{E}(\bar{X}_i|\Theta_i)] = \mathrm{E}[\mu(\Theta_i)] = \mu$$

and

$$\begin{aligned}\mathrm{Var}(\bar{X}_i) &= \mathrm{Var}[\mathrm{E}(\bar{X}_i|\Theta_i)] + \mathrm{E}[\mathrm{Var}(\bar{X}_i|\Theta_i)] \\ &= \mathrm{Var}[\mu(\Theta_i)] + \mathrm{E}\left[\frac{v(\Theta_i)}{n}\right] = a + \frac{v}{n}.\end{aligned}$$

Therefore, $\bar{X}_1, \ldots, \bar{X}_r$ are independent with common mean μ and common variance $a + v/n$. Their sample average is $\bar{X} = r^{-1}\sum_{i=1}^{r}\bar{X}_i$. Consequently, an unbiased estimator of $a + v/n$ is $(r-1)^{-1}\sum_{i=1}^{r}\left(\bar{X}_i - \bar{X}\right)^2$. Because we already have an unbiased estimator of v given above, an unbiased estimator of a is given by

$$\begin{aligned}\hat{a} &= \frac{1}{r-1}\sum_{i=1}^{r}\left(\bar{X}_i - \bar{X}\right)^2 - \frac{\hat{v}}{n} \\ &= \frac{1}{r-1}\sum_{i=1}^{r}\left(\bar{X}_i - \bar{X}\right)^2 - \frac{1}{rn(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2. \quad (16.58)\end{aligned}$$

□

These estimators might look familiar. Consider a one-factor analysis of variance in which each policyholder represents a treatment. The estimator

for v (16.57) is the within (also called the error) mean square. The first term in the estimator for a (16.58) is the between (also called the treatment) mean square divided by n. The hypothesis that all treatments have the same mean is accepted when the between mean square is small relative to the within mean square—that is, when $\hat{a}$ is small relative to $\hat{v}$. But that implies $\hat{Z}$ will be near zero and little credibility will be given to each $\bar{X}_i$. This is as it should be when the policyholders are essentially identical.

Due to the subtraction in (16.58), it is possible that $\hat{a}$ could be negative. When that happens, it is customary to set $\hat{a} = \hat{Z} = 0$. This case is equivalent to the F test statistic in the analysis of variance being less than 1, a case that always leads to an acceptance of the hypothesis of equal means.

Example 16.35 (Example 16.34 continued) *As a numerical illustration, suppose we have $r = 2$ policyholders with $n = 3$ years experience for each. Let the losses be $\mathbf{x}_1 = (3,5,7)^T$ and $\mathbf{x}_2 = (6,12,9)^T$. Estimate the Bühlmann credibility premiums for each policyholder.*

We have

$$\bar{X}_1 = \tfrac{1}{3}(3+5+7) = 5, \ \bar{X}_2 = \tfrac{1}{3}(6+12+9) = 9$$

and so $\bar{X} = \frac{1}{2}(5+9) = 7$. Then $\hat{\mu} = 7$. We next have

$$\begin{aligned} \hat{v}_1 &= \tfrac{1}{2}[(3-5)^2 + (5-5)^2 + (7-5)^2] = 4, \\ \hat{v}_2 &= \tfrac{1}{2}[(6-9)^2 + (12-9)^2 + (9-9)^2] = 9, \end{aligned}$$

and so $\hat{v} = \frac{1}{2}(4+9) = \frac{13}{2}$. Then

$$\hat{a} = [(5-7)^2 + (9-7)^2] - \tfrac{1}{3}\hat{v} = \tfrac{35}{6}.$$

Next, $\hat{k} = \hat{v}/\hat{a} = \frac{39}{35}$ and the estimated credibility factor is $\hat{Z} = 3/(3+\hat{k}) = \frac{35}{48}$. The estimated credibility premiums are

$$\begin{aligned} \hat{Z}\bar{X}_1 + (1-\hat{Z})\hat{\mu} &= \left(\tfrac{35}{48}\right)(5) + \left(\tfrac{13}{48}\right)(7) = \tfrac{133}{24}, \\ \hat{Z}\bar{X}_2 + (1-\hat{Z})\hat{\mu} &= \left(\tfrac{35}{48}\right)(9) + \left(\tfrac{13}{48}\right)(7) = \tfrac{203}{24} \end{aligned}$$

for policyholders 1 and 2 respectively. □

We now turn to the more general Bühlmann–Straub setup described earlier in this section. We have $\mathrm{E}(X_{ij}) = \mathrm{E}[\mathrm{E}(X_{ij}|\Theta_i)] = \mathrm{E}[\mu(\Theta_i)] = \mu$. Thus,

$$\mathrm{E}(\bar{X}_i|\Theta_i) = \sum_{j=1}^{n_i} \frac{m_{ij}}{m_i} \mathrm{E}(X_{ij}|\Theta_i) = \sum_{j=1}^{n_i} \frac{m_{ij}}{m_i} \mu(\Theta_i) = \mu(\Theta_i),$$

implying that

$$\mathrm{E}(\bar{X}_i) = \mathrm{E}[\mathrm{E}(\bar{X}_i|\Theta_i)] = \mathrm{E}[\mu(\Theta_i)] = \mu.$$

Finally,

$$\mathrm{E}(\bar{X}) = \frac{1}{m}\sum_{i=1}^{r} m_i \mathrm{E}(\bar{X}_i) = \frac{1}{m}\sum_{i=1}^{r} m_i \mu = \mu$$

and so an obvious unbiased estimator of μ is

$$\hat{\mu} = \bar{X}. \tag{16.59}$$

Now, $\mathrm{E}(X_{ij}|\Theta_i) = \mu(\Theta_i)$ and $\mathrm{Var}(X_{ij}|\Theta_i) = v(\Theta_i)/m_{ij}$ for $j = 1,\ldots,n_i$. Consider

$$\hat{v}_i = \frac{\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{n_i - 1}, \quad i = 1,\ldots,r. \tag{16.60}$$

Condition on Θ_i and use (16.12) with $\beta = 0$ and $\alpha = v(\Theta_i)$. Then $\mathrm{E}(\hat{v}_i|\Theta_i) = v(\Theta_i)$. But this means that, unconditionally,

$$\mathrm{E}(\hat{v}_i) = \mathrm{E}[\mathrm{E}(\hat{v}_i|\Theta_i)] = \mathrm{E}[v(\Theta_i)] = v$$

and so $\hat{v}_i$ is unbiased for v for $i = 1,\ldots,r$. Another unbiased estimator for v is then the weighted average $\hat{v} = \sum_{i=1}^{r} w_i \hat{v}_i$, where $\sum_{i=1}^{r} w_i = 1$. If we choose weights proportional to $n_i - 1$, we weight the original X_{ij}s by m_{ij}. That is, with $w_i = (n_i - 1)/\sum_{i=1}^{r}(n_i - 1)$, we obtain an unbiased estimator of v, namely,

$$\hat{v} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^{r}(n_i - 1)}. \tag{16.61}$$

We now turn to estimation of a. Recall that for fixed i the random variables $X_{i1},\ldots,X_{in_i}$ are independent, conditional on Θ_i. Thus,

$$\begin{aligned} \mathrm{Var}(\bar{X}_i|\Theta_i) &= \sum_{j=1}^{n_i}\left(\frac{m_{ij}}{m_i}\right)^2 \mathrm{Var}(X_{ij}|\Theta_i) = \sum_{j=1}^{n_i}\left(\frac{m_{ij}}{m_i}\right)^2 \frac{v(\Theta_i)}{m_{ij}} \\ &= \frac{v(\Theta_i)}{m_i^2}\sum_{j=1}^{n_i} m_{ij} = \frac{v(\Theta_i)}{m_i}. \end{aligned}$$

But this means that, unconditionally,

$$\begin{aligned} \mathrm{Var}(\bar{X}_i) &= \mathrm{Var}[\mathrm{E}(\bar{X}_i|\Theta_i)] + \mathrm{E}[\mathrm{Var}(\bar{X}_i|\Theta_i)] \\ &= \mathrm{Var}[\mu(\Theta_i)] + \mathrm{E}\left[\frac{v(\Theta_i)}{m_i}\right] = a + \frac{v}{m_i}. \end{aligned} \tag{16.62}$$

To summarize, $\bar{X}_1,\ldots,\bar{X}_r$ are independent with common mean μ and variances $\mathrm{Var}(\bar{X}_i) = a + v/m_i$. Furthermore, $\bar{X} = m^{-1}\sum_{i=1}^{r} m_i \bar{X}_i$. Now, (16.12) may again be used with $\beta = a$ and $\alpha = v$ to yield

$$\mathrm{E}\left[\sum_{i=1}^{r} m_i\left(\bar{X}_i - \bar{X}\right)^2\right] = a\left(m - m^{-1}\sum_{i=1}^{r} m_i^2\right) + v(r-1).$$

An unbiased estimator for a may be obtained by replacing v by an unbiased estimator $\hat{v}$ and "solving" for a. That is, an unbiased estimator of a is

$$\hat{a} = \left(m - m^{-1} \sum_{i=1}^{r} m_i^2 \right)^{-1} \left[\sum_{i=1}^{r} m_i (\bar{X}_i - \bar{X})^2 - \hat{v}(r-1) \right] \tag{16.63}$$

with $\hat{v}$ given by (16.61). An alternative form of (16.63) is given in Exercise 16.67.

Some remarks are in order at this point. Equations (16.59), (16.61), and (16.63) provide unbiased estimators for μ, v, and a, respectively. They are nonparametric, requiring no distributional assumptions. They are certainly not the only (unbiased) estimators which could be used, and it is possible that $\hat{a} < 0$. In this case, a is likely to be close to 0, and it makes sense to set $\hat{Z} = 0$. Furthermore, the ordinary Bühlmann estimators of Example 16.34 are recovered with $m_{ij} = 1$ and $n_i = n$. Furthermore, as may be seen from Example 16.41, these estimators are essentially maximum likelihood estimators in the case where $X_{ij}|\Theta_i$ and Θ_i are both normally distributed, and thus the estimators have good statistical properties.

There is one problem using the formulas developed above. In the past, the data from the ith policyholder was collected on an exposure of m_i. Total losses on all policyholders was $TL = \sum_{i=1}^{r} m_i \bar{X}_i$. If we had charged the credibility premium as given above, the total premium would have been

$$\begin{aligned} TP &= \sum_{i=1}^{r} m_i [\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i)\hat{\mu}] \\ &= \sum_{i=1}^{r} m_i (1 - \hat{Z}_i)(\hat{\mu} - \bar{X}_i) + \sum_{i=1}^{r} m_i \bar{X}_i \\ &= \sum_{i=1}^{r} m_i \frac{\hat{k}}{m_i + \hat{k}} (\hat{\mu} - \bar{X}_i) + \sum_{i=1}^{r} m_i \bar{X}_i. \end{aligned}$$

It is often desirable for TL to equal TP. The reason is that any premium increases that will meet the approval of regulators will be based on the total claim level from past experience. While credibility adjustments make both practical and theoretical sense, it is usually a good idea to keep the total unchanged. For this to happen, we need

$$0 = \sum_{i=1}^{r} m_i \frac{\hat{k}}{m_i + \hat{k}} (\hat{\mu} - \bar{X}_i)$$

or

$$\hat{\mu} \sum_{i=1}^{r} \hat{Z}_i = \sum_{i=1}^{r} \hat{Z}_i \bar{X}_i$$

or

$$\hat{\mu} = \frac{\sum_{i=1}^{r} \hat{Z}_i \bar{X}_i}{\sum_{i=1}^{r} \hat{Z}_i}. \tag{16.64}$$

Table 16.7 Data for Example 16.36

	Policyholder	Year 1	Year 2	Year 3	Year 4
Total claims	1	–	10,000	13,000	–
No. in group		–	50	60	75
Total claims	2	18,000	21,000	17,000	–
No. in group		100	110	105	90

That is, rather than using (16.59) to compute $\hat{\mu}$, use a credibility-weighted average of the individual sample means. Either method provides an unbiased estimator (given the $\hat{Z}_i$s), but this latter one has the advantage of preserving total claims. It should be noted that when using (16.63), the value of $\bar{X}$ from (16.54) should still be used. It can also be derived by least squares arguments. Finally, from Example 16.7 and noting the form of $\text{Var}(\bar{X}_j)$ in (16.62), the weights in (16.64) provide the smallest unconditional variance for $\hat{\mu}$.

Example 16.36 *Past data on two group policyholders are available and are given in Table* 16.7. *Determine the estimated credibility premium to be charged to each group in year* 4.

We first need to determine the average claims per person for each group in each past year. We have $n_1 = 2$ years experience for group 1 and $n_2 = 3$ for group 2. It is immaterial which past years' data we have for policyholder 1, so for notational purposes we will choose

$$m_{11} = 50 \text{ and } X_{11} = \frac{10{,}000}{50} = 200.$$

Similarly,

$$m_{12} = 60 \text{ and } X_{12} = \frac{13{,}000}{60} = 216.67.$$

Then

$$\begin{aligned} m_1 &= m_{11} + m_{12} = 50 + 60 = 110, \\ \bar{X}_1 &= \frac{10{,}000 + 13{,}000}{110} = 209.09. \end{aligned}$$

For policyholder 2,

$$\begin{aligned} m_{21} &= 100, \; X_{21} = \frac{18{,}000}{100} = 180, \\ m_{22} &= 110, \; X_{22} = \frac{21{,}000}{110} = 190.91, \\ m_{23} &= 105, \; X_{23} = \frac{17{,}000}{105} = 161.90. \end{aligned}$$

Then

$$\begin{aligned} m_2 &= m_{21}+m_{22}+m_{23}=100+110+105=315, \\ \bar{X}_2 &= \frac{18{,}000+21{,}000+17{,}000}{315}=177.78. \end{aligned}$$

Now, $m=m_1+m_2=110+315=425$. The overall mean is

$$\hat{\mu}=\bar{X}=\frac{10{,}000+13{,}000+18{,}000+21{,}000+17{,}000}{425}=185.88.$$

The alternative estimate of μ (16.64) cannot be computed until later.

Now,

$$\begin{aligned} \hat{v} &= \frac{\begin{array}{c}50(200-209.09)^2+60(216.67-209.09)^2+100(180-177.78)^2 \\ +110(190.91-177.78)^2+105(161.90-177.78)^2\end{array}}{(2-1)+(3-1)} \\ &= 17{,}837.87 \end{aligned}$$

and so

$$\begin{aligned} \hat{a} &= \frac{110(209.09-185.88)^2+315(177.78-185.88)^2-(17{,}837.87)(1)}{425-\left(110^2+315^2\right)/425} \\ &= 380.76. \end{aligned}$$

Then $\hat{k}=\hat{v}/\hat{a}=46.85$. The estimated credibility factors for the two policyholders are

$$\hat{Z}_1=\frac{110}{110+46.85}=0.70, \quad \hat{Z}_2=\frac{315}{315+46.85}=0.87.$$

Per individual the estimated credibility premium for policyholder 1 is

$$\hat{Z}_1\bar{X}_1+(1-\hat{Z}_1)\hat{\mu}=(0.70)(209.09)+(0.30)(185.88)=202.13$$

and so the total estimated credibility premium for the whole group is

$$75(202.13)=15,159.75.$$

For policyholder 2,

$$\hat{Z}_2\bar{X}_2+(1-\hat{Z}_2)\hat{\mu}=(0.87)(177.78)+(0.13)(185.88)=178.83$$

and the total estimated credibility premium is

$$90(178.83)=16,094.70.$$

For the alternative estimator we would use

$$\hat{\mu}=\frac{0.70(209.09)+0.87(177.78)}{0.70+0.87}=191.74.$$

The credibility premiums are

$$0.70(209.09)+0.30(191.74)=203.89, \quad 0.87(177.78)+0.13(191.74)=179.59.$$

The total past credibility premium is $110(203.89)+315(179.59)=78{,}998.75$. Except for rounding error, this matches the actual total losses of 79,000. □

The above analysis assumes that the parameters μ, v, and a are all unknown and need to be estimated, and this may not always be the case. Also, it is assumed that $n_i>1$ and $r>1$. If $n_i=1$, so that there is only one exposure unit's experience for policyholder i, it is difficult to obtain information on the process variance $v(\Theta_i)$ and thus v. Similarly, if $r=1$, there is only one policyholder and it is difficult to obtain information on the variance of the hypothetical means a. In these situations, stronger assumptions are needed such as knowledge of one or more of the parameters (e.g., the pure premium or manual rate μ, discussed below) or parametric assumptions which imply functional relationships between the parameters (discussed in Sections 16.5.2 and 16.5.3).

To illustrate these ideas, suppose, for example, that the manual rate μ may be already known, but estimates of a and v may be needed. In that case, (16.61) can still be used to estimate v as it is unbiased whether μ is known or not. (Why is $\left[\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\mu)^2\right]/n_i$ not unbiased for v in this case?) Similarly, (16.63) is still an unbiased estimator for a. However, if μ is known, an alternative unbiased estimator for a is

$$\tilde{a}=\sum_{i=1}^{r}\frac{m_i}{m}(\bar{X}_i-\mu)^2-\frac{r}{m}\hat{v},$$

where $\hat{v}$ is given by (16.61). To see this, note that

$$\begin{aligned} \mathrm{E}(\tilde{a}) &= \sum_{i=1}^{r}\frac{m_i}{m}E[(\bar{X}_i-\mu)^2]-\frac{r}{m}E(\hat{v}) \\ &= \sum_{i=1}^{r}\frac{m_i}{m}\operatorname{Var}(\bar{X}_i)-\frac{r}{m}v \\ &= \sum_{i=1}^{r}\frac{m_i}{m}\left(a+\frac{v}{m_i}\right)-\frac{r}{m}v=a. \end{aligned}$$

If there are data on only one policyholder, an approach like this is necessary. Clearly, (16.60) provides an estimator for v based on data from policyholder i alone, and an unbiased estimator for a based on data from policyholder i alone is

$$\tilde{a}_i=(\bar{X}_i-\mu)^2-\frac{\hat{v}_i}{m_i}=(\bar{X}_i-\mu)^2-\frac{\sum_{j=1}^{n_i}m_{ij}(X_{ij}-\bar{X}_i)^2}{m_i(n_i-1)},$$

which is unbiased because $\mathrm{E}[(\bar{X}_i - \mu)^2] = \mathrm{Var}(\bar{X}_i) = a + v/m_i$ and $\mathrm{E}(\hat{v}_i) = v$.

Example 16.37 *For a group policyholder, we have the following data available:*

	Year 1	Year 2	Year 3
Total claims	60,000	70,000	–
No. in group	125	150	200

If the manual rate per person is 500 *per year, estimate the total credibility premium for year* 3.

In the above notation, we have (assuming for notational purposes that this group is policyholder i) $m_{i1} = 125$, $X_{i1} = 60{,}000/125 = 480$, $m_{i2} = 150$, $X_{i2} = 70{,}000/150 = 466.67$, $m_i = m_{i1} + m_{i2} = 275$, and $\bar{X}_i = (60{,}000 + 70{,}000)/275 = 472.73$. Then

$$\hat{v}_i = \frac{125(480 - 472.73)^2 + 150(466.67 - 472.73)^2}{2 - 1} = 12{,}115.15,$$

and with $\mu = 500$, $\tilde{a}_i = (472.73 - 500)^2 - (12{,}115.15/275) = 699.60$. We then estimate k by $\hat{v}_i/\tilde{a}_i = 17.32$. The estimated credibility factor is $m_i/(m_i + \hat{v}_i/\tilde{a}_i) = 275/(275 + 17.32) = 0.94$. The estimated credibility premium per person is then $0.94(472.73) + 0.06(500) = 474.37$ and the estimated total credibility premium for year 3 is $200(474.37) = 94{,}874$. □

It is instructive to note that estimation of the parameters a and v based on data from a single policyholder (as in the example above) is not advised unless there is no alternative because the estimators $\hat{v}_i$ and $\tilde{a}_i$ have high variability. In particular, we are effectively estimating a from one observation ($\bar{X}_i$). It is strongly suggested that an attempt be made to obtain more data.

16.5.2 Semiparametric estimation

In some situations it may be reasonable to assume a parametric form for the conditional distribution $f_{X_{ij}|\Theta}(x_{ij}|\theta_i)$. The situation at hand may suggest that such an assumption is reasonable or prior information may imply its appropriateness.

For example, in dealing with numbers of claims, it may be reasonable to assume that the number of claims $m_{ij}X_{ij}$ for policyholder i in year j is Poisson distributed with mean $m_{ij}\theta_i$ given $\Theta_i = \theta_i$. Thus $\mathrm{E}(m_{ij}X_{ij}|\Theta_i) = \mathrm{Var}(m_{ij}X_{ij}|\Theta_i) = m_{ij}\Theta_i$, implying that $\mu(\Theta_i) = v(\Theta_i) = \Theta_i$ and so $\mu = v$ in this case. Rather than use (16.61) to estimate v, we could use $\hat{\mu} = \bar{X}$ to estimate v.

Example 16.38 *In the past year, the distribution of automobile insurance policyholders by number of claims is given below.*

No. of claims	No. of insureds
0	1,563
1	271
2	32
3	7
4	2
Total	1,875

For each policyholder, obtain a credibility estimate for the number of claims next year based on the past year's experience, assuming a (conditional) Poisson distribution of number of claims for each policyholder.

Assume that we have $r = 1{,}875$ policyholders, $n_i = 1$ year experience on each, and exposures $m_{ij} = 1$. For policyholder i (where $i = 1, \ldots, 1{,}875$) assume that $X_{i1}|\Theta_i = \theta_i$ is Poisson distributed with mean θ_i so that $\mu(\theta_i) = v(\theta_i) = \theta_i$ and $\mu = v$. As in Example 16.34,

$$\begin{aligned} \bar{X} &= \frac{1}{1{,}875}\left(\sum_{i=1}^{1{,}875} X_{i1}\right) \\ &= \frac{0(1{,}563) + 1(271) + 2(32) + 3(7) + 4(2)}{1{,}875} = 0.194. \end{aligned}$$

Now,

$$\begin{aligned} \mathrm{Var}(X_{i1}) &= \mathrm{Var}[\mathrm{E}(X_{i1}|\Theta_i)] + \mathrm{E}[\mathrm{Var}(X_{i1}|\Theta_i)] \\ &= \mathrm{Var}[\mu(\Theta_i)] + \mathrm{E}[v(\Theta_i)] = a + v = a + \mu. \end{aligned}$$

Thus an unbiased estimator of $a + v$ is the sample variance

$$\begin{aligned} \frac{\sum_{i=1}^{1{,}875}\left(X_{i1} - \bar{X}\right)^2}{1{,}874} &= \frac{\begin{array}{c}1{,}563(0 - 0.194)^2 + 271(1 - 0.194)^2 \\ +32(2 - 0.194)^2 + 7(3 - 0.194)^2 + 2(4 - 0.194)^2\end{array}}{1{,}874} \\ &= 0.226. \end{aligned}$$

Thus $\hat{a} = 0.226 - 0.194 = 0.032$ and $\hat{k} = 0.194/0.032 = 6.06$ and the credibility factor Z is $1/(1 + 6.06) = 0.14$. The estimated credibility premium for the number of claims for each policyholder is $(0.14)X_{i1} + (0.86)(0.194)$, where X_{i1} is 0, 1, 2, 3, or 4, depending on the policyholder. □

Note that in this case $v = \mu$ identically, so that only one year's experience per policyholder is needed.

Example 16.39 *Suppose we are interested in the probability that an individual in a group makes a claim (e.g., group life insurance), and the probability is believed to vary by policyholder. Then $m_{ij}X_{ij}$ could represent the number of the m_{ij} individuals in year j for policyholder i who made a claim. Develop a credibility model for this situation.*

If the claim probability is θ_i for policyholder i, then a reasonable model to describe this effect is that $m_{ij}X_{ij}$ is binomially distributed with parameters m_{ij} and θ_i, given $\Theta_i = \theta_i$. Then

$$\mathrm{E}(m_{ij}X_{ij}|\Theta_i) = m_{ij}\Theta_i \quad \text{and} \quad \mathrm{Var}(m_{ij}X_{ij}|\Theta_i) = m_{ij}\Theta_i(1-\Theta_i)$$

and so $\mu(\Theta_i) = \Theta_i$ with $v(\Theta_i) = \Theta_i(1-\Theta_i)$. Thus

$$\begin{aligned} \mu &= \mathrm{E}(\Theta_i), \quad v = \mu - \mathrm{E}[(\Theta_i)^2], \\ a &= \mathrm{Var}(\Theta_i) = \mathrm{E}[(\Theta_i)^2] - \mu^2 = \mu - v - \mu^2. \end{aligned}$$ □

In these examples there is a functional relationship between the parameters μ, v, and a which follows from the parametric assumptions made, and this often facilitates estimation of parameters.

