

6. PARAMETER ESTIMATION AND MODEL VALIDATION

6.1 Overview

As throughout this handbook, general explanations are given in Roman typeface, with **boldface** used for new terms where they are introduced or defined. Arial font is used for examples, and for any extended discussion that applies only to a particular example.

6.1.1 Introduction

Probabilistic risk assessment (PRA) analyzes accident sequences in terms of initiating events, basic events, and occasionally recovery events.

This handbook is concerned with estimating the frequencies of initiating events, the probabilities of basic events, and the distributions of recovery times and other durations. These estimates are propagated through logical relations to produce an estimated frequency of the undesirable end state, such as core damage. Moreover, the uncertainties in the parameter estimates must be quantified, and this must be done in a way that allows the uncertainty in the final estimate to be quantified.

Two approaches to estimating parameters are the Bayesian method and the frequentist, or classical, method. The two approaches are summarized here, and also in Appendix B.

In the Bayesian setting, probability is a measure of uncertainty, a quantification of degree of belief. The Bayesian methodology is used to modify uncertainty in a logically coherent way, so that “degree of belief” is rational, not merely personal opinion. In this methodology, each unknown parameter is assigned an initial **prior** probability distribution. This does not mean that the parameter varies randomly, but only that it is unknown, with the probability distribution modeling belief concerning the true value. Based on data, the analyst’s prior belief about the parameter is updated, using Bayes’ Theorem. The final inference statement uses the **posterior** distribution of the parameter to quantify the final uncertainty about the parameter. It is conditional on the observed data. Siu and Kelly (1998) give a simple but thorough introduction to Bayesian estimation in the PRA context.

The frequentist setting is quite different. A parameter is an unknown constant, and the data are modeled as occurring randomly. A frequency or rate, λ , represents the long term average rate at which the event occurs, the average number of events per unit time. Similarly, a probability of failure on demand, p , represents the long term fraction of failures in a large number of demands. The only random variability is in the data that happen to have been generated by the process. When quantifying uncertainty in an estimate, a frequentist asks questions such as, “Under similar conditions, what other data sets might have been generated? From data set to data set, how much variation would be seen in the parameter estimate? For any one data set, how far might the estimated parameter be from the true parameter?” Any prior or external information about the parameter value is ignored.

Statisticians have argued vigorously over which approach is preferable. When estimating parameters for PRA, the Bayesian approach clearly works better, for two reasons. First, data from reliable equipment are typically sparse, with few or no observed failures. In such cases, it is reasonable to draw on other sources of information. The Bayesian approach provides a mechanism for incorporating such information as prior belief. Second, the Bayesian framework allows straightforward propagation of basic event uncertainties through a logical model, to produce an uncertainty on the frequency of the undesirable end state. It assigns a probability distribution to each of the unknown parameters, draws a random sample from each, and constructs the corresponding sample for the frequency of the undesirable end state. The frequentist approach cannot handle such complicated propagation of uncertainties except by rough approximations.

Frequentist methods have their uses, however, even in PRA. Box (1980) writes “sampling theory [the frequentist approach] is needed for exploration and ultimate *criticism* of an entertained model in the light of current data, while Bayes’ theory is needed for *estimation* of parameters conditional on the adequacy of the entertained model.” This viewpoint agrees with current PRA practice. The primary use of the frequentist approach is in preliminary examination of the data, to check the correctness of model assumptions and to decide on what model to use. For example, frequentist methods can help the analyst decide whether data sets

6.

may be pooled or whether a trend is present. Goodness-of-fit tests and calculation of statistical significance are commonly used frequentist tools in this context. Then Bayesian methods are used for estimating the parameters.

Table 6.1 summarizes the above points.

6.1.2 Uncertainties Other Than Parametric Uncertainty

The above discussion might suggest that uncertainty in the value of parameters is the only uncertainty there is