16.5.3 Parametric estimation

If fully parametric assumptions are made with respect to $f_{X_{ij}|\Theta}(x_{ij}|\theta_i)$ and $\pi(\theta_i)$ for $i = 1,\ldots,r$ and $j = 1,\ldots,n_i$, then the full battery of parametric estimation techniques are available in addition to the nonparametric methods discussed earlier. In particular, maximum likelihood estimation is straightforward (at least in principle) and is now discussed. For policyholder i, the joint density of $\mathbf{X}_i = (X_{i1},\ldots,X_{in_i})^T$ is, by conditioning on Θ_i, given for $i = 1,\ldots,r$ by

$$f_{\mathbf{X}_i}(\mathbf{x}_i) = \int \left[\prod_{j=1}^{n_i} f_{X_{ij}|\Theta}(x_{ij}|\theta_i)\right] \pi(\theta_i)\, d\theta_i. \tag{16.65}$$

The likelihood function is given by

$$L = \prod_{i=1}^{r} f_{\mathbf{X}_i}(\mathbf{x}_i). \tag{16.66}$$

Maximum likelihood estimators of the parameters are then chosen to maximize L or equivalently $\ln L$.

Example 16.40 *As a simple example, suppose that $n_i = n$ for $i = 1,\ldots,r$ and $m_{ij} = 1$. Let $X_{ij}|\Theta_i$ be Poisson distributed with mean Θ_i, that is,*

$$f_{X_{ij}|\Theta}(x_{ij}|\theta_i) = \frac{\theta_i^{x_{ij}} e^{-\theta_i}}{x_{ij}!}, \quad x_{ij} = 0,1,\ldots,$$

and let Θ_i be exponentially distributed with mean μ,

$$\pi(\theta_i) = \frac{1}{\mu} e^{-\theta_i/\mu}, \quad \theta_i > 0.$$

Determine the maximum likelihood estimator of μ.

Equation (16.65) becomes

$$\begin{aligned} f_{\mathbf{X}_i}(\mathbf{x}_i) &= \int_0^\infty \left(\prod_{j=1}^{n} \frac{\theta_i^{x_{ij}} e^{-\theta_i}}{x_{ij}!}\right) \frac{1}{\mu} e^{-\theta_i/\mu}\, d\theta_i \\ &= \left(\prod_{j=1}^{n} x_{ij}!\right)^{-1} \frac{1}{\mu} \int_0^\infty \theta_i^{\sum_{j=1}^n x_{ij}} e^{-\theta_i(n+1/\mu)}\, d\theta_i \\ &= C(\mathbf{x}_i)\mu^{-1}\left(n+\frac{1}{\mu}\right)^{-\sum_{j=1}^n x_{ij}-1} \int_0^\infty \frac{\beta\,(\beta\theta_i)^{\alpha-1} e^{-\beta\theta_i}}{\Gamma(\alpha)}\, d\theta_i, \end{aligned}$$

where $C(\mathbf{x}_i)$ may be expressed in combinatorial notation as

$$C(\mathbf{x}_i) = \binom{\sum_{j=1}^n x_{ij}}{x_{i1}\ x_{i2}\ \cdots\ x_{in}};$$

$$\beta = n + \frac{1}{\mu},$$

and

$$\alpha = \sum_{j=1}^{n} x_{ij} + 1.$$

The integral is that of a gamma density with parameters α and $1/\beta$ and therefore equals 1, and so

$$f(\mathbf{x}_i) = C(\mathbf{x}_i)\mu^{-1}\left(n+\frac{1}{\mu}\right)^{-\sum_{j=1}^n x_{ij}-1}.$$

Substitution into (16.66) yields

$$L(\mu) \propto \mu^{-r}\left(n+\frac{1}{\mu}\right)^{-\sum_{i=1}^r \sum_{j=1}^n x_{ij}-r}.$$

Thus

$$l(\mu) = \ln L(\mu) = -r\ln\mu - \left(r + \sum_{i=1}^{r}\sum_{j=1}^{n} x_{ij}\right)\ln\left(n+\frac{1}{\mu}\right) + c,$$

where c is a constant which does not depend on μ. Differentiating yields

$$l'(\mu) = -\frac{r}{\mu} - \frac{r + \sum_{i=1}^{r}\sum_{j=1}^{n} x_{ij}}{n + \frac{1}{\mu}} \left(-\frac{1}{\mu^2}\right).$$

The maximum likelihood estimator $\hat{\mu}$ of μ is found by setting $l'(\hat{\mu}) = 0$, which yields

$$\frac{r}{\hat{\mu}} = \frac{r + \sum_{i=1}^{r}\sum_{j=1}^{n} x_{ij}}{\hat{\mu}(\hat{\mu}n + 1)}$$

and so

$$\hat{\mu}n + 1 = 1 + \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{n} x_{ij}$$

or

$$\hat{\mu} = \frac{1}{nr}\sum_{i=1}^{r}\sum_{j=1}^{n} x_{ij}.$$

But this is the same as the nonparametric estimate obtained in Example 16.34. An explanation is in order. We have $\mu(\theta_i) = \theta_i$ by the Poisson assumption and so $\mathrm{E}[\mu(\Theta_i)] = \mathrm{E}(\Theta_i)$, which is the same μ as was used in the exponential distribution $\pi(\theta_i)$.

Furthermore, $v(\theta_i) = \theta_i$ as well (by the Poisson assumption), and so $v = \mathrm{E}[v(\Theta_i)] = \mu$. Also, $a = \mathrm{Var}[\mu(\Theta_i)] = \mathrm{Var}(\Theta_i) = \mu^2$ by the exponential assumption for $\pi(\theta_i)$. Thus the maximum likelihood estimators of v and a are $\hat{\mu}$ and $\hat{\mu}^2$ by the invariance of maximum likelihood estimation under a parameter transformation. Similarly, the maximum likelihood estimators of $k = v/a$, the credibility factor Z, and the credibility premium $Z\bar{X}_i + (1-Z)\mu$ are $\hat{k} = \hat{\mu}^{-1} = \bar{X}^{-1}$, $\hat{Z} = n/(n + \hat{\mu}^{-1})$, and $\hat{Z}\bar{X}_i + (1-\hat{Z})\hat{\mu}$, respectively. We mention also that credibility is exact in this model so that the Bayesian premium is equal to the credibility premium. □

Example 16.41 *Suppose that $n_i = n$ for all i and $m_{ij} = 1$. Assume that $X_{ij}|\Theta_i \sim N(\Theta_i, v)$,*

$$f_{X_{ij}|\Theta}(x_{ij}|\theta_i) = (2\pi v)^{-1/2} \exp\left[-\frac{1}{2v}(x_{ij} - \theta_i)^2\right], \quad -\infty < x_{ij} < \infty,$$

and $\Theta_i \sim N(\mu, a)$, so that

$$\pi(\theta_i) = (2\pi a)^{-1/2} \exp\left[-\frac{1}{2a}(\theta_i - \mu)^2\right], \quad -\infty < \theta_i < \infty.$$

Determine the maximum likelihood estimators of the parameters.

We have $\mu(\theta_i) = \theta_i$ and $v(\theta_i) = v$. Thus $\mu = \mathrm{E}[\mu(\Theta_i)]$, $v = \mathrm{E}[v(\Theta_i)]$, and $a = \mathrm{Var}[\mu(\Theta_i)]$, consistent with previous use of μ, v, and a. We shall now

derive maximum likelihood estimators of μ, v, and a. To begin with, consider $\bar{X}_i = n^{-1}\sum_{j=1}^{n} X_{ij}$. Conditional on Θ_i, the X_{ij} are independent $N(\Theta_i, v)$ random variables, implying that $\bar{X}_i|\Theta_i \sim N(\Theta_i, v/n)$. Because $\Theta_i \sim N(\mu, a)$, it follows from Example 4.30 that unconditionally $\bar{X}_i \sim N(\mu, a + v/n)$. Hence the density of $\bar{X}_i$ is, with $w = a + v/n$,

$$f(\bar{x}_i) = (2\pi w)^{-1/2} \exp\left[-\frac{1}{2w}(\bar{x}_i - \mu)^2\right], \quad -\infty < \bar{x}_i < \infty.$$

On the other hand, by conditioning on Θ_i, we have

$$\begin{aligned} f(\bar{x}_i) &= \int_{-\infty}^{\infty} (2\pi v/n)^{-1/2} \exp\left[-\frac{n}{2v}(\bar{x}_i - \theta_i)^2\right] \\ &\quad \times (2\pi a)^{-1/2} \exp\left[-\frac{1}{2a}(\theta_i - \mu)^2\right] d\theta_i. \end{aligned}$$

Ignoring terms not involving μ, v, or a, this means that $f(\bar{x}_i)$ is proportional to

$$v^{-1/2}a^{-1/2} \int_{-\infty}^{\infty} \exp\left[-\frac{n}{2v}(\bar{x}_i - \theta_i)^2 - \frac{1}{2a}(\theta_i - \mu)^2\right] d\theta_i.$$

Now (16.65) yields

$$\begin{aligned} f(\mathbf{x}_i) &= \int_{-\infty}^{\infty} \left\{\prod_{j=1}^{n} (2\pi v)^{-1/2} \exp\left[-\frac{1}{2v}(x_{ij} - \theta_i)^2\right]\right\} (2\pi a)^{-1/2} \\ &\quad \times \exp\left[-\frac{1}{2a}(\theta_i - \mu)^2\right] d\theta_i, \end{aligned}$$

which is proportional to

$$v^{-n/2}a^{-1/2} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2v}\sum_{j=1}^{n}(x_{ij} - \theta_i)^2 - \frac{1}{2a}(\theta_i - \mu)^2\right] d\theta_i.$$

Now use the identity (16.10) restated as

$$\sum_{j=1}^{n}(x_{ij} - \theta_i)^2 = \sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2 + n(\bar{x}_i - \theta_i)^2,$$

which means that $f(\mathbf{x}_i)$ is proportional to

$$v^{-n/2}a^{-1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2v}\left[\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2 + n(\bar{x}_i - \theta_i)^2\right] - \frac{1}{2a}(\theta_i - \mu)^2\right\} d\theta_i,$$

itself proportional to

$$v^{-(n-1)/2}\exp\left[-\frac{1}{2v}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\right]f(\bar{x}_i)$$

using the second expression for the density $f(\bar{x}_i)$ of $\bar{X}_i$ given above. Then (16.66) yields

$$L \propto v^{-r(n-1)/2}\exp\left[-\frac{1}{2v}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\right]\prod_{i=1}^{r}f(\bar{x}_i).$$

Let us now invoke the invariance of maximum likelihood estimators under a parameter transformation and use μ, v, and $w = a + v/n$ rather than μ, v, and a. This means that

$$L \propto L_1(v)L_2(\mu, w),$$

where

$$L_1(v) = v^{-r(n-1)/2}\exp\left[-\frac{1}{2v}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\right]$$

and

$$L_2(\mu, w) = \prod_{i=1}^{r}f(\bar{x}_i) = \prod_{i=1}^{r}\left\{(2\pi w)^{-1/2}\exp\left[-\frac{1}{2w}(\bar{x}_i-\mu)^2\right]\right\}.$$

The maximum likelihood estimator $\hat{v}$ of v can be found by maximizing $L_1(v)$ alone and the mle $(\hat{\mu}, \hat{w})$ of (μ, w) can be found by maximizing $L_2(\mu, w)$. Taking logarithms, we obtain

$$\begin{aligned} l_1(v) &= -\frac{r(n-1)}{2}\ln v - \frac{1}{2v}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2, \\ l_1'(v) &= -\frac{r(n-1)}{2v} + \frac{1}{2v^2}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2, \end{aligned}$$

and with $l'(\hat{v}) = 0$ we have

$$\hat{v} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2}{r(n-1)}.$$

Because $L_2(\mu, w)$ is the usual normal likelihood, the mles are simply the empirical mean and variance. That is,

$$\hat{\mu} = \frac{1}{r}\sum_{i=1}^{r}\bar{X}_i = \frac{1}{nr}\sum_{i=1}^{r}\sum_{j=1}^{n}X_{ij} = \bar{X}$$

and

$$\hat{w} = \frac{1}{r}\sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2.$$

But $a = w - v/n$ and so the maximum likelihood estimator of a is

$$\hat{a} = \frac{1}{r}\sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2 - \frac{1}{rn(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2.$$

It is instructive to note that the maximum likelihood estimators $\hat{\mu}$ and $\hat{v}$ are exactly the nonparametric unbiased estimators in the Bühlmann model of Example 16.34. The maximum likelihood estimator $\hat{a}$ is almost the same as the nonparametric unbiased estimator, the only difference being the divisor r rather than $r-1$ in the first term. □

16.5.4 Notes and References

In this section a simple approach was employed to find parameter estimates. No attempt was made to find optimum estimators in the sense of minimum variance. A good deal of research has been done on this problem. See Goovaerts and Hoogstad [46] for more details and further references.

16.5.5 Exercises

16.60 Past claims data on a portfolio of policyholders are given in Table 16.8. Estimate the Bühlmann credibility premium for each of the three policyholders for year 4.

16.61 Past data on a portfolio of group policyholders are given in Table 16.9. Estimate the Bühlmann–Straub credibility premiums to be charged to each group in year 4.

16.62 For the situation in Exercise 16.9, estimate the Bühlmann credibility premium for the next year for the policyholder.

Table 16.8 Data for Exercise 16.60

	Year		
Policyholder	1	2	3
1	750	800	650
2	625	600	675
3	900	950	850

Table 16.9 Data for Exercise 16.61

	Policyholder	Year 1	2	3	4
Claims	1	–	20,000	25,000	–
No. in group		–	100	120	110
Claims	2	19,000	18,000	17,000	–
No. in group		90	75	70	60
Claims	3	26,000	30,000	35,000	–
No. in group		150	175	180	200

16.63 Consider the Bühlmann model in Example 16.34.

(a) Prove that $\text{Var}(X_{ij}) = a + v$.

(b) If $\{X_{ij} : \ i = 1, \ldots, r \text{ and } j = 1, \ldots, n\}$ are unconditionally independent for all i and j, argue that an unbiased estimator of $a + v$ is

$$\frac{1}{nr-1}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2.$$

(c) Prove the algebraic identity

$$\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2 + n\sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2.$$

(d) Show that, conditionally,

$$\text{E}\left[\frac{1}{nr-1}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2\right] = (v+a) - \frac{n-1}{nr-1}a.$$

(e) Comment on the implications of (b) and (d).

16.64 The distribution of automobile insurance policyholders by number of claims is given in Table 16.10.

Assuming a (conditional) Poisson distribution for the number of claims per policyholder, estimate the Bühlmann credibility premiums for the number of claims next year.

16.65 Suppose that, given Θ, $X_1, \ldots, X_n$ are independently geometrically distributed with pf

$$f_{X_j|\Theta}(x_j|\theta) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{x_j}, \quad x_j = 0, 1, \ldots .$$

Table 16.10 Data for Exercise 16.64

No. of claims	No. of insureds
0	2,500
1	250
2	30
3	5
4	2
Total	2,787

(a) Show that $\mu(\theta) = \theta$ and $v(\theta) = \theta(1+\theta)$.

(b) Prove that $a = v - \mu - \mu^2$.

(c) Rework Exercise 16.64 assuming a (conditional) geometric distribution.

16.66 Suppose that

$$\Pr(m_{ij}X_{ij} = t_{ij}|\Theta_i = \theta_i) = \frac{(m_{ij}\theta_i)^{t_{ij}}e^{-m_{ij}\theta_i}}{t_{ij}!}$$

and

$$\pi(\theta_i) = \frac{1}{\mu}e^{-\theta_i/\mu}, \quad \theta_i > 0.$$

Write down the equation satisfied by the maximum likelihood estimator $\hat{\mu}$ of μ for Bühlmann–Straub-type data.

16.67 (a) Prove the algebraic identity

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}m_{ij}(X_{ij}-\bar{X})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}m_{ij}(X_{ij}-\bar{X}_i)^2 + \sum_{i=1}^{r}m_i(\bar{X}_i-\bar{X})^2.$$

(b) Use part (a) and (16.61) to show that (16.63) may be expressed as

$$\hat{a} = m_*^{-1}\left[\frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}m_{ij}(X_{ij}-\bar{X})^2}{\sum_{i=1}^{r}n_i - 1} - \hat{v}\right]$$

where

$$m_* = \frac{\sum_{i=1}^{r}m_i\left(1-\frac{m_i}{m}\right)}{\sum_{i=1}^{r}n_i - 1}.$$

16.68 (*) A group of 340 insureds in a high-crime area submit the 210 theft claims in a one-year period as given in Table 16.11.

Table 16.11 Data for Exercise 16.68

Number of claims	Number of insureds
0	200
1	80
2	50
3	10

Each insured is assumed to have a Poisson distribution for the number of thefts, but the mean of such a distribution may vary from one insured to another. If a particular insured experienced two claims in the observation period, determine the Bühlmann credibility estimate for the number of claims for this insured in the next period.

17

Simulation

17.1 BASICS OF SIMULATION

Simulation has had an on-again, off-again history in actuarial practice. For example, in the 1970s, aggregate loss calculations were commonly done by simulation because the analytical methods available at the time were not adequate. However, the typical simulation often took a full day on the company's mainframe computer, a serious drag on resources. In the 1980s analytic methods such as Heckman–Meyers and the recursive formula were developed and were found to be significantly faster and more accurate. Today, desktop computers have sufficient power to run complex simulations that allow for the analysis of models not suitable for current analytic approaches.

In a similar vein, as investment vehicles become more complex, contracts have interest-sensitive components, and market fluctuations seem to be more pronounced, analysis of future cash flows must be done on a stochastic basis. In order to accommodate the complexities of the products and interest rate models, simulation has become the technique of choice.

In this chapter we will provide some illustrations of how simulation can solve problems such as those mentioned above. It is not our intention to cover the subject in great detail, but rather to give the reader an idea of how simulation can help. Study of simulation texts such as Herzog and Lord [53],

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

Ripley [110], and Ross [115] will provide many important additional insights. In addition, simulation can also be an aid in evaluating some of the statistical techniques covered in earlier chapters. This will also be covered here with an emphasis on the bootstrap method.

17.1.1 The simulation approach

The beauty of simulation is that once the model is created little additional creative thought is required.[1] The entire process can be summarized in four steps, where the goal is to determine values relating to the distribution of a random variable S.

1. Build a model for S which depends on random variables $X, Y, Z, \ldots$, where their distributions and any dependencies are known.
2. For $j = 1, \ldots, n$ generate pseudorandom values $x_j, y_j, z_j, \ldots$ and then compute s_j using the model from step 1.
3. The cdf of S may be approximated by $F_n(s)$, the empirical cdf based on the pseudorandom sample $s_1, \ldots, s_n$.
4. Compute quantities of interest, such as the mean, variance, percentiles, or probabilities, using the empirical cdf.

Two questions remain. First, what does it mean to generate a pseudorandom variable? Consider a random variable X with cdf $F_X(x)$. This is the real random variable produced by some phenomenon of interest. For example, it may be the result of the experiment "collect one automobile bodily injury medical payment at random and record its value." We assume that the cdf is known. For example, it may be the Pareto cdf, $F_X(x) = 1 - \left(\frac{1{,}000}{1{,}000+x}\right)^3$. Now consider a second random variable, X^*, resulting from some other process, but with the same Pareto distribution. A random sample from X^*, say $x_1^*, \ldots, x_n^*$, would be impossible to distinguish from one taken from X. That is, given the n numbers, we could not tell if they arose from automobile claims or something else. This means that, instead of learning about X by observing automobile claims, we could learn about it by observing X^*. Obtaining a random sample from a Pareto distribution is still probably difficult, so we have not yet accomplished much.

We can make some progress by making a concession. Let us accept as a replacement for a random sample from X^* a sequence of numbers $x_1^{**}, \ldots, x_n^{**}$ which is not a random sample at all, but simply a sequence of numbers which

[1] This is not entirely true. A great deal of creativity may be employed in designing an efficient simulation. The brute force approach used here will work; it just may take your computer longer to produce the answer.

may not be independent, or even random, but was generated by some known process that is related to the random variable X^*. Such a sequence is called a pseudorandom sequence because anyone who did not know its origin could not distinguish it from a random sample from X^* (and therefore from X). Such a sequence will be satisfactory for our purposes.

The field of developing processes for generating pseudorandom sequences of numbers has been well developed. One fact that makes it easier to do this is that it is sufficient to be able to generate such sequences for the uniform distribution on the interval $(0, 1)$. That is because, if U has the uniform$(0, 1)$ distribution, then $X = F_X^{-1}(U)$ will have $F_X(x)$ as its cdf. Therefore, we simply obtain uniform pseudorandom numbers $u_1^{**}, \ldots, u_n^{**}$ and then let $x_j^{**} = F_X^{-1}(u_j^{**})$. This is called the inversion method of generating random variates. Specific methods for particular distributions have been developed but will not be discussed here. There is a considerable literature on the best ways to generate pseudorandom uniform numbers and a variety of tests proposed to evaluate them. Readers are cautioned to ensure that one being used is a good one.

Example 17.1 *Generate* 10,000 *pseudo-Pareto* (*with* $\alpha = 3$, *and* $\theta = 1{,}000$) *variates and verify that they are indistinguishable from real Pareto observations.*

The pseudouniform values were obtained using the built-in generator supplied with a commercial programming language. The pseudo-Pareto values are calculated from

$$u^{**} = 1 - \left(\frac{1{,}000}{1{,}000 + x^{**}}\right)^3.$$

That is,

$$x^{**} = 1{,}000[(1 - u^{**})^{-1/3} - 1].$$

So, for example, if the first value generated is $u_1^{**} = 0.54246$, we have $x_1^{**} = 297.75$. This was repeated 10,000 times. The results are displayed in Table 17.1, where a chi-square goodness-of-fit test is conducted. The expected counts are calculated using the Pareto distribution with $\alpha = 3$ and $\theta = 1{,}000$. Because the parameters are known, there are nine degrees of freedom. At a significance level of 5% the critical value is 16.92, and we conclude that the pseudorandom sample could have been a random sample from this Pareto distribution. □

When the distribution function of X is continuous and strictly increasing, the equation $u = F_X(x)$ will have a unique solution for any u. In that case the inversion method reduces to solving the equation. In other cases some care must be taken. Suppose $F_X(x)$ jumps at $x = c$ so that $F_X(c-) = a$ and $F_X(c) = b > a$. If the uniform number is such that $a \le u < b$, the equation has no solution. In that situation choose c as the simulated value.

Table 17.1 Chi-square test of simulated Pareto observations

Interval	Observed	Expected	Chi square
0–100	2,519	2,486.85	0.42
100–250	2,348	2,393.15	0.85
250–500	2,196	2,157.04	0.70
500–750	1,071	1,097.07	0.62
750–1,000	635	615.89	0.59
1,000–1,500	589	610.00	0.72
1,500–2,500	409	406.76	0.01
2,500–5,000	192	186.94	0.14
5,000–10,000	36	38.78	0.20
10,000–	5	7.51	0.84
Total	10,000	10,000	5.10

Example 17.2 *Suppose*

$$F_X(x) = \begin{cases} 0.5x, & 0 \le x < 1, \\ 0.5 + 0.25x & 1 \le x \le 2. \end{cases}$$

Determine the simulated values of x resulting from the uniform numbers 0.3, 0.6, *and* 0.9.

In the first interval, the distribution function ranges from 0 to 0.5 and in the second interval from 0.75 to 1. With $u = 0.3$ in the first interval, solve $0.3 = 0.5x$ for $x = 0.6$. With the distribution function jumping from 0.5 to 0.75 at $x = 1$, any u in that interval will lead to a simulated value of 1, so for $u = 0.6$, the simulated value is $x = 1$. Note that $\Pr(0.5 \le U < 0.75) = 0.25$, so the value of $x = 1$ will be simulated 25% of the time, matching its true probability. Finally, with 0.9 in the second interval, solve $0.9 = 0.5 + 0.25x$ for $x = 1.6$. Figure 17.1 illustrates this process, showing how drawing vertical bars on the function makes the inversion obvious. □

It is also possible for the distribution function to be constant over some interval. In that case the equation $u = F_X(x)$ will have multiple solutions over that interval. Our convention (to be justified shortly) is to choose the largest possible value in the interval.

Example 17.3 *Suppose*

$$F_X(x) = \begin{cases} 0.5x, & 0 \le x < 1, \\ 0.5, & 1 \le x < 2, \\ 0.5x - 0.5 & 2 \le x < 3. \end{cases}$$

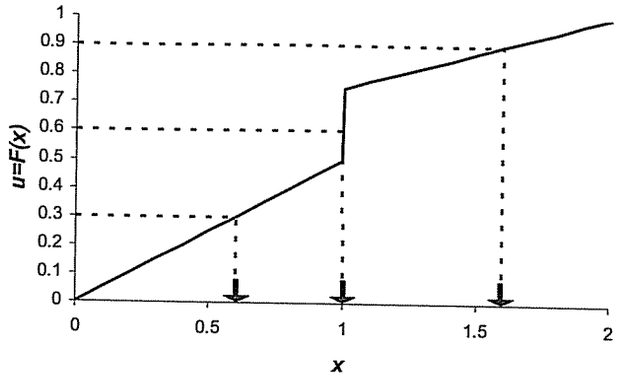


Fig. 17.1 Inversion of the distribution function for Example 17.2

Determine the simulated values of x resulting from the uniform numbers 0.3, 0.5, *and* 0.9.

The first interval covers values of the distribution function from 0 to 0.5 and the final interval covers the range 0.5 to 1. For $u = 0.3$, use the first interval and solve $0.3 = 0.5x$ for $x = 0.6$. The function is constant at 0.5 from 1 to 2 and so for $u = 0.5$ choose the largest value, $x = 2$. For $u = 0.9$ use the final interval and solve $0.9 = 0.5x - 0.5$ for $x = 2.8$. □

Discrete distributions have both features. The distribution function jumps at the possible values of the variable and is constant in between.

Example 17.4 *Simulate values from a binomial distribution with $m = 4$ and $q = 0.5$ using the uniform numbers* 0.3, 0.6875, *and* 0.95.

The distribution function is

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 0.0625, & 0 \le x < 1, \\ 0.3125, & 1 \le x < 2, \\ 0.6875, & 2 \le x < 3, \\ 0.9375, & 3 \le x < 4, \\ 1, & x \ge 4. \end{cases}$$

For $u = 0.3$, the function is jumping at $x = 1$. For $u = 0.6875$, the function is constant from 2 to 3 (as the limiting value of the interval) and so $x = 3$. For $u = 0.95$ the function is jumping at $x = 4$. It is usually easier to present the simulation algorithm using a table based on the distribution function. Then

a simple table lookup function (such as the VLOOKUP function in Excel®) can be used to obtain simulated values. For this example, the table is as follows.

For u in this range,	the simulated value is
$0 \le u < 0.0625$,	0
$0.0625 \le u < 0.3125$,	1
$0.3125 \le u < 0.6875$,	2
$0.6875 \le u < 0.9375$,	3
$0.9375 \le u < 1$,	4

Many random number generators can produce a value of 0 but not a value of 1 (though some produce neither one). This is the motivation for choosing the largest value in an interval where the cdf is constant. □

The second question is: What value of n should be used? We know that any consistent estimator will be arbitrarily close to the true value with high probability as the sample size is increased. In particular, empirical estimators have this attribute. With a little effort we should be able to determine the value of n that will get us as close as we want with a specified probability. Often, the central limit theorem will help, as in the following example.

Example 17.5 (Example 17.1 continued) *Use simulation to estimate the mean, $F_X(1{,}000)$, and $\pi_{0.9}$, the 90th percentile of the Pareto distribution with $\alpha = 3$ and $\theta = 1{,}000$. In each case, stop the simulations when you are 95% confident that the answer is within ±1% of the true value.*

In this example we know the values. Here, $\mu = 500$, $F_X(1{,}000) = 0.875$, and $\pi_{0.9} = 1{,}154.43$. For instructional purposes we will behave as if we do not know these values.

The empirical estimate of μ is $\bar{x}$. The central limit theorem tells us that for a sample of size n

$$\begin{aligned} 0.95 &= \Pr(0.99\mu \le \bar{X}_n \le 1.01\mu) \\ &= \Pr\left(-\frac{0.01\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le \frac{0.01\mu}{\sigma/\sqrt{n}}\right) \\ &\doteq \Pr\left(-\frac{0.01\mu}{\sigma/\sqrt{n}} \le Z \le \frac{0.01\mu}{\sigma/\sqrt{n}}\right), \end{aligned}$$

where Z has the standard normal distribution. Our goal is achieved when

$$\frac{0.01\mu}{\sigma/\sqrt{n}} = 1.96, \tag{17.1}$$

which means $n = 38{,}416(\sigma/\mu)^2$. Because we do not know the values of σ and μ, we estimate them with the sample standard deviation and mean. The estimates improve with n, so our stopping rule is to cease simulating when

$$n \ge \frac{38{,}416 s^2}{\bar{x}^2}.$$

For a particular simulation conducted by the authors, the criterion was met when $n = 106{,}934$, at which point $\bar{x} = 501.15$, a relative error of 0.23%, well within our goal.

We now turn to the estimation of $F_X(1{,}000)$. The empirical estimator is the sample proportion below 1,000, say P_n/n, where P_n is the number below 1,000 after n simulations. The central limit theorem tells us that P_n/n is approximately normal with mean $F_X(1{,}000)$ and variance $F_X(1{,}000)[1 - F_X(1{,}000)]/n$. Arguing as above, the requirement will be met when

$$n \ge 38{,}416\frac{n - P_n}{P_n}.$$

For our simulation, the criterion was met at $n = 5{,}548$, at which point the estimate was $4{,}848/5{,}548 = 0.87383$, which has a relative error of 0.13%.

Finally, for $\pi_{0.9}$, begin with

$$0.95 = \Pr(Y_a \le \pi_{0.9} \le Y_b),$$

where $Y_1 \le Y_2 \le \cdots \le Y_n$ are the order statistics from the simulated sample, $a = [0.9n - 1.96\sqrt{0.9(0.1)n}]$, $b = [0.9n + 1.96\sqrt{0.9(0.1)n}] + 1$ (where $[\cdot]$ is the greatest integer function and the 1 is added for conservatism), and the process terminates when both

$$\hat{\pi}_{0.9} - Y_a \le 0.01\hat{\pi}_{0.9}$$

and

$$Y_b - \hat{\pi}_{0.9} \le 0.01\hat{\pi}_{0.9}.$$

For the example, this occurred when $n = 126{,}364$, and the estimated 90th percentile is 1,153.97, with a relative error of 0.04%. □

17.1.2 Exercises

17.1 Use the inversion method to simulate three values from the Poisson(3) distribution. Use 0.1247, 0.9321, and 0.6873 for the uniform random numbers.

17.2 Use the uniform random numbers 0.2, 0.5, and 0.7 to simulate values from

$$f_X(x) = \begin{cases} 0.25, & 0 \le x \le 2, \\ 0.1, & 4 \le x \le 9, \\ 0, & \text{otherwise.} \end{cases}$$

17.3 Demonstrate that $0.95 = \Pr(Y_a \leq \pi_{0.9} \leq Y_b)$ for Y_a and Y_b as defined in Example 17.5.

17.4 You are simulating observations from an exponential distribution with $\theta = 100$. How many simulations are needed to be 90% certain of being within 2% of each of the mean and the probability of being below 200? Conduct the required number of simulations and note if the 2% goal has been reached.

17.5 Simulate 1,000 observations from a gamma distribution with $\alpha = 2$ and $\theta = 500$. Perform the chi-square goodness-of-fit and Kolmogorov–Smirnov tests to see if the simulated values were actually from that distribution.

17.6 (*) To estimate $\mathrm{E}(X)$, you have simulated five observations from the random variable X. The values are 1, 2, 3, 4, and 5. Your goal is to have the standard deviation of the estimate of $\mathrm{E}(X)$ be less than 0.05. Estimate the total number of simulations needed.

17.2 EXAMPLES OF SIMULATION IN ACTUARIAL MODELING

17.2.1 Aggregate loss calculations

The analytic methods presented in Chapter 6 have two features in common. First, they are exact up to the level of the approximation introduced. For recursion and the FFT, that involves replacing the true severity distribution with an arithmetized approximation. For Heckman–Meyers a histogram approximation is required. Furthermore, Heckman–Meyers requires a numerical integration. In each case, the errors can be reduced to near zero by increasing the number of points used. Second, both recursion and inversion assume that aggregate claims can be written as $S = X_1 + \cdots + X_N$ with $N, X_1, X_2, \ldots$ independent and the X_js identically distributed.

There is no need to be concerned about the first feature because the approximation error can be made as small as desired. However, the second restriction may prevent the model from reflecting reality. In this section we indicate some common ways in which the independence or identical distribution assumptions may fail to hold and then demonstrate how simulation can lead to a solution. When the X_js are i.i.d. it does not matter how we go about labeling the losses—that is, which loss is called X_1, which one X_2, and so on. With the assumption removed, the labels become important. Because S is the aggregate loss for one year, time is a factor. One way of identifying the losses is to let X_1 be the first loss, X_2 be the second loss, and so on. Then let T_j be the random variable that records the time of the jth loss. Without going into much detail about the claims-paying process, we do want to note that T_j may be the time at which the loss occurred, the time it was reported, or the time payment was made. In the latter two cases it may be that $T_j > 1$, which occurs when the report of the loss or the payment of the claim takes place at a time subsequent to the end of the time period of the coverage, usually one year. If the timing of the losses is important, we will need to know the joint distribution of $(T_1, T_2, \ldots, X_1, X_2, \ldots)$.

17.2.2 Examples of lack of independence or identical distributions

There are two common ways to have the assumption fail to hold. One is through accounting for time (and in particular the time value of money) and the other is through coverage modifications. The latter may have a time factor as well. The following examples provide some illustrations.

Example 17.6 (Time value of loss payments) *Suppose the quantity of interest, S, is the present value of all payments made in respect of a policy issued today and covering loss events that occur in the next year. Develop a model for S.*

Let T_j be the time of the payment of the jth loss. While T_j records the time of the payment, the subscripts are selected in order of the loss events. Let $T_j = C_j + L_j$ where C_j is the time of the event and L_j is the time from occurrence to payment. Assume they are independent and the L_js are independent of each other. Let the time between events, $C_j - C_{j-1}$ (where $C_0 = 0$), be i.i.d. with an exponential distribution with mean 0.2 years.

Let X_j be the amount paid at time T_j on the loss that occurred at time C_j. Assume that X_j and C_j are independent (the amount of the claim does not depend on when in the year it occurred) but X_j and L_j are positively correlated (a specific distributional model will be specified when the example is continued). This is reasonable because the more expensive losses may take longer to settle.

Finally, let V_t be a random variable that represents the value, which, if invested today, will accumulate to 1 in t years. It is independent of all X_j, C_j, and L_j. But clearly, for $s \neq t$, V_s and V_t are dependent.

We then have

$$S = \sum_{j=1}^{N} X_j V_{T_j}$$

where $N = \max_{C_j<1}\{j\}$. The various dependencies were established in the development of the random variables. □

Example 17.7 (Out-of-pocket maximum) *Suppose there is a deductible, d, on individual losses. However, in the course of a year, the policyholder will pay no more than u. Develop a model for the insurer's aggregate payments.*

Let X_j be the amount of the jth loss. Here the assignment of j does not matter. Let $W_j = X_j \wedge d$ be the amount paid by the policyholder due to the deductible and let $Y_j = X_j - W_j$ be the amount paid by the insurer. Then $R = W_1 + \cdots + W_N$ is the total amount paid by the policyholder prior to

imposing the out-of-pocket maximum. Then the amount actually paid by the policyholder is $R_u = R \wedge u$. Let $S = X_1 + \cdots + X_N$ be the total losses, and then the aggregate amount paid by the insurer is $T = S - R_u$. Note that the distributions of S and R_u are based on i.i.d. severity distributions. The analytic methods described earlier can be used to obtain their distributions. But because they are dependent, their individual distributions cannot be combined to produce the distribution of T. There is also no way to write T as a random sum of i.i.d. variables. At the beginning of the year, it appears that T will be the sum of i.i.d. Y_js, but at some point the Y_js may be replaced by X_js as the out-of-pocket maximum is reached. □

17.2.3 Simulation analysis of the two examples

We now complete the two examples using the simulation approach. The models have been selected arbitrarily, but we should assume they were determined by a careful estimation process using the techniques presented earlier in this text.

Example 17.8 (Example 17.6 continued) *The model is completed with the following specifications. The amount of a payment (X_j) has the Pareto distribution with parameters $\alpha = 3$ and $\theta = 1{,}000$. The time from the occurrence of a claim to its payment (L_j) has a Weibull distribution with $\tau = 1.5$ and $\theta = \ln(X_j)/6$. This models the dependence by having the scale parameter depend on the size of the loss. The discount factor will be modeled by assuming that, for $t > s$, $[\ln(V_s/V_t)]/(t-s)$ has a normal distribution with mean 0.06 and variance $0.0004(t-s)$. We do not need to specify a model for the number of losses. Instead, we use the model given earlier for the time between losses. Use simulation to determine the expected present value of aggregate payments.*

The mechanics of a single simulation will be done in detail, and that should indicate how the process is to be done. Begin by generating i.i.d. exponential interloss times until their sum exceeds 1 (in order to obtain one year's worth of claims). The individual variates are generated from pseudouniform numbers using

$$u = 1 - e^{-5x},$$

which yields

$$x = -0.2\ln(1-u).$$

For the first simulation, the uniform pseudorandom numbers and the corresponding x values are (0.25373, 0.0585), (0.46750, 0.1260), (0.23709, 0.0541), (0.75780, 0.2836), and (0.96642, 0.6788). At this point the simulated xs total 1.2010 and therefore there are four loss events, occurring at times $c_1 = 0.0585$, $c_2 = 0.1845$, $c_3 = 0.2386$, and $c_4 = 0.5222$.

The four loss amounts are found from inverting the Pareto cdf. That is,

$$x = 1{,}000[(1-u)^{-1/3} - 1].$$

The four pseudouniform numbers are 0.71786, 0.47779, 0.61084, and 0.68579. This produces the four losses $x_1 = 524.68$, $x_2 = 241.80$, $x_3 = 369.70$, and $x_4 = 470.93$.

The times from occurrence to payment have a Weibull distribution. The equation to solve is

$$u = 1 - e^{-[6l/\ln(x)]^{1.5}},$$

where x is the loss. Solving for the lag time l yields

$$l = \tfrac{1}{6}\ln(x)[-\ln(1-u)]^{2/3}.$$

For the first lag we have $u = 0.23376$ and so

$$l_1 = \tfrac{1}{6}\ln(524.68)[-\ln 0.76624]^{2/3} = 0.4320.$$

Similarly, with the next three values of u being 0.85799, 0.12951, and 0.72085, we have $l_2 = 1.4286$, $l_3 = 0.2640$, and $l_4 = 1.2068$. The payment times of the four losses are the sum of c_j and l_j, namely $t_1 = 0.4905$, $t_2 = 1.6131$, $t_3 = 0.5026$, and $t_4 = 1.7290$.

Finally, we generate the discount factors. They must be generated in order of increasing t_j so we first obtain $v_{0.4905}$. We begin with a normal variate with mean 0.06 and variance $0.0004(0.4905) = 0.0001962$. Using inversion, the simulated value is $0.0592 = [\ln(1/v_{0.4905})]/0.4905$ and so $v_{0.4905} = 0.9714$. Note that for the first value we have $s = 0$, and $v_0 = 1$. For the second value we require a normal variate with mean 0.06 and variance $(0.5026 - 0.4905)(0.0004) = 0.00000484$. The simulated value is

$$0.0604 = \frac{\ln(0.9714/v_{0.5026})}{0.0121} \text{ for } v_{0.5026} = 0.9707.$$

For the next two payments, we have

$$0.0768 = \frac{\ln(0.9707/v_{1.6131})}{1.1105} \text{ for } v_{1.6131} = 0.8913,$$

$$0.0628 = \frac{\ln(0.8913/v_{1.7290})}{0.1159} \text{ for } v_{1.7290} = 0.8848.$$

We are now ready to determine the first simulated value of the aggregate present value. It is

$$\begin{aligned} s_1 &= 524.68(0.9714) + 241.80(0.8913) + 369.70(0.9707) + 470.93(0.8848) \\ &= 1{,}500.74. \end{aligned}$$

The process was then repeated until there was 95% confidence that the estimated mean was within 1% of the true mean. This took 26,944 simulations, producing a sample mean of 2,299.16. □

Example 17.9 (Example 17.7 continued) *For this example, set the deductible d at 250 and the out-of-pocket maximum at $u = 1{,}000$. Assume that the*

Table 17.2 Negative binomial cumulative probabilities

n	$F_N(n)$	n	$F_N(n)$
0	0.03704	8	0.76589
1	0.11111	9	0.81888
2	0.20988	10	0.86127
3	0.31962	11	0.89467
4	0.42936	12	0.92064
5	0.53178	13	0.94062
6	0.62282	14	0.95585
7	0.70086	15	0.96735

number of losses has the negative binomial distribution with $r = 3$ and $\beta = 2$. Further assume that individual losses have the Weibull distribution with $\tau = 2$ and $\theta = 600$. Determine the 95th percentile of the insurer's losses.

In order to simulate the negative binomial claim counts, we require the cdf of the negative binomial distribution. There is no closed form, but a table can be constructed, and one appears here as Table 17.2. The number of losses for the year is generated by obtaining one pseudouniform value—for example, $u = 0.47515$—and then determining the smallest entry in the table that is larger than 0.47515. The simulated value appears to its left. In this case our first simulation produced $n = 5$ losses.

The amounts of the five losses are obtained from the Weibull distribution. Inversion of the cdf produces

$$x = 600[-\ln(1-u)]^{1/2}.$$

The five simulated values are 544.04, 453.67, 217.87, 681.98, and 449.83. The total loss is 2,347.39. The policyholder pays 250.00+250.00+217.87+250.00+250.00 = 1,217.87, but the out-of-pocket maximum limits this to 1,000. Thus our first simulated value has the insurer paying 1,347.39.

The goal was set to be 95% confident that the estimated 95th percentile would be within 2% of the true value. This required 11,476 simulations, producing an estimated 95th percentile of 6,668.18. □

17.2.4 Statistical analyses

Simulation can help in a variety of ways when analyzing data. Two will be discussed here, both of which have to do with evaluating a statistical procedure. The first is the determination of the p-value (or critical value) for a hypothesis test. The second is to evaluate the mean-squared error of an estimator. We begin with the hypothesis testing situation.

Example 17.10 *It is conjectured that losses have a lognormal distribution. One hundred observations have been collected and the Kolmogorov–Smirnov test statistic is* 0.06272. *Determine the p-value for this test, first with the null hypothesis being that the distribution is lognormal with $\mu = 7$ and $\sigma = 1$ and then with the parameters unspecified.*

For the null hypothesis with each parameter specified, one simulation involves first simulating 100 lognormal observations from the specified lognormal distribution. Then the Kolmogorov–Smirnov test statistic is calculated. The estimated p-value is the proportion of simulations for which the test statistic exceeds 0.06272. After 1000 simulations, the estimate of the p-value is 0.836.

With the parameters unspecified, it is not clear which lognormal distribution should be used. It turned out that for the observations actually collected $\hat{\mu} = 7.2201$ and $\hat{\sigma} = 0.80893$. These were used as the basis for each simulation. The only change is that after the simulated observations have been obtained, the results are compared to a lognormal distribution with parameters estimated (by maximum likelihood) from the simulated data set. For 1,000 simulations, the test statistic exceeded 0.06272 491 times, for an estimated p-value of 0.491.

As indicated in Section 13.4.1, not specifying the parameters makes a considerable difference in the interpretation of the test statistic. □

When testing hypotheses, p-values and significance levels are calculated assuming the null hypothesis to be true. In other situations, there is no known population distribution from which to simulate. For such situations, a technique called the bootstrap (see [33] for thorough coverage of this subject) may help. The key is to use the empirical distribution from the data as the population from which to simulate values. Theoretical arguments show that at least asymptotically the bootstrap estimate will converge to the true value. This is reasonable because as the sample size increases the empirical distribution becomes more and more like the true distribution. The following example shows how the bootstrap works and also indicates that, at least in the case illustrated, it gives a reasonable answer.

Example 17.11 *A sample (with replacement) of size* 3 *from a population produced the values* 2, 3, *and* 7. *Determine the bootstrap estimate of the mean-squared error of the sample mean as an estimator of the population mean.*

The bootstrap approach assumes that the population places probability $\frac{1}{3}$ on each of the three values 2, 3, and 7. The mean of this distribution is 4. From this population there are 27 samples of size 3 that might be drawn. Sample means can be 2 (sample values 2, 2, 2, with probability $\frac{1}{27}$), $\frac{7}{3}$ (sample values 2, 2, 3, 2, 3, 2, and 3, 2, 2, with probability $\frac{3}{27}$), and so on, up to 7 with

probability $\frac{1}{27}$. The mean-squared error is

$$\frac{(2-4)^2(1/27)+(\frac{7}{3}-4)^2(3/27)+\cdots+(7-4)^2}{27}=\frac{14}{9}.$$

The usual approach is to note that the sample mean is unbiased and therefore

$$\text{MSE}(\bar{X})=\text{Var}(\bar{X})=\sigma^2/n.$$

With the variance unknown, a reasonable choice is to use the sample variance. With a denominator of n, for this example, the estimated mean-squared error is

$$\frac{\frac{1}{3}[(2-4)^2+(3-4)^2+(7-4)^2]}{3}=\frac{14}{9},$$

the same as the bootstrap estimate. □

In many situations, determination of the mean-squared error is not so easy, and then the bootstrap becomes an extremely useful tool. While simulation was not needed for the example, note that an original sample size of 3 led to 27 possible bootstrap values. Once the sample size gets beyond 6, it becomes impractical to enumerate all the cases. In that case, simulating observations from the empirical distribution becomes the only feasible choice.

Example 17.12 *In Example* 11.3 *an empirical model for time to death was obtained. The empirical probabilities are* 0.0333, 0.0744, 0.0343, 0.0660, 0.0344, *and* 0.0361 *that death is at times* 0.8, 2.9, 3.1, 4.0, 4.1, *and* 4.8 *respectively. The remaining* 0.7215 *probability is that the person will be alive five years from now. The expected present value for a five-year term insurance policy that pays* 1,000 *at the moment of death is estimated as*

$$1000(0.0333v^{0.8}+\cdots+0.0361v^{4.8})=223.01,$$

where $v=1.07^{-1}$. *Simulate* 10,000 *bootstrap samples to estimate the mean-squared error of this estimator.*

A method for conducting a bootstrap simulation with the Kaplan–Meier estimate is given by Efron [31]. Rather than simulate from the empirical distribution (as given by the Kaplan–Meier estimate), simulate from the original sample. In this example, that means assigning probability $\frac{1}{40}$ to each of the original observations. Then each bootstrap observation is a left-truncation point along with the accompanying censored or uncensored value. After 40 such observations are recorded, the Kaplan–Meier estimate is constructed from the bootstrap sample and then the quantity of interest computed. This is relatively easy because the bootstrap estimate can place probability only at the six original points. Ten thousand simulations were quickly done. The mean was 222.05 and the mean-squared error was 4,119. Efron also noted that the bootstrap estimate of the variance of $\hat{S}(t)$ is asymptotically equal to Greenwood's estimate, thus giving credence to both methods. □

17.2.5 Exercises

17.7 (*) Insurance for a city's snow removal costs covers four winter months. There is a deductible of 10,000 per month. Monthly costs are independent and normally distributed with $\mu=15{,}000$ and $\sigma=2{,}000$. Monthly costs are simulated using the inversion method. For one simulation of a year's payments the four uniform pseudorandom numbers are 0.5398, 0.1151, 0.0013, and 0.7881. Calculate the insurer's cost for this simulated year.

17.8 (*) After one period, the price of a stock is X times its price at the beginning of the period, where X has a lognormal distribution with $\mu=0.01$ and $\sigma=0.02$. The price at time 0 is 100. The inversion method is used to simulate price movements. The pseudouniform random numbers are 0.1587 and 0.9332 for periods 1 and 2. Determine the simulated prices at the end of each of the first two periods.

17.9 (*) You have insured 100 people, each age 70. Each person has probability 0.03318 of dying in the next year and the deaths are independent. Therefore, the number of deaths has a binomial distribution with $m=100$ and $q=0.03318$. Use the inversion method to determine the simulated number of deaths in the next year based on $u=0.18$.

17.10 (*) For a surplus process, claims occur according to a Poisson process at the rate of two per year. Thus the time between claims has the exponential distribution with $\theta=2$. Claims have a Pareto distribution with $\alpha=2$ and $\theta=1{,}000$. The initial surplus is 2,000 and premiums are collected at a rate of 2,200. Ruin occurs any time the surplus is negative, at which time no further premiums are collected or claims paid. All simulations are done with the inversion method. For the time between claims, use 0.83, 0.54, 0.48, and 0.14 as the pseudorandom numbers. For claim amounts use 0.89, 0.36, 0.70, and 0.61. Determine the surplus at time 1.

17.11 (*) You are given a random sample of size 2 from some distribution. The values are 1 and 3. You plan to estimate the population variance with the estimator $[(X_1-\bar{X})^2+(X_2-\bar{X})^2]/2$. Determine the bootstrap estimate of the mean-squared error of this estimator.

17.12 A sample of three items from the uniform(0,10) distribution produced the following values: 2, 4, and 7.

(a) Calculate the Kolmogorov–Smirnov test statistic for the null hypothesis that the data came from the uniform(0,10) distribution.

(b) Simulate 10,000 samples of size 3 from the uniform(0,10) distribution and compute the Kolmogorov–Smirnov test statistic for each. The proportion of times the value equals or exceeds your answer to part (a) is an estimate of the p-value.

17.13 A sample of three items from the uniform$(0, \theta)$ distribution produced the following values: 2, 4, and 7. Consider the estimator of θ,

$$\hat{\theta} = \tfrac{4}{3}\max(x_1, x_2, x_3).$$

From example 9.15 the mean-squared error of this unbiased estimator was shown to be $\theta^2/15$.

(a) Estimate the mean-squared error by replacing θ with its estimate.

(b) Obtain the bootstrap estimate of the variance of the estimator. (It is not possible to use the bootstrap to estimate the mean-squared error because you cannot obtain the true value of θ from the empirical distribution, but you can obtain the expected value of the estimator.)

Appendix A An inventory of continuous distributions

A.1 INTRODUCTION

Descriptions of the models are given below. First a few mathematical preliminaries are presented that indicate how the various quantities can be computed.

The incomplete gamma function[1] is given by

$$\Gamma(\alpha; x) = \frac{1}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-t}\, dt, \quad \alpha > 0,\ x > 0$$

$$\text{with } \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt, \quad \alpha > 0.$$

[1]Some references, such as [3], denote this integral $P(\alpha, x)$ and define $\Gamma(\alpha, x) = \int_x^\infty t^{\alpha-1} e^{-t}\, dt$. Note that this definition does not normalize by dividing by $\Gamma(\alpha)$. When using software to evaluate the incomplete gamma function, be sure to note how it is defined.

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

Also, define

$$G(\alpha;x)=\int_x^\infty t^{\alpha-1}e^{-t}\,dt,\quad x>0.$$

At times we will need this integral for nonpositive values of α. Integration by parts produces the relationship

$$G(\alpha;x)=-\frac{x^\alpha e^{-x}}{\alpha}+\frac{1}{\alpha}G(\alpha+1;x).$$

This can be repeated until the first argument of G is $\alpha+k$, a positive number. Then it can be evaluated from

$$G(\alpha+k;x)=\Gamma(\alpha+k)[1-\Gamma(\alpha+k;x)].$$

However, if α is a negative integer or zero, the value of $G(0;x)$ is needed. It is

$$G(0;x)=\int_x^\infty t^{-1}e^{-t}\,dt=E_1(x),$$

which is called the **exponential integral**. A series expansion for this integral is

$$E_1(x)=-0.57721566490153-\ln x-\sum_{n=1}^{\infty}\frac{(-1)^n x^n}{n(n!)}.$$

When α is a positive integer, the incomplete gamma function can be evaluated exactly as given in the following theorem.

Theorem A.1 *For integer α,*

$$\Gamma(\alpha;x)=1-\sum_{j=0}^{\alpha-1}\frac{x^j e^{-x}}{j!}.$$

Proof: For $\alpha=1$, $\Gamma(1;x)=\int_0^x e^{-t}\,dt=1-e^{-x}$, and so the theorem is true for this case. The proof is completed by induction. Assume it is true for $\alpha=1,\ldots,n$. Then

$$\begin{aligned}
\Gamma(n+1;x) &= \frac{1}{n!}\int_0^x t^n e^{-t}\,dt\\
&= \frac{1}{n!}\left(-t^n e^{-t}\Big|_0^x+\int_0^x nt^{n-1}e^{-t}\,dt\right)\\
&= \frac{1}{n!}\left(-x^n e^{-x}\right)+\Gamma(n;x)\\
&= -\frac{x^n e^{-x}}{n!}+1-\sum_{j=0}^{n-1}\frac{x^j e^{-x}}{j!}\\
&= 1-\sum_{j=0}^{n}\frac{x^j e^{-x}}{j!}.
\end{aligned}$$

The incomplete beta function is given by

$$\beta(a,b;x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x t^{a-1}(1-t)^{b-1}\,dt,\quad a>0,\ b>0,\ 0<x<1,$$

and when $b<0$ (but $a>1+\lfloor -b\rfloor$), repeated integration by parts produces

$$\begin{aligned}
\Gamma(a)\Gamma(b)\beta(a,b;x) &= -\Gamma(a+b)\left[\frac{x^{a-1}(1-x)^b}{b}\right.\\
&\quad+\frac{(a-1)x^{a-2}(1-x)^{b+1}}{b(b+1)}+\cdots\\
&\quad\left.+\frac{(a-1)\cdots(a-r)x^{a-r-1}(1-x)^{b+r}}{b(b+1)\cdots(b+r)}\right]\\
&\quad+\frac{(a-1)\cdots(a-r-1)}{b(b+1)\cdots(b+r)}\Gamma(a-r-1)\\
&\quad\times\Gamma(b+r+1)\beta(a-r-1,b+r+1;x),
\end{aligned}$$

where r is the smallest integer such that $b+r+1>0$. The first argument must be positive (that is, $a-r-1>0$).

Numerical approximations for both the incomplete gamma and the incomplete beta function are available in many statistical computing packages as well as in many spreadsheets because they are just the distribution functions of the gamma and beta distributions. The following approximations are taken from [3]. The suggestion regarding using different formulas for small and large x when evaluating the incomplete gamma function is from [107]. That reference also contains computer subroutines for evaluating these expressions. In particular, it provides an effective way of evaluating continued fractions.

For $x\le\alpha+1$ use the series expansion

$$\Gamma(\alpha;x)=\frac{x^\alpha e^{-x}}{\Gamma(\alpha)}\sum_{n=0}^{\infty}\frac{x^n}{\alpha(\alpha+1)\cdots(\alpha+n)}$$

while for $x>\alpha+1$ use the continued-fraction expansion

$$1-\Gamma(\alpha;x)=\frac{x^\alpha e^{-x}}{\Gamma(\alpha)}\cfrac{1}{x+\cfrac{1-\alpha}{1+\cfrac{1}{x+\cfrac{2-\alpha}{1+\cfrac{2}{x+\cdots}}}}}.$$

The incomplete gamma function can also be used to produce cumulative probabilities from the standard normal distribution. Let $\Phi(z)=\Pr(Z\le z)$,

where Z has the standard normal distribution. Then, for $z \geq 0$, $\Phi(z) = 0.5 + \Gamma(0.5; z^2/2)/2$ while, for $z < 0$, $\Phi(z) = 1 - \Phi(-z)$.

The incomplete beta function can be evaluated by the series expansion

$$\begin{aligned}\beta(a,b;x) &= \frac{\Gamma(a+b)x^a(1-x)^b}{a\Gamma(a)\Gamma(b)} \\ &\quad \times \left[1 + \sum_{n=0}^{\infty} \frac{(a+b)(a+b+1)\cdots(a+b+n)}{(a+1)(a+2)\cdots(a+n+1)} x^{n+1}\right].\end{aligned}$$

The gamma function itself can be found from

$$\begin{aligned}\ln\Gamma(\alpha) &\doteq (\alpha - \tfrac{1}{2})\ln\alpha - \alpha + \frac{\ln(2\pi)}{2} \\ &\quad + \frac{1}{12\alpha} - \frac{1}{360\alpha^3} + \frac{1}{1{,}260\alpha^5} - \frac{1}{1{,}680\alpha^7} + \frac{1}{1{,}188\alpha^9} - \frac{691}{360{,}360\alpha^{11}} \\ &\quad + \frac{1}{156\alpha^{13}} - \frac{3{,}617}{122{,}400\alpha^{15}} + \frac{43{,}867}{244{,}188\alpha^{17}} - \frac{174{,}611}{125{,}400\alpha^{19}}.\end{aligned}$$

For values of α above 10 the error is less than 10^{-19}. For values below 10 use the relationship

$$\ln\Gamma(\alpha) = \ln\Gamma(\alpha+1) - \ln\alpha.$$

The distributions are presented in the following way. First the name is given along with the parameters. Many of the distributions have other names, which are noted in parentheses. Next the density function $f(x)$ and distribution function $F(x)$ are given. For some distributions, formulas for starting values are given. Within each family the distributions are presented in decreasing order with regard to the number of parameters. The Greek letters used are selected to be consistent. Any Greek letter that is not used in the distribution means that that distribution is a special case of one with more parameters but with the missing parameters set equal to 1. Unless specifically indicated, all parameters must be positive.

Except for two distributions, inflation can be recognized by simply inflating the scale parameter θ. That is, if X has a particular distribution, then cX has the same distribution type, with all parameters unchanged except θ is changed to $c\theta$. For the lognormal distribution, μ changes to $\mu + \ln(c)$ with σ unchanged, while for the inverse Gaussian both μ and θ are multiplied by c.

For several of the distributions, starting values are suggested. They are not necessarily good estimators, just places from which to start an iterative procedure to maximize the likelihood or other objective function. These are found by either the methods of moments or percentile matching. The quantities used are:

$$\text{Moments:} \quad m = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad t = \frac{1}{n}\sum_{i=1}^{n} x_i^2,$$

$$\text{Percentile matching:} \quad p = \text{25th percentile}, \quad q = \text{75th percentile}.$$

For grouped data or data that have been truncated or censored, these quantities may have to be approximated. Because the purpose is to obtain starting values and not a useful estimate, it is often sufficient to just ignore modifications. For three- and four-parameter distributions, starting values can be obtained by using estimates from a special case, then making the new parameters equal to 1. An all-purpose starting value rule (for when all else fails) is to set the scale parameter (θ) equal to the mean and set all other parameters equal to 2.

All the distributions listed here (and many more) are discussed in great detail in [73]. In many cases, alternatives to maximum likelihood estimators are presented.

A.2 TRANSFORMED BETA FAMILY

A.2.1 Four-parameter distribution

A.2.1.1 Transformed beta—$\alpha, \theta, \gamma, \tau$ (generalized beta of the second kind, Pearson Type VI)

$$\begin{aligned}f(x) &= \frac{\Gamma(\alpha+\tau)}{\Gamma(\alpha)\Gamma(\tau)} \frac{\gamma(x/\theta)^{\gamma\tau}}{x[1+(x/\theta)^{\gamma}]^{\alpha+\tau}}, \\ F(x) &= \beta(\tau,\alpha;u), \quad u = \frac{(x/\theta)^{\gamma}}{1+(x/\theta)^{\gamma}}, \\ \mathrm{E}[X^k] &= \frac{\theta^k\Gamma(\tau+k/\gamma)\Gamma(\alpha-k/\gamma)}{\Gamma(\alpha)\Gamma(\tau)}, \quad -\tau\gamma < k < \alpha\gamma, \\ \mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(\tau+k/\gamma)\Gamma(\alpha-k/\gamma)}{\Gamma(\alpha)\Gamma(\tau)}\beta(\tau+k/\gamma,\alpha-k/\gamma;u) \\ &\quad + x^k[1-F(x)], \quad k > -\tau\gamma\,, \\ \text{Mode} &= \theta\left(\frac{\tau\gamma-1}{\alpha\gamma+1}\right)^{1/\gamma}, \quad \tau\gamma > 1, \text{ else } 0.\end{aligned}$$

A.2.2 Three-parameter distributions

A.2.2.1 Generalized Pareto—α, θ, τ (beta of the second kind)

$$\begin{aligned}f(x) &= \frac{\Gamma(\alpha+\tau)}{\Gamma(\alpha)\Gamma(\tau)} \frac{\theta^{\alpha}x^{\tau-1}}{(x+\theta)^{\alpha+\tau}}, \\ F(x) &= \beta(\tau,\alpha;u), \quad u = \frac{x}{x+\theta},\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(\tau+k)\Gamma(\alpha-k)}{\Gamma(\alpha)\Gamma(\tau)}, \quad -\tau < k < \alpha, \\
\mathrm{E}[X^k] &= \frac{\theta^k\tau(\tau+1)\cdots(\tau+k-1)}{(\alpha-1)\cdots(\alpha-k)} \quad \text{if } k \text{ is an integer}, \\
\mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(\tau+k)\Gamma(\alpha-k)}{\Gamma(\alpha)\Gamma(\tau)}\beta(\tau+k,\alpha-k;u), \\
&\quad + x^k[1-F(x)], \quad k > -\tau\,, \\
\text{Mode} &= \theta\frac{\tau-1}{\alpha+1}, \quad \tau > 1, \text{ else } 0.
\end{aligned}$$

A.2.2.2 *Burr*—α, θ, γ (Burr Type XII, Singh–Maddala)

$$\begin{aligned}
f(x) &= \frac{\alpha\gamma(x/\theta)^\gamma}{x[1+(x/\theta)^\gamma]^{\alpha+1}}, \\
F(x) &= 1-u^\alpha, \quad u = \frac{1}{1+(x/\theta)^\gamma}, \\
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(1+k/\gamma)\Gamma(\alpha-k/\gamma)}{\Gamma(\alpha)}, \quad -\gamma < k < \alpha\gamma, \\
\mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(1+k/\gamma)\Gamma(\alpha-k/\gamma)}{\Gamma(\alpha)}\beta(1+k/\gamma,\alpha-k/\gamma;1-u) \\
&\quad + x^k u^\alpha, \quad k > -\gamma\,, \\
\text{Mode} &= \theta\left(\frac{\gamma-1}{\alpha\gamma+1}\right)^{1/\gamma}, \quad \gamma > 1, \text{ else } 0.
\end{aligned}$$

A.2.2.3 *Inverse Burr*—τ, θ, γ (Dagum)

$$\begin{aligned}
f(x) &= \frac{\tau\gamma(x/\theta)^{\gamma\tau}}{x[1+(x/\theta)^\gamma]^{\tau+1}}, \\
F(x) &= u^\tau, \quad u = \frac{(x/\theta)^\gamma}{1+(x/\theta)^\gamma}, \\
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(\tau+k/\gamma)\Gamma(1-k/\gamma)}{\Gamma(\tau)}, \quad -\tau\gamma < k < \gamma, \\
\mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(\tau+k/\gamma)\Gamma(1-k/\gamma)}{\Gamma(\tau)}\beta(\tau+k/\gamma,1-k/\gamma;u) \\
&\quad + x^k[1-u^\tau], \quad k > -\tau\gamma\,, \\
\text{Mode} &= \theta\left(\frac{\tau\gamma-1}{\gamma+1}\right)^{1/\gamma}, \quad \tau\gamma > 1, \text{ else } 0.
\end{aligned}$$

A.2.3 Two-parameter distributions

A.2.3.1 *Pareto*—α, θ (Pareto Type II, Lomax)

$$\begin{aligned}
f(x) &= \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}, \\
F(x) &= 1-\left(\frac{\theta}{x+\theta}\right)^\alpha, \\
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(k+1)\Gamma(\alpha-k)}{\Gamma(\alpha)}, \quad -1 < k < \alpha, \\
\mathrm{E}[X^k] &= \frac{\theta^k k!}{(\alpha-1)\cdots(\alpha-k)} \quad \text{if } k \text{ is an integer} \\
\mathrm{E}[X\wedge x] &= \frac{\theta}{\alpha-1}\left[1-\left(\frac{\theta}{x+\theta}\right)^{\alpha-1}\right], \quad \alpha \neq 1, \\
\mathrm{E}[X\wedge x] &= -\theta\ln\left(\frac{\theta}{x+\theta}\right), \quad \alpha = 1, \\
\mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(k+1)\Gamma(\alpha-k)}{\Gamma(\alpha)}\beta[k+1,\alpha-k;x/(x+\theta)] \\
&\quad + x^k\left(\frac{\theta}{x+\theta}\right)^\alpha, \quad \text{all } k, \\
\text{Mode} &= 0, \\
\hat\alpha &= 2\frac{t-m^2}{t-2m^2}, \quad \hat\theta = \frac{mt}{t-2m^2}.
\end{aligned}$$

A.2.3.2 *Inverse Pareto*—τ, θ

$$\begin{aligned}
f(x) &= \frac{\tau\theta x^{\tau-1}}{(x+\theta)^{\tau+1}}, \\
F(x) &= \left(\frac{x}{x+\theta}\right)^\tau, \\
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(\tau+k)\Gamma(1-k)}{\Gamma(\tau)}, \quad -\tau < k < 1, \\
\mathrm{E}[X^k] &= \frac{\theta^k(-k)!}{(\tau-1)\cdots(\tau+k)} \quad \text{if } k \text{ is a negative integer}, \\
\mathrm{E}[(X\wedge x)^k] &= \theta^k\tau\int_0^{x/(x+\theta)} y^{\tau+k-1}(1-y)^{-k}dy \\
&\quad + x^k\left[1-\left(\frac{x}{x+\theta}\right)^\tau\right], \quad k > -\tau\,, \\
\text{Mode} &= \theta\frac{\tau-1}{2}, \quad \tau > 1, \text{ else } 0.
\end{aligned}$$

A.2.3.3 Loglogistic—γ, θ (Fisk)

$$
\begin{aligned}
f(x) &= \frac{\gamma(x/\theta)^\gamma}{x[1+(x/\theta)^\gamma]^2}, \\
F(x) &= u, \quad u = \frac{(x/\theta)^\gamma}{1+(x/\theta)^\gamma}, \\
\mathrm{E}[X^k] &= \theta^k \Gamma(1+k/\gamma)\Gamma(1-k/\gamma), \quad -\gamma < k < \gamma, \\
\mathrm{E}[(X \wedge x)^k] &= \theta^k \Gamma(1+k/\gamma)\Gamma(1-k/\gamma)\beta(1+k/\gamma, 1-k/\gamma; u) \\
&\quad + x^k(1-u), \quad k > -\gamma\,, \\
\text{Mode} &= \theta\left(\frac{\gamma-1}{\gamma+1}\right)^{1/\gamma}, \quad \gamma > 1, \text{ else } 0, \\
\hat{\gamma} &= \frac{2\ln(3)}{\ln(q)-\ln(p)}, \quad \hat{\theta} = \exp\left(\frac{\ln(q)+\ln(p)}{2}\right).
\end{aligned}
$$

A.2.3.4 Paralogistic—α, θ This is a Burr distribution with $\gamma = \alpha$.

$$
\begin{aligned}
f(x) &= \frac{\alpha^2(x/\theta)^\alpha}{x[1+(x/\theta)^\alpha]^{\alpha+1}}, \\
F(x) &= 1-u^\alpha, \quad u = \frac{1}{1+(x/\theta)^\alpha}, \\
\mathrm{E}[X^k] &= \frac{\theta^k \Gamma(1+k/\alpha)\Gamma(\alpha-k/\alpha)}{\Gamma(\alpha)}, \quad -\alpha < k < \alpha^2, \\
\mathrm{E}[(X \wedge x)^k] &= \frac{\theta^k \Gamma(1+k/\alpha)\Gamma(\alpha-k/\alpha)}{\Gamma(\alpha)}\beta(1+k/\alpha, \alpha-k/\alpha; 1-u) \\
&\quad + x^k u^\alpha, \quad k > -\alpha\,, \\
\text{Mode} &= \theta\left(\frac{\alpha-1}{\alpha^2+1}\right)^{1/\alpha}, \quad \alpha > 1, \text{ else } 0
\end{aligned}
$$

Starting values can use estimates from the loglogistic (use γ for α) or Pareto (use α) distributions.

A.2.3.5 Inverse paralogistic—τ, θ This is an inverse Burr distribution with $\gamma = \tau$.

$$
\begin{aligned}
f(x) &= \frac{\tau^2(x/\theta)^{\tau^2}}{x[1+(x/\theta)^\tau]^{\tau+1}}, \\
F(x) &= u^\tau, \quad u = \frac{(x/\theta)^\tau}{1+(x/\theta)^\tau}, \\
\mathrm{E}[X^k] &= \frac{\theta^k \Gamma(\tau+k/\tau)\Gamma(1-k/\tau)}{\Gamma(\tau)}, \quad -\tau^2 < k < \tau, \\
\mathrm{E}[(X \wedge x)^k] &= \frac{\theta^k \Gamma(\tau+k/\tau)\Gamma(1-k/\tau)}{\Gamma(\tau)}\beta(\tau+k/\tau, 1-k/\tau; u) \\
&\quad + x^k[1-u^\tau], \quad k > -\tau^2\,, \\
\text{Mode} &= \theta(\tau-1)^{1/\tau}, \quad \tau > 1, \text{ else } 0.
\end{aligned}
$$

Starting values can use estimates from the loglogistic (use γ for τ) or inverse Pareto (use τ) distributions.

A.3 TRANSFORMED GAMMA FAMILY

A.3.1 Three-parameter distributions

A.3.1.1 Transformed gamma—α, θ, τ (generalized gamma)

$$
\begin{aligned}
f(x) &= \frac{\tau u^\alpha e^{-u}}{x\Gamma(\alpha)}, \quad u = (x/\theta)^\tau, \\
F(x) &= \Gamma(\alpha; u), \\
\mathrm{E}[X^k] &= \frac{\theta^k \Gamma(\alpha+k/\tau)}{\Gamma(\alpha)}, \quad k > -\alpha\tau, \\
\mathrm{E}[(X \wedge x)^k] &= \frac{\theta^k \Gamma(\alpha+k/\tau)}{\Gamma(\alpha)}\Gamma(\alpha+k/\tau; u) \\
&\quad + x^k[1-\Gamma(\alpha; u)], \quad k > -\alpha\tau\,, \\
\text{Mode} &= \theta\left(\frac{\alpha\tau-1}{\tau}\right)^{1/\tau}, \quad \alpha\tau > 1, \text{ else } 0.
\end{aligned}
$$

A.3.1.2 Inverse transformed gamma—α, θ, τ (inverse generalized gamma)

$$\begin{aligned}
f(x) &= \frac{\tau u^{\alpha} e^{-u}}{x\Gamma(\alpha)}, \quad u = (\theta/x)^{\tau}, \\
F(x) &= 1 - \Gamma(\alpha; u), \\
\mathrm{E}[X^k] &= \frac{\theta^k \Gamma(\alpha - k/\tau)}{\Gamma(\alpha)}, \quad k < \alpha\tau, \\
\mathrm{E}[(X \wedge x)^k] &= \frac{\theta^k \Gamma(\alpha - k/\tau)}{\Gamma(\alpha)}[1 - \Gamma(\alpha - k/\tau; u)] + x^k \Gamma(\alpha; u) \\
&= \frac{\theta^k G(\alpha - k/\tau; u)}{\Gamma(\alpha)} + x^k \Gamma(\alpha; u), \quad \text{all } k, \\
\text{Mode} &= \theta \left(\frac{\tau}{\alpha\tau + 1} \right)^{1/\tau}.
\end{aligned}$$

A.3.2 Two-parameter distributions

A.3.2.1 Gamma—α, θ

$$\begin{aligned}
f(x) &= \frac{(x/\theta)^{\alpha} e^{-x/\theta}}{x\Gamma(\alpha)}, \\
F(x) &= \Gamma(\alpha; x/\theta), \\
\mathrm{E}[X^k] &= \frac{\theta^k \Gamma(\alpha + k)}{\Gamma(\alpha)}, \quad k > -\alpha, \\
\mathrm{E}[X^k] &= \theta^k (\alpha + k - 1) \cdots \alpha \quad \text{if } k \text{ is an integer}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[(X \wedge x)^k] &= \frac{\theta^k \Gamma(\alpha + k)}{\Gamma(\alpha)} \Gamma(\alpha + k; x/\theta) + x^k [1 - \Gamma(\alpha; x/\theta)], \quad k > -\alpha \\
\mathrm{E}[(X \wedge x)^k] &= \alpha(\alpha + 1) \cdots (\alpha + k - 1)\theta^k \Gamma(\alpha + k; x/\theta) \\
&\quad + x^k [1 - \Gamma(\alpha; x/\theta)] \quad \text{if } k \text{ is an integer}, \\
M(t) &= (1 - \theta t)^{-\alpha}, \quad t < 1/\theta, \\
\text{Mode} &= \theta(\alpha - 1), \quad \alpha > 1, \text{ else } 0, \\
\hat{\alpha} &= \frac{m^2}{t - m^2}, \quad \hat{\theta} = \frac{t - m^2}{m}.
\end{aligned}$$

A.3.2.2 Inverse gamma—α, θ (Vinci)

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

$$\begin{aligned}
\mathrm{E}[X^k] &= \frac{\theta^k \Gamma(\alpha - k)}{\Gamma(\alpha)}, \quad k < \alpha, \\
\mathrm{E}[X^k] &= \frac{\theta^k}{(\alpha - 1) \cdots (\alpha - k)} \quad \text{if } k \text{ is an integer}, \\
\mathrm{E}[(X \wedge x)^k] &= \frac{\theta^k \Gamma(\alpha - k)}{\Gamma(\alpha)}[1 - \Gamma(\alpha - k; \theta/x)] + x^k \Gamma(\alpha; \theta/x) \\
&= \frac{\theta^k G(\alpha - k; \theta/x)}{\Gamma(\alpha)} + x^k \Gamma(\alpha; \theta/x), \text{ all } k, \\
\text{Mode} &= \theta/(\alpha + 1), \\
\hat{\alpha} &= \frac{2t - m^2}{t - m^2}, \quad \hat{\theta} = \frac{mt}{t - m^2}.
\end{aligned}$$

A.3.2.3 Weibull—θ, τ

$$\begin{aligned}
f(x) &= \frac{\tau(x/\theta)^{\tau} e^{-(x/\theta)^{\tau}}}{x}, \\
F(x) &= 1 - e^{-(x/\theta)^{\tau}}, \\
\mathrm{E}[X^k] &= \theta^k \Gamma(1 + k/\tau), \quad k > -\tau, \\
\mathrm{E}[(X \wedge x)^k] &= \theta^k \Gamma(1 + k/\tau)\Gamma[1 + k/\tau; (x/\theta)^{\tau}] + x^k e^{-(x/\theta)^{\tau}}, \quad k > -\tau, \\
\text{Mode} &= \theta \left(\frac{\tau - 1}{\tau} \right)^{1/\tau}, \quad \tau > 1, \text{ else } 0, \\
\hat{\theta} &= \exp\left(\frac{g \ln(p) - \ln(q)}{g - 1} \right), \quad g = \frac{\ln(\ln(4))}{\ln(\ln(4/3))}, \\
\hat{\tau} &= \frac{\ln(\ln(4))}{\ln(q) - \ln(\hat{\theta})}.
\end{aligned}$$

A.3.2.4 Inverse Weibull—θ, τ (log-Gompertz)

$$\begin{aligned}
f(x) &= \frac{\tau(\theta/x)^{\tau} e^{-(\theta/x)^{\tau}}}{x}, \\
F(x) &= e^{-(\theta/x)^{\tau}}, \\
\mathrm{E}[X^k] &= \theta^k \Gamma(1 - k/\tau), \quad k < \tau, \\
\mathrm{E}[(X \wedge x)^k] &= \theta^k \Gamma(1 - k/\tau)\{1 - \Gamma[1 - k/\tau; (\theta/x)^{\tau}]\} \\
&\quad + x^k \left[1 - e^{-(\theta/x)^{\tau}} \right], \\
&= \theta^k G[1 - k/\tau; (\theta/x)^{\tau}] + x^k \left[1 - e^{-(\theta/x)^{\tau}} \right], \quad \text{all } k,
\end{aligned}$$

$$\begin{aligned}
\text{Mode} &= \theta\left(\frac{\tau}{\tau+1}\right)^{1/\tau}, \\
\hat{\theta} &= \exp\left(\frac{g\ln(q)-\ln(p)}{g-1}\right), \quad g=\frac{\ln(\ln(4))}{\ln(\ln(4/3))}, \\
\hat{\tau} &= \frac{\ln(\ln(4))}{\ln(\hat{\theta})-\ln(p)}.
\end{aligned}$$

A.3.3 One-parameter distributions

A.3.3.1 Exponential—θ

$$\begin{aligned}
f(x) &= \frac{e^{-x/\theta}}{\theta}, \\
F(x) &= 1-e^{-x/\theta}, \\
\mathrm{E}[X^k] &= \theta^k\Gamma(k+1), \quad k>-1, \\
\mathrm{E}[X^k] &= \theta^k k! \quad \text{if } k \text{ is an integer}, \\
\mathrm{E}[X\wedge x] &= \theta(1-e^{-x/\theta}), \\
\mathrm{E}[(X\wedge x)^k] &= \theta^k\Gamma(k+1)\Gamma(k+1;x/\theta)+x^k e^{-x/\theta}, \quad k>-1, \\
\mathrm{E}[(X\wedge x)^k] &= \theta^k k!\Gamma(k+1;x/\theta)+x^k e^{-x/\theta} \quad \text{if } k>-1 \text{ is an integer}, \\
M(t) &= (1-\theta t)^{-1}, \quad t<1/\theta, \\
\text{Mode} &= 0, \\
\hat{\theta} &= m.
\end{aligned}$$

A.3.3.2 Inverse exponential—θ

$$\begin{aligned}
f(x) &= \frac{\theta e^{-\theta/x}}{x^2}, \\
F(x) &= e^{-\theta/x}, \\
\mathrm{E}[X^k] &= \theta^k\Gamma(1-k), \quad k<1, \\
\mathrm{E}[(X\wedge x)^k] &= \theta^k G(1-k;\theta/x)+x^k(1-e^{-\theta/x}), \quad \text{all } k, \\
\text{Mode} &= \theta/2, \\
\hat{\theta} &= -q\ln(3/4).
\end{aligned}$$

A.4 OTHER DISTRIBUTIONS

A.4.1.1 Lognormal—μ,σ (μ can be negative)

$$\begin{aligned}
f(x) &= \frac{1}{x\sigma\sqrt{2\pi}}\exp(-z^2/2)=\phi(z)/(\sigma x), \quad z=\frac{\ln x-\mu}{\sigma}, \\
F(x) &= \Phi(z),
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[X^k] &= \exp\left(k\mu+\tfrac{1}{2}k^2\sigma^2\right), \\
\mathrm{E}[(X\wedge x)^k] &= \exp\left(k\mu+\tfrac{1}{2}k^2\sigma^2\right)\Phi\left(\frac{\ln x-\mu-k\sigma^2}{\sigma}\right)+x^k[1-F(x)], \\
\text{Mode} &= \exp(\mu-\sigma^2), \\
\hat{\sigma} &= \sqrt{\ln(t)-2\ln(m)}, \quad \hat{\mu}=\ln(m)-\tfrac{1}{2}\hat{\sigma}^2.
\end{aligned}$$

A.4.1.2 Inverse Gaussian—μ, θ

$$\begin{aligned}
f(x) &= \left(\frac{\theta}{2\pi x^3}\right)^{1/2}\exp\left(-\frac{\theta z^2}{2x}\right), \quad z=\frac{x-\mu}{\mu}, \\
F(x) &= \Phi\left[z\left(\frac{\theta}{x}\right)^{1/2}\right]+\exp\left(\frac{2\theta}{\mu}\right)\Phi\left[-y\left(\frac{\theta}{x}\right)^{1/2}\right], \quad y=\frac{x+\mu}{\mu}, \\
\mathrm{E}[X] &= \mu, \quad Var[X]=\mu^3/\theta, \\
\mathrm{E}[X\wedge x] &= x-\mu z\Phi\left[z\left(\frac{\theta}{x}\right)^{1/2}\right]-\mu y\exp(2\theta/\mu)\Phi\left[-y\left(\frac{\theta}{x}\right)^{1/2}\right], \\
M(t) &= \exp\left[\frac{\theta}{\mu}\left(1-\sqrt{1-\frac{2\mu^2}{\theta}t}\right)\right], \quad t<\frac{\theta}{2\mu^2}, \\
\hat{\mu} &= m, \quad \hat{\theta}=\frac{m^3}{t-m^2}.
\end{aligned}$$

A.4.1.3 log-t—r, μ, σ (μ can be negative)

Let Y have a t distribution with r degrees of freedom. Then $X=\exp(\sigma Y+\mu)$ has the log-t distribution. Positive moments do not exist for this distribution. Just as the t distribution has a heavier tail than the normal distribution, this distribution has a heavier tail than the lognormal distribution.

$$\begin{aligned}
f(x) &= \frac{\Gamma\left(\frac{r+1}{2}\right)}{x\sigma\sqrt{\pi r}\Gamma\left(\frac{r}{2}\right)\left[1+\frac{1}{r}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right]^{(r+1)/2}}, \\
F(x) &= F_r\left(\frac{\ln x-\mu}{\sigma}\right) \text{ with } F_r(t) \text{ the cdf of a } t \text{ distribution with } r \text{ d.f.},
\end{aligned}$$

$$F(x) = \begin{cases} \frac{1}{2}\beta\left[\frac{r}{2},\frac{1}{2};\frac{r}{r+\left(\frac{\ln x-\mu}{\sigma}\right)^2}\right], & 0<x\le e^{\mu}, \\ 1-\frac{1}{2}\beta\left[\frac{r}{2},\frac{1}{2};\frac{r}{r+\left(\frac{\ln x-\mu}{\sigma}\right)^2}\right], & x\ge e^{\mu}. \end{cases}$$

A.4.1.4 Single-parameter Pareto—α, θ

$$\begin{aligned} f(x) &= \frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}, \quad x>\theta, \\ F(x) &= 1-\left(\frac{\theta}{x}\right)^{\alpha}, \quad x>\theta, \end{aligned}$$

$$\begin{aligned} \mathrm{E}[X^k] &= \frac{\alpha\theta^k}{\alpha-k}, \quad k<\alpha, \\ \mathrm{E}[(X\wedge x)^k] &= \frac{\alpha\theta^k}{\alpha-k}-\frac{k\theta^{\alpha}}{(\alpha-k)x^{\alpha-k}}, \\ \text{Mode} &= \theta, \\ \hat{\alpha} &= \frac{m}{m-\theta}. \end{aligned}$$

Note: Although there appears to be two parameters, only α is a true parameter. The value of θ must be set in advance.

A.5 DISTRIBUTIONS WITH FINITE SUPPORT

For these two distributions, the scale parameter θ is assumed known.

A.5.1.1 Generalized beta—a, b, θ, τ

$$\begin{aligned} f(x) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}u^a(1-u)^{b-1}\frac{\tau}{x}, \quad 0<x<\theta, \quad u=(x/\theta)^{\tau}, \\ F(x) &= \beta(a,b;u), \\ \mathrm{E}[X^k] &= \frac{\theta^k\Gamma(a+b)\Gamma(a+k/\tau)}{\Gamma(a)\Gamma(a+b+k/\tau)}, \quad k>-a\tau, \\ \mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(a+b)\Gamma(a+k/\tau)}{\Gamma(a)\Gamma(a+b+k/\tau)}\beta(a+k/\tau,b;u)+x^k[1-\beta(a,b;u)]. \end{aligned}$$

A.5.1.2 beta—a, b, θ

$$\begin{aligned} f(x) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}u^a(1-u)^{b-1}\frac{1}{x}, \quad 0<x<\theta, \quad u=x/\theta, \\ F(x) &= \beta(a,b;u), \\ \mathrm{E}[X^k] &= \frac{\theta^k\Gamma(a+b)\Gamma(a+k)}{\Gamma(a)\Gamma(a+b+k)}, \quad k>-a, \\ \mathrm{E}[X^k] &= \frac{\theta^k a(a+1)\cdots(a+k-1)}{(a+b)(a+b+1)\cdots(a+b+k-1)} \quad \text{if } k \text{ is an integer}, \\ \mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k a(a+1)\cdots(a+k-1)}{(a+b)(a+b+1)\cdots(a+b+k-1)}\beta(a+k,b;u) \\ &\quad +x^k[1-\beta(a,b;u)], \\ \hat{a} &= \frac{\theta m^2-mt}{\theta t-\theta m^2}, \quad \hat{b}=\frac{(\theta m-t)(\theta-m)}{\theta t-\theta m^2}. \end{aligned}$$

Appendix B
An inventory of discrete distributions

B.1 INTRODUCTION

The 16 models fall into three classes. The divisions are based on the algorithm by which the probabilities are computed. For some of the more familiar distributions these formulas will look different from the ones you may have learned, but they produce the same probabilities. After each name, the parameters are given. All parameters are positive unless otherwise indicated. In all cases, p_k is the probability of observing k losses.

For finding moments, the most convenient form is to give the factorial moments. The jth factorial moment is $\mu_{(j)} = \mathrm{E}[N(N-1)\cdots(N-j+1)]$. We have $\mathrm{E}[N] = \mu_{(1)}$ and $\mathrm{Var}(N) = \mu_{(2)} + \mu_{(1)} - \mu_{(1)}^2$.

The estimators which are presented are not intended to be useful estimators but rather for providing starting values for maximizing the likelihood (or other) function. For determining starting values, the following quantities are

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

used [where n_k is the observed frequency at k (if, for the last entry, n_k represents the number of observations at k or more, assume it was at exactly k) and n is the sample size]:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{\infty} kn_k, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{\infty} k^2 n_k - \hat{\mu}^2.$$

When the method of moments is used to determine the starting value, a circumflex (e.g., $\hat{\lambda}$) is used. For any other method, a tilde (e.g., $\tilde{\lambda}$) is used. When the starting value formulas do not provide admissible parameter values, a truly crude guess is to set the product of all λ and β parameters equal to the sample mean and set all other parameters equal to 1. If there are two λ and/or β parameters, an easy choice is to set each to the square root of the sample mean.

The last item presented is the probability generating function,

$$P(z) = \mathrm{E}[z^N].$$

B.2 THE $(a, b, 0)$ CLASS

The distributions in this class have support on $0, 1, \ldots$. For this class, a particular distribution is specified by setting p_0 and then using $p_k = (a + b/k)p_{k-1}$. Specific members are created by setting p_0, a, and b. For any member, $\mu_{(1)} = (a+b)/(1-a)$, and for higher j, $\mu_{(j)} = (aj+b)\mu_{(j-1)}/(1-a)$. The variance is $(a+b)/(1-a)^2$.

B.2.1.1 Poisson—λ

$$\begin{aligned} p_0 &= e^{-\lambda}, \quad a = 0, \quad b = \lambda, \\ p_k &= \frac{e^{-\lambda}\lambda^k}{k!}, \\ \mathrm{E}[N] &= \lambda, \quad \mathrm{Var}[N] = \lambda, \\ \hat{\lambda} &= \hat{\mu}, \\ P(z) &= e^{\lambda(z-1)}. \end{aligned}$$

B.2.1.2 Geometric—β

$$\begin{aligned} p_0 &= \frac{1}{1+\beta}, \quad a = \frac{\beta}{1+\beta}, \quad b = 0, \\ p_k &= \frac{\beta^k}{(1+\beta)^{k+1}}, \\ \mathrm{E}[N] &= \beta, \quad \mathrm{Var}[N] = \beta(1+\beta), \\ \hat{\beta} &= \hat{\mu}, \\ P(z) &= [1-\beta(z-1)]^{-1}. \end{aligned}$$

This is a special case of the negative binomial with $r = 1$.

B.2.1.3 Binomial—$q, m, (0 < q < 1, m$ an integer)

$$\begin{aligned} p_0 &= (1-q)^m, \quad a = -\frac{q}{1-q}, \quad b = \frac{(m+1)q}{1-q}, \\ p_k &= \binom{m}{k} q^k (1-q)^{m-k}, \quad k = 0, 1, \ldots, m, \\ \mathrm{E}[N] &= mq, \quad \mathrm{Var}[N] = mq(1-q), \\ \hat{q} &= \hat{\mu}/m, \\ P(z) &= [1+q(z-1)]^m. \end{aligned}$$

B.2.1.4 Negative binomial—β, r

$$\begin{aligned} p_0 &= (1+\beta)^{-r}, \quad a = \frac{\beta}{1+\beta}, \quad b = \frac{(r-1)\beta}{1+\beta}, \\ p_k &= \frac{r(r+1)\cdots(r+k-1)\beta^k}{k!(1+\beta)^{r+k}}, \\ \mathrm{E}[N] &= r\beta, \quad \mathrm{Var}[N] = r\beta(1+\beta), \\ \hat{\beta} &= \frac{\hat{\sigma}^2}{\hat{\mu}} - 1, \quad \hat{r} = \frac{\hat{\mu}^2}{\hat{\sigma}^2 - \hat{\mu}}, \\ P(z) &= [1-\beta(z-1)]^{-r}. \end{aligned}$$

B.3 THE $(a, b, 1)$ CLASS

To distinguish this class from the $(a, b, 0)$ class, the probabilities are denoted $\Pr(N = k) = p_k^M$ or $\Pr(N = k) = p_k^T$ depending on which subclass is being represented. For this class, p_0^M is arbitrary (that is, it is a parameter) and then p_1^M or p_1^T is a specified function of the parameters a and b. Subsequent probabilities are obtained recursively as in the $(a, b, 0)$ class: $p_k^M = (a + b/k)p_{k-1}^M$, $k = 2, 3, \ldots$, with the same recursion for p_k^T There are two subclasses of this class. When discussing their members, we often refer to the "corresponding" member of the $(a, b, 0)$ class. This refers to the member of that class with the same values for a and b. The notation p_k will continue to be used for probabilities for the corresponding $(a, b, 0)$ distribution.

B.3.1 The zero-truncated subclass

The members of this class have $p_0^T = 0$ and therefore it need not be estimated. These distributions should only be used when a value of zero is impossible. The first factorial moment is $\mu_{(1)} = (a+b)/[(1-a)(1-p_0)]$, where p_0 is the value for the corresponding member of the $(a, b, 0)$ class. For the logarithmic

distribution (which has no corresponding member), $\mu_{(1)} = \beta/\ln(1+\beta)$. Higher factorial moments are obtained recursively with the same formula as with the $(a,b,0)$ class. The variance is $(a+b)[1-(a+b+1)p_0]/[(1-a)(1-p_0)]^2$.For those members of the subclass which have corresponding $(a,b,0)$ distributions, $p_k^T = p_k/(1-p_0)$.

B.3.1.1 Zero-truncated Poisson—λ

$$\begin{aligned}
p_1^T &= \frac{\lambda}{e^\lambda - 1}, \quad a = 0, \quad b = \lambda, \\
p_k^T &= \frac{\lambda^k}{k!(e^\lambda - 1)}, \\
\mathrm{E}[N] &= \lambda/(1-e^{-\lambda}), \quad \mathrm{Var}[N] = \lambda[1-(\lambda+1)e^{-\lambda}]/(1-e^{-\lambda})^2, \\
\tilde{\lambda} &= \ln(n\hat{\mu}/n_1), \\
P(z) &= \frac{e^{\lambda z} - 1}{e^\lambda - 1}.
\end{aligned}$$

B.3.1.2 Zero-truncated geometric—β

$$\begin{aligned}
p_1^T &= \frac{1}{1+\beta}, \quad a = \frac{\beta}{1+\beta}, \quad b = 0, \\
p_k^T &= \frac{\beta^{k-1}}{(1+\beta)^k}, \\
\mathrm{E}[N] &= 1+\beta, \quad \mathrm{Var}[N] = \beta(1+\beta), \\
\hat{\beta} &= \hat{\mu} - 1, \\
P(z) &= \frac{[1-\beta(z-1)]^{-1} - (1+\beta)^{-1}}{1-(1+\beta)^{-1}}.
\end{aligned}$$

This is a special case of the zero-truncated negative binomial with $r = 1$.

B.3.1.3 Logarithmic—β

$$\begin{aligned}
p_1^T &= \frac{\beta}{(1+\beta)\ln(1+\beta)}, \quad a = \frac{\beta}{1+\beta}, \quad b = -\frac{\beta}{1+\beta}, \\
p_k^T &= \frac{\beta^k}{k(1+\beta)^k \ln(1+\beta)}, \\
\mathrm{E}[N] &= \beta/\ln(1+\beta), \quad \mathrm{Var}[N] = \frac{\beta[1+\beta-\beta/\ln(1+\beta)]}{\ln(1+\beta)}, \\
\tilde{\beta} &= \frac{n\hat{\mu}}{n_1} - 1 \quad \text{or} \quad \frac{2(\hat{\mu}-1)}{\hat{\mu}}, \\
P(z) &= 1 - \frac{\ln[1-\beta(z-1)]}{\ln(1+\beta)}.
\end{aligned}$$

This is a limiting case of the zero-truncated negative binomial as $r \to 0$.

B.3.1.4 Zero-truncated binomial—q, m, $(0 < q < 1$, m an integer)

$$\begin{aligned}
p_1^T &= \frac{m(1-q)^{m-1}q}{1-(1-q)^m}, \quad a = -\frac{q}{1-q}, \quad b = \frac{(m+1)q}{1-q}, \\
p_k^T &= \frac{\binom{m}{k}q^k(1-q)^{m-k}}{1-(1-q)^m}, \quad k = 1,2,\ldots,m, \\
\mathrm{E}[N] &= \frac{mq}{1-(1-q)^m}, \\
\mathrm{Var}[N] &= \frac{mq[(1-q)-(1-q+mq)(1-q)^m]}{[1-(1-q)^m]^2}, \\
\tilde{q} &= \frac{\hat{\mu}}{m}, \\
P(z) &= \frac{[1+q(z-1)]^m - (1-q)^m}{1-(1-q)^m}.
\end{aligned}$$

B.3.1.5 Zero-truncated negative binomial—β, r, $(r > -1, r \neq 0)$

$$\begin{aligned}
p_1^T &= \frac{r\beta}{(1+\beta)^{r+1} - (1+\beta)}, \quad a = \frac{\beta}{1+\beta}, \quad b = \frac{(r-1)\beta}{1+\beta}, \\
p_k^T &= \frac{r(r+1)\cdots(r+k-1)}{k![(1+\beta)^r - 1]}\left(\frac{\beta}{1+\beta}\right)^k, \\
\mathrm{E}[N] &= \frac{r\beta}{1-(1+\beta)^{-r}}, \\
Var[N] &= \frac{r\beta[(1+\beta)-(1+\beta+r\beta)(1+\beta)^{-r}]}{[1-(1+\beta)^{-r}]^2}, \\
\tilde{\beta} &= \frac{\hat{\sigma}^2}{\hat{\mu}} - 1, \quad \tilde{r} = \frac{\hat{\mu}^2}{\hat{\sigma}^2 - \hat{\mu}}, \\
P(z) &= \frac{[1-\beta(z-1)]^{-r} - (1+\beta)^{-r}}{1-(1+\beta)^{-r}}.
\end{aligned}$$

This distribution is sometimes called the extended truncated negative binomial distribution because the parameter r can extend below 0.

B.3.2 The zero-modified subclass

A zero-modified distribution is created by starting with a truncated distribution and then placing an arbitrary amount of probability at zero. This probability, p_0^M, is a parameter. The remaining probabilities are adjusted accordingly. Values of p_k^M can be determined from the corresponding zero-truncated distribution as $p_k^M = (1-p_0^M)p_k^T$ or from the corresponding $(a,b,0)$ distribution as $p_k^M = (1-p_0^M)p_k/(1-p_0)$. The same recursion used for the zero-truncated subclass applies.

The mean is $1-p_0^M$ times the mean for the corresponding zero-truncated distribution. The variance is $1-p_0^M$ times the zero-truncated variance plus $p_0^M(1-p_0^M)$ times the square of the zero-truncated mean. The probability generating function is $P^M(z)=p_0^M+(1-p_0^M)P(z)$, where $P(z)$ is the probability generating function for the corresponding zero-truncated distribution.

The maximum likelihood estimator of p_0^M is always the sample relative frequency at 0.

B.4 THE COMPOUND CLASS

Members of this class are obtained by compounding one distribution with another. That is, let N be a discrete distribution, called the **primary distribution** and let $M_1, M_2, \ldots$ be identically and independently distributed with another discrete distribution, called the **secondary distribution**. The compound distribution is $S=M_1+\cdots+M_N$. The probabilities for the compound distributions are found from

$$p_k=\frac{1}{1-af_0}\sum_{y=1}^{k}(a+by/k)f_y p_{k-y}$$

for $n=1,2,\ldots$, where a and b are the usual values for the primary distribution [which must be a member of the $(a,b,0)$ class] and f_y is p_y for the secondary distribution. The only two primary distributions used here are Poisson (for which $p_0=\exp[-\lambda(1-f_0)]$) and geometric [for which $p_0=1/[1+\beta-\beta f_0]$]. Because this information completely describes these distributions, only the names and starting values are given below.

The moments can be found from the moments of the individual distributions:

$$\mathrm{E}[S]=\mathrm{E}[N]\mathrm{E}[M] \quad \text{and} \quad \mathrm{Var}[S]=\mathrm{E}[N]\,\mathrm{Var}[M]+\mathrm{Var}[N]\mathrm{E}[M]^2.$$

The probability generating function is $P(z)=P_{\text{primary}}[P_{\text{secondary}}(z)]$.

In the following list the primary distribution is always named first. For the first, second, and fourth distributions, the secondary distribution is the $(a,b,0)$ class member with that name. For the third and the last three distributions (the Poisson–ETNB and its two special cases) the secondary distribution is the zero-truncated version.

B.4.1 Some compound distributions

B.4.1.1 Poisson–binomial—λ, q, m ($0<q<1$, m an integer)

$$\hat{q}=\frac{\hat{\sigma}^2/\hat{\mu}-1}{m-1}, \quad \hat{\lambda}=\frac{\hat{\mu}}{m\hat{q}} \quad \text{or} \quad \tilde{q}=0.5,\ \tilde{\lambda}=\frac{2\hat{\mu}}{m}.$$

B.4.1.2 Poisson–Poisson—λ_1, λ_2 The parameter λ_1 is for the primary Poisson distribution, and λ_2 is for the secondary Poisson distribution. This distribution is also called the **Neyman Type A**.

$$\tilde{\lambda}_1=\tilde{\lambda}_2=\sqrt{\hat{\mu}}.$$

B.4.1.3 Geometric–extended truncated negative binomial—β_1, β_2, r ($r>-1$) The parameter β_1 is for the primary geometric distribution. The last two parameters are for the secondary distribution, noting that for $r=0$ the secondary distribution is logarithmic. The truncated version is used so that the extension of r is available.

$$\tilde{\beta}_1=\tilde{\beta}_2=\sqrt{\hat{\mu}}.$$

B.4.1.4 Geometric–Poisson—β, λ

$$\tilde{\beta}=\tilde{\lambda}=\sqrt{\hat{\mu}}.$$

B.4.1.5 Poisson–extended truncated negative binomial—λ, β, ($r>-1, r\neq 0$) When $r=0$ the secondary distribution is logarithmic, resulting in the negative binomial distribution.

$$\tilde{r} = \frac{\hat{\mu}(K-3\hat{\sigma}^2+2\hat{\mu})-2(\hat{\sigma}^2-\hat{\mu})^2}{\hat{\mu}(K-3\hat{\sigma}^2+2\hat{\mu})-(\hat{\sigma}^2-\hat{\mu})^2}, \quad \tilde{\beta}=\frac{\hat{\sigma}^2-\hat{\mu}}{\hat{\mu}(1+\hat{r})}, \quad \tilde{\lambda}=\frac{\hat{\mu}}{\hat{r}\hat{\beta}},$$

or,

$$\tilde{r} = \frac{\hat{\sigma}^2 n_1/n-\hat{\mu}^2 n_0/n}{(\hat{\sigma}^2-\hat{\mu}^2)(n_0/n)\ln(n_0/n)-\hat{\mu}(\hat{\mu}n_0/n-n_1/n)},$$

$$\tilde{\beta}=\frac{\hat{\sigma}^2-\hat{\mu}}{\hat{\mu}(1+\hat{r})}, \quad \tilde{\lambda}=\frac{\hat{\mu}}{\hat{r}\hat{\beta}}$$

where

$$K=\frac{1}{n}\sum_{k=0}^{\infty}k^3 n_k-3\hat{\mu}\frac{1}{n}\sum_{k=0}^{\infty}k^2 n_k+2\hat{\mu}^3.$$

This distribution is also called the **generalized Poisson–Pascal**.

B.4.1.6 Polya–Aeppli—λ, β

$$\hat{\beta}=\frac{\hat{\sigma}^2-\hat{\mu}}{2\hat{\mu}}, \quad \hat{\lambda}=\frac{\hat{\mu}}{1+\hat{\beta}}.$$

This is a special case of the Poisson–extended truncated negative binomial with $r=1$. It is actually a Poisson–truncated geometric.

B.4.1.7 Poisson–inverse Gaussian—λ, β

$$\tilde{\lambda} = -\ln(n_0/n),\ \tilde{\beta} = \frac{4(\hat{\mu} - \hat{\lambda})}{\hat{\mu}}.$$

This is a special case of the Poisson–extended truncated negative binomial with $r = -0.5$.

B.5 A HIERARCHY OF DISCRETE DISTRIBUTIONS

The following table indicates which distributions are special or limiting cases of others. For the special cases, one parameter is set equal to a constant to create the special case. For the limiting cases, two parameters go to infinity or zero in some special way.

Distribution	Is a special case of	Is a limiting case of
Poisson	ZM Poisson	Negative binomial, Poisson–binomial, Poisson–inv. Gaussian, Polya–Aeppli, Neyman–A
ZT Poisson	ZM Poisson	ZT negative binomial
ZM Poisson		ZM negative binomial
Geometric	Negative binomial ZM geometric	Geometric–Poisson
ZT geometric	ZT negative binomial	
ZM geometric	ZM negative binomial	
Logarithmic		ZT negative binomial
ZM logarithmic		ZM negative binomial
Binomial	ZM binomial	
Negative binomial	ZM negative binomial	Poisson–ETNB
Poisson–inverse Gaussian	Poisson–ETNB	
Polya–Aeppli	Poisson–ETNB	
Neyman–A		Poisson–ETNB

Appendix C Frequency and severity relationships

Let N^L be the number of losses random variable and let X be the severity random variable. If there is a deductible of d imposed, there are two ways to modify X. One is to create Y^L, the amount paid per loss:

$$Y^L = \begin{cases} 0, & X \le d, \\ X - d, & X > d. \end{cases}$$

In this case the appropriate frequency distribution continues to be N^L.

An alternative approach is to create Y^P, the amount paid per payment:

$$Y^P = \begin{cases} \text{undefined}, & X \le d, \\ X - d, & X > d. \end{cases}$$

In this case the frequency random variable must be altered to reflect the

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

number of payments. Let this variable be N^P. Assume that for each loss the probability is $v = 1 - F_X(d)$ that a payment will result. Further assume that the incidence of making a payment is independent of the number of losses. Then $N^P = L_1 + L_2 + \cdots + L_N$, where L_j is 0 with probability $1 - v$ and is 1 with probability v. Probability generating functions yield the following relationships:

N^L	Parameters for N^P
Poisson	$\lambda^* = v\lambda$
ZM Poisson	$p_0^{M*} = \dfrac{p_0^M - e^{-\lambda} + e^{-v\lambda} - p_0^M e^{-v\lambda}}{1 - e^{-\lambda}}$, $\lambda^* = v\lambda$
Binomial	$q^* = vq$
ZM binomial	$p_0^{M*} = \dfrac{p_0^M - (1-q)^m + (1-vq)^m - p_0^M(1-vq)^m}{1-(1-q)^m}$ $q^* = vq$
Negative binomial	$\beta^* = v\beta$, $r^* = r$
ZM neg. binomial	$p_0^{M*} = \dfrac{p_0^M - (1+\beta)^{-r} + (1+v\beta)^{-r} - p_0^M(1+v\beta)^{-r}}{1-(1+\beta)^{-r}}$ $\beta^* = v\beta$, $r^* = r$
ZM logarithmic	$p_0^{M*} = 1 - (1 - p_0^M)\ln(1+v\beta)/\ln(1+\beta)$ $\beta^* = v\beta$

The geometric distribution is not presented as it is a special case of the negative binomial with $r = 1$. For zero truncated distributions, the above is still used as the distribution for N^P will now be zero modified. For compound distributions, modify only the secondary distribution. For ETNB secondary distributions the parameter for the primary distribution is multiplied by $1 - p_0^{M*}$ as obtained above while the secondary distribution remains zero truncated (however, $\beta^* = v\beta$).

There are occasions in which frequency data are collected which provide a model for N^P. There would have to have been a deductible d in place and therefore v is available. It is possible to recover the distribution for N^L, although there is no guarantee that reversing the process will produce a legitimate probability distribution. The solutions are the same as above, only now $v = 1/[1 - F_X(d)]$.

Now suppose the current frequency model is N^d, which is appropriate for a deductible of d. Now suppose the deductible is to be changed to d^*. The new frequency for payments is N^{d^*} and is of the same type. Then use the table with $v = [1 - F_X(d^*)]/[1 - F_X(d)]$.

Appendix D
The recursive formula

The recursive formula is (where the frequency distribution is a member of the $(a, b, 1)$ class),

$$f_S(x) = \frac{[p_1 - (a+b)p_0]f_X(x) + \sum_{y=1}^{x\wedge m}\left(a + \frac{by}{x}\right) f_X(y) f_S(x-y)}{1 - af_X(0)},$$

where $f_S(x) = \Pr(S = x)$, $x = 0, 1, 2, \ldots$, $f_X(x) = \Pr(X = x)$, $x = 0, 1, 2, \ldots$, $p_0 = \Pr(N = 0)$, and $p_1 = \Pr(N = 1)$. Note that the severity distribution (X) must place probability on non-negative integers. The formula must be initialized with the value of $f_S(0)$. These values are given in Table D.1. It should be noted that, if N is a member of the $(a, b, 0)$ class, $p_1 - (a+b)p_0 = 0$ and so the first term will vanish. If N is a member of the compound class, the recursion must be run twice. The first pass uses the secondary distribution for p_0, p_1, a, and b. The second pass uses the output from the first pass as $f_X(x)$ and uses the primary distribution for p_0, p_1, a, and b.

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

Table D.1 Starting values ($f_S(0)$) for recursions

Distribution	$f_S(0)$
Poisson	$\exp[\lambda(f_0-1)]$
Geometric	$[1+\beta(1-f_0)]^{-1}$
Binomial	$[1+q(f_0-1)]^m$
Negative binomial	$[1+\beta(1-f_0)]^{-r}$
ZM Poisson	$p_0^M+(1-p_0^M)\frac{\exp(\lambda f_0)-1}{\exp(\lambda)-1}$
ZM geometric	$p_0^M+(1-p_0^M)\frac{f_0}{1+\beta(1-f_0)}$
ZM binomial	$p_0^M+(1-p_0^M)\frac{[1+q(f_0-1)]^m-(1-q)^m}{1-(1-q)^m}$
ZM negative binomial	$p_0^M+(1-p_0^M)\frac{[1+\beta(1-f_0)]^{-r}-(1+\beta)^{-r}}{1-(1+\beta)^{-r}}$
ZM logarithmic	$p_0^M+(1-p_0^M)\left\{1-\frac{\ln[1+\beta(1-f_0)]}{\ln(1+\beta)}\right\}$

Appendix E Discretization of the severity distribution

There are two relatively simple ways to discretize the severity distribution. One is the method of rounding and the other is a mean-preserving method.

E.1 THE METHOD OF ROUNDING

This method has two features: All probabilities are positive, and the probabilities add to 1. Let h be the span and let Y be the discretized version of X. If there are no modifications, then

$$\begin{aligned} f_j &= \Pr(Y=jh)=\Pr\left[\left(j-\tfrac{1}{2}\right)h\le X<\left(j+\tfrac{1}{2}\right)h\right] \\ &= F_X\left[\left(j+\tfrac{1}{2}\right)h\right]-F_X\left[\left(j-\tfrac{1}{2}\right)h\right]. \end{aligned}$$

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

The recursive formula is then used with $f_X(j) = f_j$. Suppose a deductible of d, limit of u, and coinsurance of α are to be applied. If the modifications are to be applied before the discretization, then

$$\begin{aligned} g_0 &= \frac{F_X(d+h/2) - F_X(d)}{1 - F_X(d)}, \\ g_j &= \frac{F_X[d+(j+1/2)h] - F_X[d+(j-1/2)h]}{1 - F_X(d)}, \\ & \qquad j = 1, \ldots, \frac{u-d}{h} - 1, \\ g_{(u-d)/h} &= \frac{1 - F_X(u-h/2)}{1 - F_X(d)}, \end{aligned}$$

where $g_j = \Pr(Z = j\alpha h)$ and Z is the modified distribution. This method does not require that the limits be multiples of h but does require that $u-d$ be a multiple of h. This method gives the probabilities of payments per payment.

Finally, if there is truncation from above at u, change all denominators to $F_X(u) - F_X(d)$ and also change the numerator of $g_{(u-d)/h}$ to $F_X(u) - F_X(u - h/2)$.

E.2 MEAN PRESERVING

This method ensures that the discretized distribution has the same mean as the original severity distribution. With no modifications the discretization is

$$\begin{aligned} f_0 &= 1 - \frac{\mathrm{E}[X \wedge h]}{h}, \\ f_j &= \frac{2\mathrm{E}[X \wedge jh] - \mathrm{E}[X \wedge (j-1)h] - \mathrm{E}[X \wedge (j+1)h]}{h}, \quad j = 1, 2, \ldots. \end{aligned}$$

For the modified distribution,

$$\begin{aligned} g_0 &= 1 - \frac{\mathrm{E}[X \wedge d + h] - \mathrm{E}[X \wedge d]}{h[1 - F_X(d)]}, \\ g_j &= \frac{2\mathrm{E}[X \wedge d + jh] - \mathrm{E}[X \wedge d + (j-1)h] - \mathrm{E}[X \wedge d + (j+1)h]}{h[1 - F_X(d)]}, \\ & \qquad j = 1, \ldots, \frac{u-d}{h} - 1, \\ g_{(u-d)/h} &= \frac{\mathrm{E}[X \wedge u] - \mathrm{E}[X \wedge u - h]}{h[1 - F_X(d)]}. \end{aligned}$$

To incorporate truncation from above, change the denominators to

$$h[F_X(u) - F_X(d)]$$

and subtract $h[1 - F_X(u)]$ from the numerators of each of g_0 and $g_{(u-d)/h}$.

E.3 UNDISCRETIZATION OF A DISCRETIZED DISTRIBUTION

Assume we have $g_0 = \Pr(S = 0)$, the true probability that the random variable is zero. Let $p_j = \Pr(S^* = jh)$, where S^* is a discretized distribution and h is the span. The following are approximations for the cdf and LEV of S, the true distribution which was discretized as S^*. They are all based on the assumption that S has a uniform distribution over the interval from $(j-\frac{1}{2})h$ to $(j+\frac{1}{2})h$ for integral j. The first interval is from 0 to $h/2$, and the probability $p_0 - g_0$ is assumed to be uniformly distributed over it. Let S^{**} be the random variable with this approximate mixed distribution. (It is continuous, except for discrete probability g_0 at zero.) The approximate distribution function can be found by interpolation as follows. First, let

$$F_j = F_{S^{**}}\left[\left(j + \tfrac{1}{2}\right)h\right] = \sum_{i=0}^{j} p_i, \quad j = 0, 1, \ldots.$$

Then, for x in the interval $(j - \frac{1}{2})h$ to $(j + \frac{1}{2})h$,

$$\begin{aligned} F_{S^{**}}(x) &= F_{j-1} + \int_{(j-1/2)h}^{x} h^{-1} p_j \, dt = F_{j-1} + \left[x - \left(j - \tfrac{1}{2}\right)h\right] h^{-1} p_j \\ &= F_{j-1} + \left[x - \left(j - \tfrac{1}{2}\right)h\right] h^{-1}(F_j - F_{j-1}) \\ &= (1-w)F_{j-1} + wF_j, \quad w = \frac{x}{h} - j + \tfrac{1}{2}. \end{aligned}$$

Because the first interval is only half as wide, the formula for $0 \le x \le h/2$ is

$$F_{S^{**}}(x) = (1-w)g_0 + wp_0, \quad w = \frac{2x}{h}.$$

It is also possible to express these formulas in terms of the discrete probabilities:

$$F_{S^{**}}(x) = \begin{cases} g_0 + \dfrac{2x}{h}[p_0 - g_0], & 0 < x \le \dfrac{h}{2}, \\ \displaystyle\sum_{i=0}^{j-1} p_i + \frac{x - (j-1/2)h}{h} p_j, & (j - \tfrac{1}{2})h < x \le (j + \tfrac{1}{2})h. \end{cases}$$

With regard to the limited expected value, expressions for the first and kth LEVs are

$$\mathrm{E}(S^{**} \wedge x) = \begin{cases} x(1-g_0) - \dfrac{x^2}{h}(p_0 - g_0), & 0 < x \le \dfrac{h}{2}, \\ \dfrac{h}{4}(p_0 - g_0) + \displaystyle\sum_{i=1}^{j-1} ihp_i + \frac{x^2 - [(j-1/2)h]^2}{2h} p_j & \\ \qquad + x[1 - F_{S^{**}}(x)], & (j - \tfrac{1}{2})h < x \le (j + \tfrac{1}{2})h, \end{cases}$$

and, for $0 < x \leq \frac{h}{2}$,

$$\mathrm{E}[(S^{**} \wedge x)^k] = \frac{2x^{k+1}}{h(k+1)}(p_0 - g_0) + x^k[1 - F_{S^{**}}(x)],$$

while for $(j - \frac{1}{2})h < x \leq (j + \frac{1}{2})h$,

$$\begin{aligned}\mathrm{E}[(S^{**} \wedge x)^k] &= \frac{(h/2)^k(p_0 - g_0)}{k+1} + \sum_{i=1}^{j-1} \frac{h^k[(i+\frac{1}{2})^{k+1} - (i-\frac{1}{2})^{k+1}]}{k+1} p_i \\ &\quad + \frac{x^{k+1} - [(j-\frac{1}{2})h]^{k+1}}{h(k+1)} p_j + x^k[1 - F_{S^{**}}(x)].\end{aligned}$$

Appendix F
Numerical optimization and solution of systems of equations

Maximizing functions can be difficult when there are many variables. A variety of numerical methods have been developed, and most any will be sufficient for the tasks set forth in this text. Here we present two options. The first is to use the Excel® Solver add-in. It is fairly reliable, though at times it may declare a maximum has been found when there is no maximum. A second option is the simplex method. This method tends to be slower but is more reliable. The final section of this Appendix shows how the solver and goal seek routines in Excel® can be used to solve systems of equations.

Loss Models: From Data to Decisions, Second Edition.
By Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot
ISBN 0-471-21577-5

F.1 MAXIMIZATION USING SOLVER

Solver is not automatically available when Excel® is installed. If it is available, you can tell because *Solver* will appear on Excel's Tools menu. If it does not, it must be added in. To do this, select *Add-ins* from the Tools menu, check the Solver box, and then click OK. If Solver does not appear on the add-in list, Solver was not installed when Excel® was installed on your machine. This will be the case if a typical (as opposed to full or custom) install was done. To install Solver, go to Add/Remove Programs in the Control Panel and modify your Microsoft Office® installation. You will not need to reinstall all of Office® to add the Solver.

Use of Solver is illustrated with an example in which maximum likelihood estimates for the gamma distribution are found for Data Set B right censored at 200. If you have not read far enough to appreciate this example, it is not important.

Begin by setting up a spreadsheet in which the parameters (alpha and theta) are in identifiable cells as is the objective function (lnL). In this example the parameters are in E1 and E2 and the objective function is in E3.[1]

Microsoft Excel - solver illustration.xls

File Edit View Insert Format Tools Data Window Help

E3 =SUM(C2:C8)

	A	B	C	D	E	F
1	x	"f(x)"	ln	alpha	1	
2	27	0.000973	-6.93476	theta	1000	
3	82	0.000921	-6.98976	lnL	-44.9125	
4	115	0.000891	-7.02276			
5	126	0.000882	-7.03376			
6	155	0.000856	-7.06276			
7	161	0.000851	-7.06876			
8	200	0.818731	-2.8			
9						
10						
11						
12						
13						

Sheet1 Sheet2 Sheet3

Ready

The formulas underlying this spreadsheet are given below.

[1] Screenshots reprinted by permission from Microsoft Corporation.

Microsoft Excel - solver illustration.xls

File Edit View Insert Format Tools Data Window Help

E7

	A	B	C	D	E
1	x	"f(x)"	ln	alpha	1
2	27	=GAMMADIST(A2,E$1,E$2,FALSE)	=LN(B2)	theta	1000
3	82	=GAMMADIST(A3,E$1,E$2,FALSE)	=LN(B3)	lnL	=SUM(C2:C8)
4	115	=GAMMADIST(A4,E$1,E$2,FALSE)	=LN(B4)		
5	126	=GAMMADIST(A5,E$1,E$2,FALSE)	=LN(B5)		
6	155	=GAMMADIST(A6,E$1,E$2,FALSE)	=LN(B6)		
7	161	=GAMMADIST(A7,E$1,E$2,FALSE)	=LN(B7)		
8	200	=1-GAMMADIST(200,E1,E2,TRUE)	=14*LN(B8)		
9					
10					

Sheet1 Sheet2 Sheet3

Ready

Note that trial values for alpha and theta have been entered (1 and 1,000). The better these guesses are, the higher the probability that Solver will succeed in finding the maximum. Selecting Solver from the Tools menu brings up the following dialog box:

Solver Parameters

Set Target Cell: E3

Equal To: Max Min Value of: 0

By Changing Cells:

E1,E2

Guess

Subject to the Constraints:

Add

Change

Delete

Solve

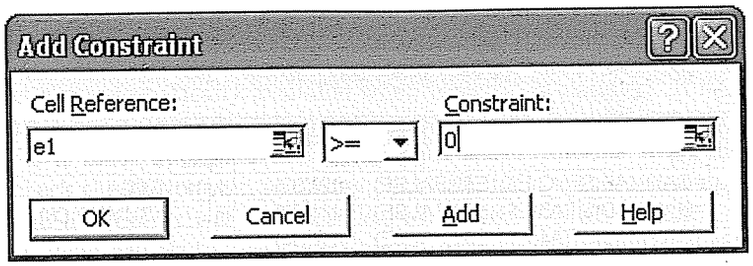
Close

Options

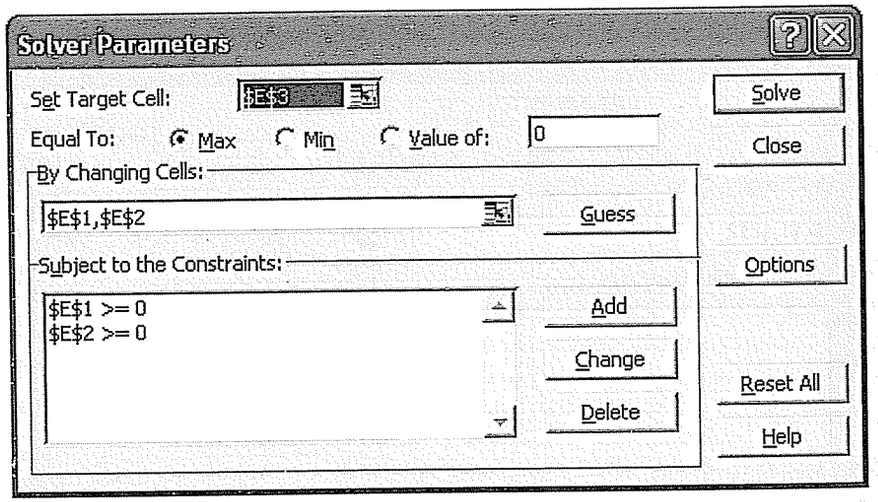
Reset All

Help

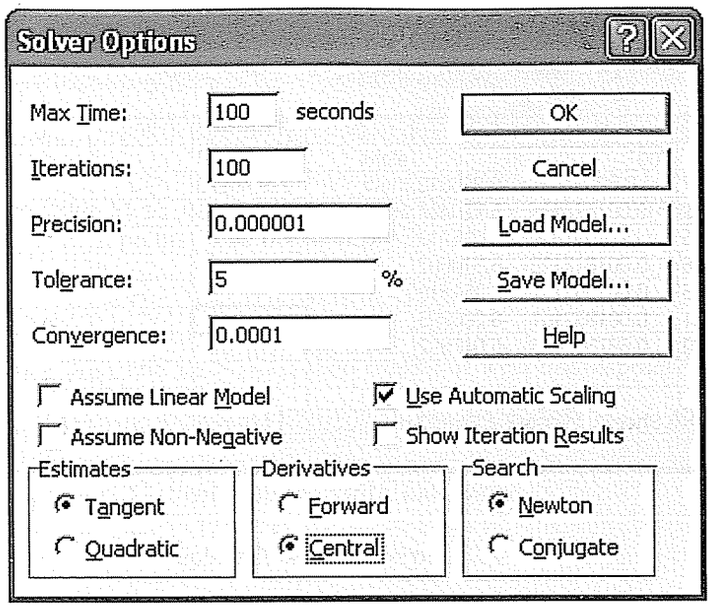
The target cell is the location of the objective function and the *By Changing Cells* box contains the location of the parameters. These cells need not be contiguous. It turns out that clicking on *Solve* will get the job done, but there are two additional items to think about. First, Solver allows for the imposition of constraints. They can be added by clicking on *Add* which brings up the following dialog box:



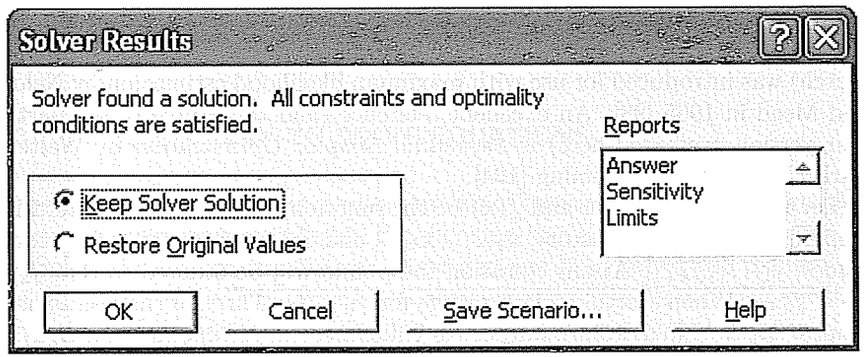
The constraint $\alpha \geq 0$ has been entered. Solver does not allow the constraint we really want, which is $\alpha > 0$. After entering a similar constraint for θ, the Solver dialog box looks like:



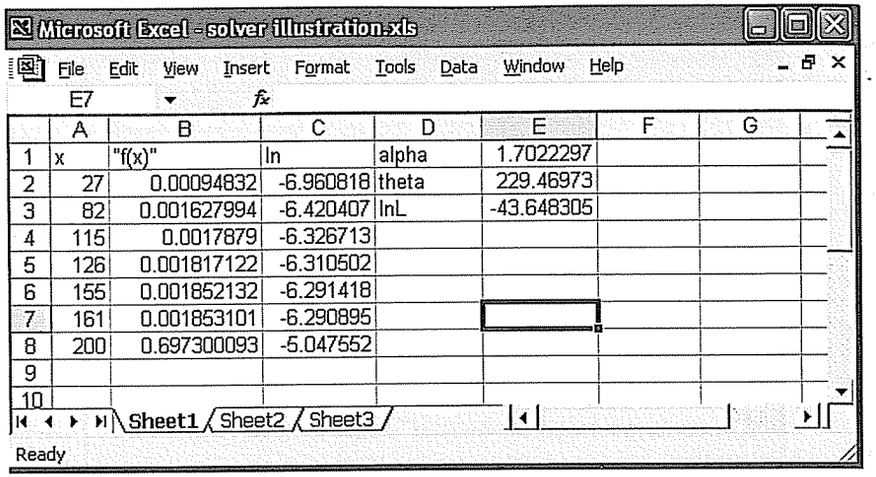
The reason adding the constraints is not needed here is that the solution Solver finds meets the constraints anyway. Clicking on *Options* brings up the following dialog box:



Two changes have been made from the default settings. The *Use Automatic Scaling* box has been checked. This improves performance when the parameters are on different scales (as is the case here). Also, *Central* approximate derivatives have been selected. Additional precision in the answer can be obtained by making the Precision, Tolerance, and Convergence numbers smaller. Clicking *OK* on the options box (no changes will be apparent in the Solver box) and then clicking *Solve* results in the following:



Clicking *OK* gives the answer.



	A	B	C	D	E	F	G
1	x	"f(x)"	ln	alpha	1.7022297		
2	27	0.00094832	-6.960818	theta	229.46973		
3	82	0.001627994	-6.420407	lnL	-43.648305		
4	115	0.0017879	-6.326713				
5	126	0.001817122	-6.310502				
6	155	0.001852132	-6.291418				
7	161	0.001853101	-6.290895				
8	200	0.697300093	-5.047552				
9							
10							

Users of Solver (or any numerical analysis routine) should always be wary of the results. The program may announce a solution when the maximum has not been found, and it may give up when there is a maximum to be found. When the program gives up, it may be necessary to provide better starting values. To verify that an announced solution is legitimate (or at least is a local maximum), it is a good idea to check the function at nearby points to see that the values are indeed smaller.

F.2 THE SIMPLEX METHOD

The method (which is not related to the simplex method from operations research) was introduced for use with maximum likelihood estimation by Nelder and Mead in 1965 [98]. An excellent reference (and the source of the particular version presented here) is *Sequential Simplex Optimization* by Walters, Parker, Morgan, and Deming [134].

Let $\mathbf{x}$ be a $k \times 1$ vector and $f(\mathbf{x})$ be the function in question. The iterative step begins with $k+1$ vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_{k+1}$, and the corresponding functional values, $f_1, \ldots, f_{k+1}$. At any iteration the points will be ordered so that $f_2 < \cdots < f_{k+1}$. When starting, also arrange for $f_1 < f_2$. Three of the points have names: $\mathbf{x}_1$ is called *worstpoint*, $\mathbf{x}_2$ is called *secondworstpoint*, and $\mathbf{x}_{k+1}$ is called *bestpoint*. It should be noted that after the first iteration these names may not perfectly describe the points. Now identify five new points. The first one, $\mathbf{y}_1$, is the center of $\mathbf{x}_2, \ldots, \mathbf{x}_{k+1}$, That is, $\mathbf{y}_1 = \sum_{j=2}^{k+1} \mathbf{x}_j/k$ and is called *midpoint*. The other four points are found as follows:

$$\begin{aligned} \mathbf{y}_2 &= 2\mathbf{y}_1 - \mathbf{x}_1, && \textit{refpoint}, \\ \mathbf{y}_3 &= 2\mathbf{y}_2 - \mathbf{x}_1, && \textit{doublepoint}, \\ \mathbf{y}_4 &= (\mathbf{y}_1 + \mathbf{y}_2)/2, && \textit{halfpoint}, \\ \mathbf{y}_5 &= (\mathbf{y}_1 + \mathbf{x}_1)/2, && \textit{centerpoint}. \end{aligned}$$

Then let $g_2, \ldots, g_5$ be the corresponding functional values, that is, $g_j = f(\mathbf{y}_j)$ (the value at $\mathbf{y}_1$ is never used). The key is to replace *worstpoint* ($\mathbf{x}_1$) with one of these points. The decision process proceeds as follows:

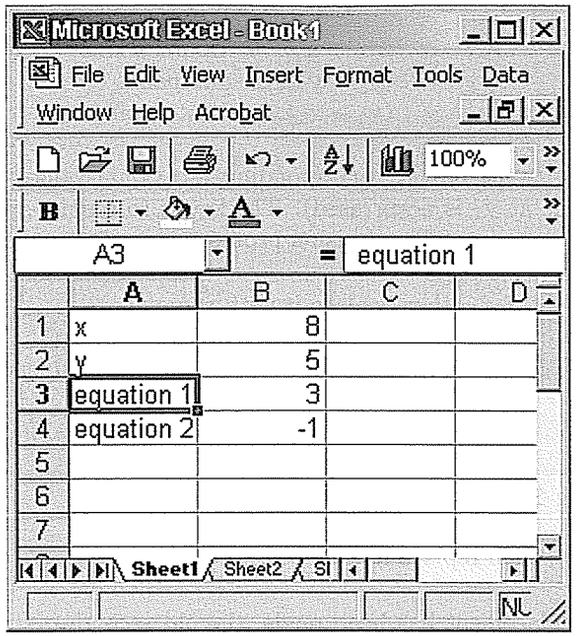
1. If $f_2 < g_2 < f_{k+1}$, then replace it with *refpoint*.
2. If $g_2 \geq f_{k+1}$ and $g_3 > f_{k+1}$, then replace it with *doublepoint*.
3. If $g_2 \geq f_{k+1}$ and $g_3 \leq f_{k+1}$, then replace it with *refpoint*.
4. If $f_1 < g_2 \leq f_2$, then replace it with *halfpoint*.
5. If $g_2 \leq f_1$, then replace it with *centerpoint*.

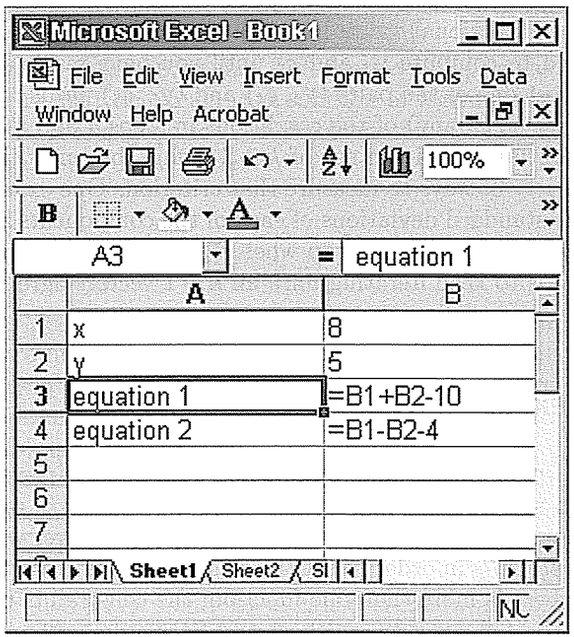
After the replacement has been made, the old *secondworstpoint* becomes the new *worstpoint*. The remaining k points are then ordered. The one with the smallest functional value becomes the new *secondworstpoint*, and the one with the largest functional value becomes the new *bestpoint*. In practice, there is no need to compute $\mathbf{y}_3$ and g_3 until you have reached step 2. Also note that at most one of the pairs $(\mathbf{y}_4, g_4)$ and $(\mathbf{y}_5, g_5)$ needs to be obtained, depending on which (if any) of the conditions in steps 4 and 5 hold.

Iterations continue until the set of $k+1$ points becomes tightly packed. There are a variety of ways to measure that criterion. One example would be to calculate the standard deviations of each of the components and then average those values. Iterations can stop when a small enough value is obtained. Another option is to keep iterating until all $k+1$ vectors agree to a specified number of significant digits.

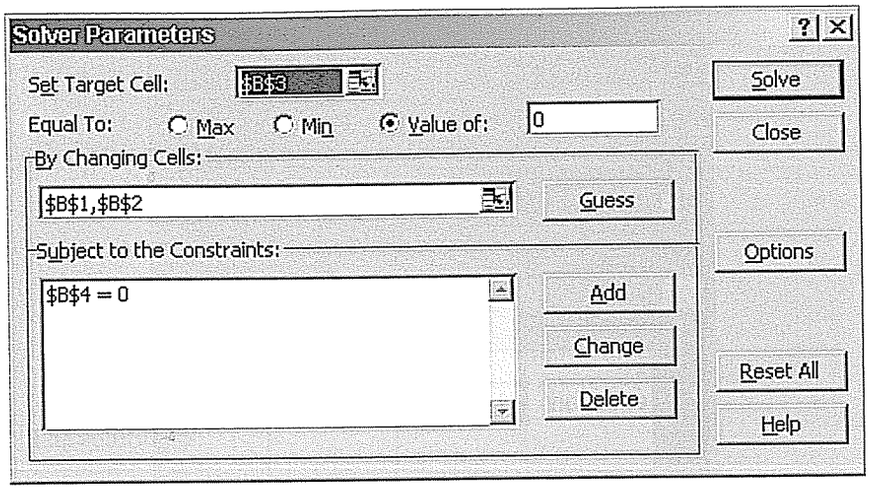
F.3 USING EXCEL® TO SOLVE EQUATIONS

In addition to maximizing and minimizing functions of several variables, Solver can also solve equations. By choosing the *Value of:* radio button in the Solver dialog box, a value can be entered and then Solver will manipulate the *By Changing Cells* in order to set the contents of the *Target Cell* equal to that value. If there is more than one function, the constraints can be used to set them up. The following spreadsheet and Solver dialog box are set up to solve the two equations $x + y = 10$ and $x - y = 4$ with starting values $x = 8$ and $y = 5$ (to illustrate that the starting values do not have to be solutions to any of the equations).

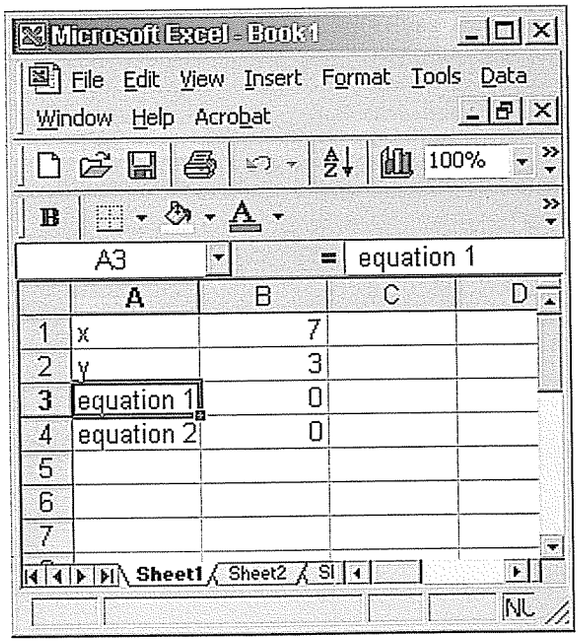




The Solver dialog box is:

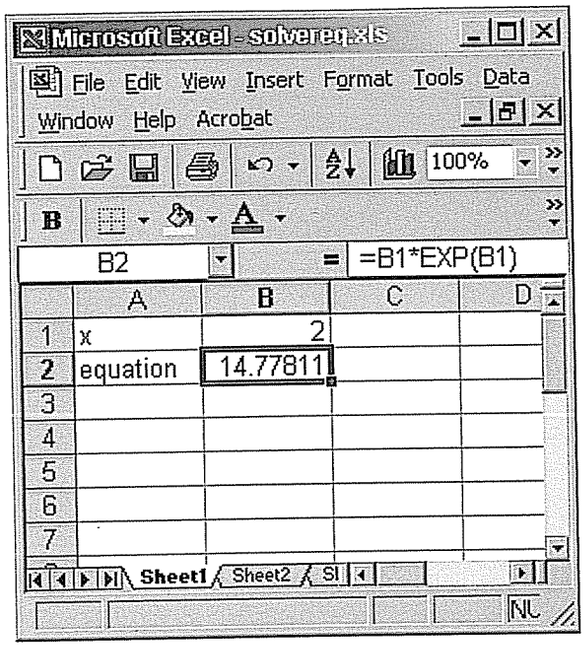


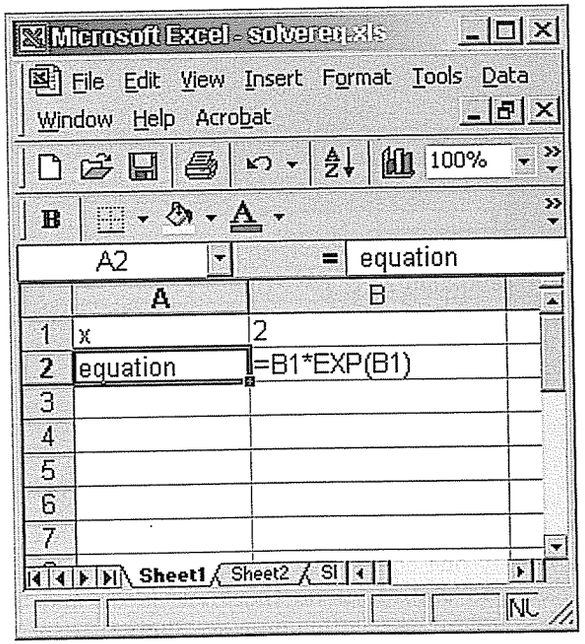
The solution is:



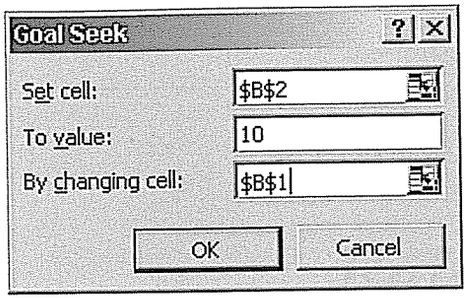
When there is only one equation with one unknown, the Goal Seek tool in Excel® is easier to use. It is on the Tools menu and is most always installed with the standard installation process. Suppose we want the solution

of $xe^x = 10$. The following simple spreadsheet sets up the problem with a starting value of $x = 2$.

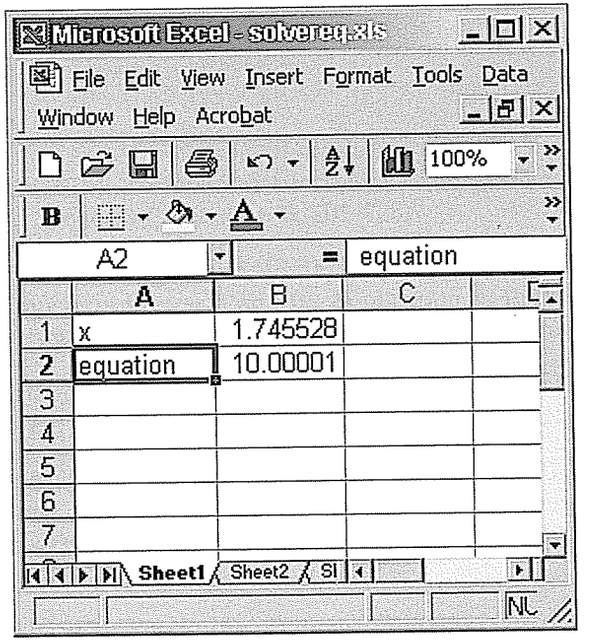




The Goal Seek dialog box is:



The solution is:



References

1. Aalen, O. (1978), "Nonparametric Inference for a Family of Counting Processes," *Annals of Statistics*, **6**, 701–726.

2. Abate, J., Choudhury, G., and Whitt, W. (2000), "An Introduction to Numerical Transform Inversion and Its Application to Probability Models," in W. Grassman, ed., *Computational Probability*, Boston: Kluwer.

3. Abramowitz, M. and Stegun, I. (1964), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Wiley.

4. Accomando, F. and Weissner, E. (1988), "Report Lag Distributions: Estimation and Application to IBNR Counts," in *Transcripts of the 1988 Casualty Loss Reserve Seminar*, Arlington, VA: Casualty Actuarial Society, 1038–1133.

5. Arnold, B. (1983), *Pareto Distributions (Statistical Distributions in Scientific Work)*, Vol. 5, Fairland, MD: International Co-operative Publishing House.

6. Bailey, A. (1942), "Sampling Theory in Casualty Insurance, Parts I and II," *Proceedings of the Casualty Actuarial Society*, **XXIX**, 50–95.

7. Bailey, A. (1943), "Sampling Theory in Casualty Insurance, Parts III through VII," *Proceedings of the Casualty Actuarial Society*, **XXX**, 31–65.

8. Bailey, A. (1950), "Credibility Procedures," *Proceedings of the Casualty Actuarial Society*, **XXXVII**, 7–23 and 94–115.

9. Bailey, W. (1992), "A Method for Determining Confidence Intervals for Trend," *Transactions of the Society of Actuaries*, **XLIV**, 11–54.

10. Baker, C. (1977), *The Numerical Treatment of Integral Equations*, Oxford: Clarendon Press.

11. Batten, R. (1978), *Mortality Table Construction*, Englewood Cliffs, NJ: Prentice-Hall.

12. Beard, R., Pentikainen, T., and Pesonen, E. (1984), *Risk Theory*, 3rd ed., London: Chapman & Hall.

13. Berger, J. (1985), *Bayesian Inference in Statistical Analysis*, 2nd ed., New York: Springer-Verlag.

14. Bertram, J. (1981), "Numerische Berechnumg von Gesamtschadenverteilungen," *Blätter der deutschen Gesellschaft Versicherungsmathematik*, B. **15.2**, 175–194.

15. Bevan, J. (1963), "Comprehensive Medical Insurance—Statistical Analysis for Ratemaking," *Proceedings of the Casualty Actuarial Society*, **L**, 111–128.

16. Bowers, N., Gerber, H., Hickman, J., Jones, D., and Nesbitt, C. (1986), *Actuarial Mathematics*, Schaumburg, IL: Society of Actuaries.

17. Brockett, P. (1991), "Information Theoretic Approach to Actuarial Science: A Unification and Extension of Relevant Theory and Applications," with discussion, *Transactions of the Society of Actuaries*, **XLIII**, 73–135.

18. Bühlmann, H. (1967), "Experience Rating and Credibility," *ASTIN Bulletin*, **4**, 199–207.

19. Bühlmann, H. (1970), *Mathematical Methods in Risk Theory*, New York: Springer-Verlag.

20. Bühlmann, H. and Straub, E. (1970), "Glaubwürdigkeit für Schadensätze (credibility for loss ratios)," *Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker*, **70**, 111–133.

21. Carlin, B. and Klugman, S. (1993), "Hierarchical Bayesian Whitaker Graduation," *Scandinavian Actuarial Journal*, 183–196.

22. Carlin, B. and Louis, T. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Boca Raton, FL: CRC Press.

23. Carriere, J. (1993), "Nonparametric Estimators of a Distribution Function Based on Mixtures of Gamma Distributions," *Actuarial Research Clearing House*, **1993.3**, 1–11.

24. Casualty Actuarial Society (1990), *Foundations of Casualty Actuarial Science*, Arlington, VA: Casualty Actuarial Society.

25. deAlba, E. (2002), "Bayesian Estimation of Outstanding Claim Reserves," *North American Actuarial Journal*, **6**, 1–20.

26. DePril, N. (1986), "On the Exact Computation of the Aggregate Claims Distribution in the Individual Life Model," *ASTIN Bulletin*, **16**, 109–112.

27. DePril, N. (1988), "Improved Approximations for the Aggregate Claims Distribution of a Life Insurance Portfolio," *Scandinavian Actuarial Journal*, 61–68.

28. DePril, N. (1989), "The Aggregate Claim Distribution in the Individual Model with Arbitrary Positive Claims," *ASTIN Bulletin*, **19**, 9–24.

29. Douglas, J. (1980), *Analysis with Standard Contagious Distributions*, Fairland, MD: International Co-operative Publishing House.

30. Dropkin, L. (1959), "Some Considerations on Automobile Rating Systems Utilizing Individual Driving Records," *Proceedings of the Casualty Actuarial Society*, **XLVI**, 165–176.

31. Efron, B. (1981), "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, **76**, 321–319.

32. Efron, B. (1986), "Why Isn't Everyone a Bayesian?" *The American Statistician*, **40**, 1–11 (including comments and reply).

33. Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

34. Ericson, W. (1969), "A Note on the Posterior Mean of a Population Mean," *Journal of the Royal Statistical Society, Series B*, **31**, 332–334.

35. Feller, W. (1968), *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed. rev., New York: Wiley.

36. Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd ed., New York: Wiley.

37. Fisher, A. (1915), "Note on the Application of Recent Mathematical-Statistical Methods to Coal Mine Accidents, with Special Reference to Catastrophes in Coal Mines in the United States," *Proceedings of the Casualty Actuarial Society*, **II**, 70–78.

38. Fisz, M. (1963), *Probability Theory and Mathematical Statistics*, New York: Wiley.

39. Frees, E., Carriere, J., and Valdez, E. (1996), "Annuity Valuation with Dependent Mortality," *Journal of Risk and Insurance*, **63**, 229–261.

40. Frees, E. and Valdez, E. (1998) "Understanding Relationships Using Copulas," *North American Actuarial Journal*, **2**, 1–25.

41. Genest, C. (1987), "Frank's Family of Bivariate Distributions," *Biometrika*, **74**, 549–555.

42. Genest, C. and McKay, J. (1986), "The Joy of Copulas: Bivariate Distributions with Uniform Marginals," *The American Statistician*, **40**, 280–283.

43. Gerber, H. (1982), "On the Numerical Evaluation of the Distribution of Aggregate Claims and Its Stop-Loss Premiums," *Insurance: Mathematics and Economics*, **1**, 13–18.

44. Gerber, H. and D. Jones (1976), "Some Practical Considerations in Connection with the Calculation of Stop-Loss Premiums," *Transactions of the Society of Actuaries*, **XXVIII**, 215–231.

45. Gillam, W. (1992), "Parametrizing the Workers Compensation Experience Rating Plan," *Proceedings of the Casualty Actuarial Society*, **LXXIX**, 21–56.

46. Goovaerts, M. J. and Hoogstad, W. J. (1987), *Credibility Theory, Surveys of Actuarial Studies No. 4*, Rotterdam: Nationale-Nederlanden.

47. Guiahi, F. (2001), "Fitting to Loss Distributions with Emphasis on Rating Variables," *CAS Forum*, **Winter 2001**, 133–174.

48. Hachemeister, C .A. (1975), "Credibility for Regression Models with Application to Trend," in P. Kahn, ed., *Credibility: Theory and Applications*, New York: Academic Press, 129–163.

49. Harwayne, F. (1959), "Merit Rating in Private Passenger Automobile Liability Insurance and the California Driver Record Study," *Proceedings of the Casualty Actuarial Society*, **XLVI**, 189–195.

50. Hayne, R. (1994), "Extended Service Contracts," *Proceedings of the Casualty Actuarial Society*, **LXXXI**, 243–302.

51. Heckman, P. and G. Meyers (1983), "The Calculation of Aggregate Loss Distributions from Claim Severity and Claim Count Distributions," *Proceedings of the Casualty Actuarial Society*, **LXX**, 22–61.

52. Herzog, T. (1999), *Introduction to Credibility Theory*, 3rd ed., Winsted, CT: ACTEX.

53. Herzog, T. and Lord, G. (2002), *Applications of Monte Carlo Methods to Finance and Insurance*, Winsted, CT: ACTEX.

54. Herzog, T. and Laverty, J. (1995), "Experience of Refinanced FHA Section 203(b) Single Family Mortages," *Actuarial Research Clearing House*, **1995.1**, 97–129.

55. Hewitt, C., Jr. (1967), "Loss Ratio Distributions—A Model," *Proceedings of the Casualty Actuarial Society*, **LIV**, 70–88.

56. Hewitt, C., Jr., and Lefkowitz, B. (1979), "Methods for Fitting Distributions to Insurance Loss Data," *Proceedings of the Casualty Actuarial Society*, **LXVI**, 139–160.

57. Hipp, G. (1938), "Special Funds Under the New York Workmen's Compensation Law," *Proceedings of the Casualty Actuarial Society*, **XXIV**, 247–275.

58. Hogg, R. and Craig, A. (1978), *Introduction to Mathematical Statistics*, 4th ed., New York: Macmillan.

59. Hogg, R. and Klugman, S. (1984), *Loss Distributions*, New York: Wiley.

60. Holgate, P. (1970), "The Modality of Some Compound Poisson Distributions," *Biometrika*, **57**, 666–667.

61. Holler, K., Sommer, D., and Trahair, G. (1999), "Something Old, Something New in Classification Ratemaking with a Novel Use of Generalized Linear Models for Credit Insurance," *CAS Forum*, **Winter 1999**, 31–84.

62. Hossack, I., Pollard, J., and Zehnwirth, B. (1983), *Introductory Statistics with Applications in General Insurance*, Cambridge: Cambridge University Press.

63. Hougaard, P. (2000), *Analysis of Multivariate Survival Data*, New York: Springer-Verlag.

64. Hutchinson, T. and Lai, C. (1990), *Continuous Bivariate Distributions, Emphasizing Applications*, Adelaide: Rumsby.

65. Hyndman, R. and Fan, Y. (1996), "Sample Quantiles in Statistical Packages," *The American Statistician*, **50**, 361–365.

66. Jewell, W. (1974), "Credibility Is Exact Bayesian for Exponential Families," *ASTIN Bulletin*, **8**, 77–90.

67. Johnson, N., Kotz, S., and Balakrishnan, N. (1994), *Continuous Univariate Distributions*, Vol. 1, 2nd ed., New York: Wiley.

68. Johnson, N., Kotz, S., and Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 2, 2nd ed., New York: Wiley.

69. Johnson, N., Kotz, S., and Kemp, A. (1993), *Univariate Discrete Distributions*, 2nd ed., New York: Wiley.

70. Kaplan, E. and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, **53**, 457–481.

71. Karlin, S. and Taylor, H. (1975), *A First Course in Stochastic Processes*, 2nd ed., New York: Academic Press.

72. Karlin, S. and Taylor, H. (1981), *A Second Course in Stochastic Processes*, New York: Academic Press.

73. Kleiber, C. and Kotz, S. (2003), *Statistical Size Distributions in Economics and Actuarial Sciences*, New York: Wiley.

74. Klein, J. and Moeschberger, M. (1997), *Survival Analysis, Techniques for Censored and Truncated Data*, New York: Springer-Verlag.

75. Klugman, S. (1981), "On the Variance and Mean Squared Error of Decrement Estimators," *Transactions of the Society of Actuaries*, **XXXIII**, 301–311.

76. Klugman, S. (1987), "Credibility for Classification Ratemaking Via the Hierarchical Linear Model," *Proceedings of the Casualty Actuarial Society*, **LXXIV**, 272–321.

77. Klugman, S. (1992), *Bayesian Statistics in Actuarial Science with Emphasis on Credibility*, Boston: Kluwer.

78. Klugman, S. and Parsa, A. (1999), "Fitting Bivariate Distributions with Copulas," *Insurance: Mathematics and Economics*, **24**, 139–148.

79. Kornya, P. (1983), "Distribution of Aggregate Claims in the Individual Risk Model," *Transactions of the Society of Actuaries*, **XXXV**, 837–858.

80. Kotz, S., Balakrishnan, N., and Johnson, N. (2000), *Continuous Multivariate Distributions*, Vol. 1, Models and Applications, New York: Wiley.

81. Lawless, J. (2003), *Statistical Models and Methods for Lifetime Data*, 2nd ed., New York: Wiley.

82. Lemaire, J. (1995), *Automobile Insurance: Actuarial Models*, 2nd ed., Boston: Kluwer.

83. Lindley, D. (1987), "The Probability Approach to the Treatment of Uncertainty in Artificial Intelligence and Expert Systems," *Statistical Science*, **2**, 17–24 (also related articles in that issue).

84. London, D. (1985), *Graduation: The Revision of Estimates*, Winsted, CT: ACTEX.

85. London, D. (1988), *Survival Models and Their Estimation*, 3rd ed., Winsted, CT: ACTEX.

86. Longley-Cook, L. (1958), "The Employment of Property and Casualty Actuaries," *Proceedings of the Casualty Actuarial Society*, **XLV**, 9–10.

87. Longley-Cook, L. (1962), "An Introduction to Credibility Theory," *Proceeding of the Casualty Actuarial Society*, **XLIX**, 194–221.

88. Luong, A. and Doray, L. (1996), "Goodness of Fit Test Statistics for the Zeta Family," *Insurance: Mathematics and Economics*, **10**, 45–53.

89. Mardia, K. (1970), *Families of Bivariate Distributions*, London: Griffin.

90. McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, New York: Chapman & Hall.

91. Meyers, G. (1984), "Empirical Bayesian Credibility for Workers' Compensation Classification Ratemaking," *Proceedings of the Casualty Actuarial Society*, **LXXI**, 96–121.

92. Meyers, G. (1994), "Quantifying the Uncertainty in Claim Severity Estimates for an Excess Layer When Using the Single Parameter Pareto," *Proceedings of the Casualty Actuarial Society*, **LXXXI**, 91–122 (including discussion).

93. Mildenhall, S. (1999), "A Systematic Relationship between Minimum Bias and Generalized Linear Models," *Proceedings of the Casualty Actuarial Society*, **LXXXVI**, 393–487.

94. Miller, M. (1949), *Elements of Graduation*, Philadelphia: The Actuarial Society of America and the American Institute of Actuaries.

95. Moore, D. (1986), "Tests of Chi-Squared Type," in D'Agostino, R. and Stephens, M., eds., *Goodness-of-Fit Techniques*, New York: Marcel Dekker, 63–95.

96. Mowbray, A. H. (1914), "How Extensive a Payroll Exposure Is Necessary to Give a Dependable Pure Premium?" *Proceedings of the Casualty Actuarial Society*, **I**, 24–30.

97. Murphy, K., Brockman, M., and Lee, P. (2000), "Using Generalized Linear Models to Build Dynamic Pricing Systems for Personal Lines Insurance," *CAS Forum*, **Winter 2000**, 107–140.

98. Nelder, J. and Mead, U. (1965), "A Simplex Method for Function Minimization," *The Computer Journal*, **7**, 308–313.

99. Nelson, W. (1972), "Theory and Applications of Hazard Plotting for Censored Failure Data," *Technometrics*, **14**, 945–965.

100. Norberg, R. (1979), "The Credibility Approach to Experience Rating," *Scandinavian Actuarial Journal*, 181–221.

101. Ntzoufras, I. and Dellaportas, P. (2002), "Bayesian Modeling of Oustanding Liabilities Incorporating Claim Count Uncertainty," *North American Actuarial Journal*, **6**, 113–128.

102. Patrik, G. (1980), "Estimating Casualty Insurance Loss Amount Distributions," *Proceedings of the Casualty Actuarial Society*, **LXVII**, 57–109.

103. Panjer, H. and Lutek, B. (1983), "Practical Aspects of Stop-Loss Calculations," *Insurance: Mathematics and Economics*, **2**, 159–177.

104. Panjer, H. and Wang, S. (1993), "On the Stability of Recursive Formulas," *ASTIN Bulletin*, **23**, 227–258.

105. Panjer, H. and Willmot, G. (1986), "Computational Aspects of Recursive Evaluation of Compound Distributions," *Insurance: Mathematics and Economics*, **5**, 113–116.

106. Panjer, H. and Willmot, G. (1992), *Insurance Risk Models*, Chicago: Society of Actuaries.

107. Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1988), *Numerical Recipes in C*, Cambridge: Cambridge University Press.

108. Rao, C. (1965), *Linear Statistical Inference and Its Applications*, New York: Wiley.

109. Rioux, J. and Klugman S. (2003), "Toward a Unified Approach to Fitting Loss Models," working paper.

110. Ripley, B. (1987), *Stochastic Simulation*, New York: Wiley.

111. Robertson, J. (1992), "The Computation of Aggregate Loss Distributions," *Proceedings of the Casualty Actuarial Society*, **LXXIX**, 57–133.

112. Rohatgi, V. (1976), *An Introduction to Probability Theory and Mathematical Statistics*, New York: Wiley.

113. Rolski, T., Schmidli, H., Schmidt, V., and Teugels, J. (1999), *Stochastic Processes for Insurance and Finance*, Chichester: Wiley.

114. Ross, S. (1996), *Stochastic Processes*, 2nd ed., New York: Wiley.

115. Ross, S. (2002), *Simulation*, 3rd ed., San Diego: Academic Press.

116. Ross, S. (2003), *Introduction to Probability Models*, 8th ed., San Diego: Academic Press.

117. Schoenberg, I. (1964), "Spline Functions and the Problem of Graduation," *Proceedings of the National Academy of Science*, **52**, 947–950.

118. Scollnik, D. (2001), "Actuarial Modeling with MCMC and BUGS," *North American Actuarial Journal*, **5**, 96–124.

119. Scollnik, D. (2002), "Modeling Size-of-Loss Distributions for Exact Data in WinBUGS," *Journal of Actuarial Practice*, **10**, 193–218.

120. Self, S. and Liang, K. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, **82**, 605–610.

121. Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, **6**, 461–464.

122. Simon, L. (1961), "Fitting Negative Binomial Distributions by the Method of Maximum Likelihood," *Proceedings of the Casualty Actuarial Society*, **XLVIII**, 45–53.

123. Society of Actuaries Committee on Actuarial Principles (1992), "Principles of Actuarial Science," *Transactions of the Society of Actuaries*, **XLIV**, 565–628.

124. Society of Actuaries Committee on Actuarial Principles (1995), "Principles Regarding Provisions for Life Risks," *Transactions of the Society of Actuaries*, **XLVII**, 775–793.

125. Stephens, M. (1986), "Tests Based on EDF Statistics," in D'Agostino, R. and Stephens, M., eds., *Goodness-of-Fit Techniques*, New York: Marcel Dekker, 97–193.

126. Sundt, B. (1986), Special issue on credibility theory, *Insurance: Abstracts and Reviews*, **2**.

127. Sundt, B. (1999), *An Introduction to Non-Life Insurance Mathematics*, 4th ed., Mannheim: University of Mannheim Press.

128. Thyrion, P. (1961), "Contribution a l'Etude du Bonus pour non Sinstre en Assurance Automobile," *ASTIN Bulletin*, **1**, 142–162.

129. Tijms, H. (1994), *Stochastic Models—An Algorithmic Approach*, Chichester: Wiley.

130. Tröbliger, A. (1961), "Mathematische Untersuchungen zur Beitragsruckgewahr in der Kraftfahrversicherung," *Blatter der Deutsche Gesellschaft fur Versicherungsmathematik*, **5**, 327–348.

131. Tukey, J. (1962), "The future of data analysis," *Annals of Mathematical Statistics*, **33**, 1–67.

132. Venter, G. (1983), "Transformed Beta and Gamma Distributions and Aggregate Losses," *Proceedings of the Casualty Actuarial Society*, **LXX**, 156–193.

133. Verrall, R. (1990), "Bayes and Empirical Bayes Estimation for the Chain Ladder Method," *ASTIN Bulletin*, **20**, 217–243.

134. Walters, F., Parker, L., Morgan, S., and Deming, S. (1991), *Sequential Simplex Optimization*, Boca Raton, FL: CRC Press.

135. Waters, H. R. (1993), *Credibility Theory*, Edinburgh: Department of Actuarial Mathematics & Statistics, Heriot-Watt University.

136. Whitney, A.W. (1918), "The Theory of Experience Rating," *Proceedings of the Casualty Actuarial Society*, **IV**, 274–292.

137. Willmot, G. (1998), "On a Class of Approximations for Ruin and Waiting Time Probabilities," *Operations Research Letters*, **22**, 27–32.

Index

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

*BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Third Edition*
BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BERNARDO and SMITH · Bayesian Theory
BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
BILLINGSLEY · Probability and Measure, *Third Edition*
BIRKES and DODGE · Alternative Methods of Regression
BLISCHKE AND MURTHY (editors) · Case Studies in Reliability and Maintenance
BLISCHKE AND MURTHY · Reliability: Modeling, Prediction, and Optimization
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
BOLLEN · Structural Equations with Latent Variables
BOROVKOV · Ergodicity and Stability of Stochastic Processes
BOULEAU · Numerical Methods for Stochastic Processes
BOX · Bayesian Inference in Statistical Analysis
BOX · R. A. Fisher, the Life of a Scientist
BOX and DRAPER · Empirical Model-Building and Response Surfaces
*BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER, and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment
BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
BROWN and HOLLANDER · Statistics: A Biomedical Introduction
BRUNNER, DOMHOF, and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation
CAIROLI and DALANG · Sequential Stochastic Optimization
CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science
CHAN · Time Series: Applications to Finance
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHATTERJEE and PRICE · Regression Analysis by Example, *Third Edition*
CHERNICK · Bootstrap Methods: A Practitioner's Guide
CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
*COCHRAN and COX · Experimental Designs, *Second Edition*
CONGDON · Applied Bayesian Modelling
CONGDON · Bayesian Statistical Modelling
CONOVER · Practical Nonparametric Statistics, *Third Edition*
COOK · Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

COVER and THOMAS · Elements of Information Theory
COX · A Handbook of Introductory Statistical Methods
*COX · Planning of Experiments
CRESSIE · Statistics for Spatial Data, *Revised Edition*
CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
*DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
DAVID and NAGARAJA · Order Statistics, *Third Edition*
*DEGROOT, FIENBERG, and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
DEMIDENKO · Mixed Models: Theory and Applications
DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
DODGE · Alternative Methods of Regression
*DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
*DOOB · Stochastic Processes
DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Third Edition*
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
*ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series
ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition,* Revised; Volume II, *Second Edition*
FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis
*FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*
FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FULLER · Introduction to Statistical Time Series, *Second Edition*
FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
GIFI · Nonlinear Multivariate Analysis
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
GROSS and HARRIS · Fundamentals of Queueing Theory, *Third Edition*
*HAHN and SHAPIRO · Statistical Models in Engineering
HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
HALD · A History of Probability and Statistics and their Applications Before 1750
HALD · A History of Mathematical Statistics from 1750 to 1930
HAMPEL · Robust Statistics: The Approach Based on Influence Functions
HANNAN and DEISTLER · The Statistical Theory of Linear Systems
HEIBERGER · Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HELLER · MACSYMA for Statisticians
HINKELMAN and KEMPTHORNE: · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design
HOAGLIN, MOSTELLER, and TUKEY · Exploratory Approach to Analysis of Variance
HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
*HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE · Multiple Comparison Procedures
HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Second Edition*
HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
HOGG and KLUGMAN · Loss Distributions
HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*
HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
HOSMER and LEMESHOW · Applied Survival Analysis: Regression Modeling of Time to Event Data
HUBER · Robust Statistics
HUBERTY · Applied Discriminant Analysis
HUNT and KENNEDY · Financial Derivatives in Theory and Practice
HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—with Commentary
HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data
IMAN and CONOVER · A Modern Approach to Statistics
JACKSON · A User's Guide to Principle Components
JOHN · Statistical Methods in Engineering and Quality Assurance
JOHNSON · Multivariate Statistical Simulation
JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JOHNSON and BHATTACHARYYA · Statistics: Principles and Methods, *Fifth Edition*
JOHNSON and KOTZ · Distributions in Statistics
JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
JOHNSON, KOTZ, and KEMP · Univariate Discrete Distributions, *Second Edition*
JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*
JUREČKOVÁ and SEN · Robust Statistical Procedures: Aymptotics and Interrelations

*Now available in a lower priced paperback edition in the Wiley Classics Library.

JUREK and MASON · Operator-Limit Distributions in Probability Theory
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*
KASS and VOS · Geometrical Foundations of Asymptotic Inference
KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS · Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
*KISH · Statistical Design for Research
KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions, *Second Edition*
KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions, *Second Edition*
KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
LACHIN · Biostatistical Methods: The Assessment of Relative Risks
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry
LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
LAWSON · Statistical Methods in Spatial Epidemiology
LE · Applied Categorical Data Analysis
LE · Applied Survival Analysis
LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
LIAO · Statistical Group Comparison
LINDVALL · Lectures on the Coupling Method
LINHART and ZUCCHINI · Model Selection
LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
LLOYD · The Statistical Analysis of Categorical Data
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
MALLER and ZHOU · Survival Analysis with Long Term Survivors
MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel
MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data

*Now available in a lower priced paperback edition in the Wiley Classics Library.

MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
MARCHETTE · Random Graphs for Statistical Pattern Recognition
MARDIA and JUPP · Directional Statistics
MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
McCULLOCH and SEARLE · Generalized, Linear, and Mixed Models
McFADDEN · Management of Data in Clinical Trials
*McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data
McLACHLAN and KRISHNAN · The EM Algorithm and Extensions
McLACHLAN and PEEL · Finite Mixture Models
McNEIL · Epidemiological Research Methods
MEEKER and ESCOBAR · Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and Regression, *Third Edition*
*MILLER · Survival Analysis, *Second Edition*
MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Third Edition*
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory
MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks
MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and Nonlinear Optimization
MURTHY, XIE, and JIANG · Weibull Models
MYERS and MONTGOMERY · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Second Edition*
MYERS, MONTGOMERY, and VINING · Generalized Linear Models. With Applications in Engineering and the Sciences
*NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
NELSON · Applied Life Data Analysis
NEWMAN · Biostatistical Methods in Epidemiology
OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
*PARZEN · Modern Probability Theory and Its Applications
PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
PIANTADOSI · Clinical Trials: A Methodologic Perspective
PORT · Theoretical Probability for Applications
POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
PRESS · Bayesian Statistics: Principles, Models, and Applications
PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
PUKELSHEIM · Optimal Experimental Design
PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
*RAO · Linear Statistical Inference and Its Applications, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
RENCHER · Linear Models in Statistics
RENCHER · Methods of Multivariate Analysis, *Second Edition*
RENCHER · Multivariate Statistical Inference with Applications
*RIPLEY · Spatial Statistics
RIPLEY · Stochastic Simulation
ROBINSON · Practical Strategies for Experimenting
ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
ROSS · Introduction to Probability and Statistics for Engineers and Scientists
ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN · Simulation and the Monte Carlo Method
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RYAN · Modern Regression Methods
RYAN · Statistical Methods for Quality Improvement, *Second Edition*
SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis
*SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
SCHOTT · Matrix Analysis for Statistics
SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
*SEARLE · Linear Models
SEARLE · Linear Models for Unbalanced Data
SEARLE · Matrix Algebra Useful for Statistics
SEARLE, CASELLA, and McCULLOCH · Variance Components
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER and LEE · Linear Regression Analysis, *Second Edition*
*SEBER · Multivariate Observations
SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
*SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: It's Only a Game!
SILVAPULLE and SEN · Constrained Statistical Inference: Order, Inequality and Shape Constraints
SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building
THOMPSON · Sampling, *Second Edition*
THOMPSON · Simulation: A Modeler's Approach

*Now available in a lower priced paperback edition in the Wiley Classics Library.

THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
TIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) · Box on Quality and Discovery: with Design, Control, and Robustness
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN BELLE · Statistical Rules of Thumb
VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*
VESTRUP · The Theory of Measures and Integration
VIDAKOVIC · Statistical Modeling by Wavelets
VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
WEERAHANDI · Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models
WEISBERG · Applied Linear Regression, *Third Edition*
WELSH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOODWORTH · Biostatistics
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization
YANG · The Construction Theory of Denumerable Markov Processes
*ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library.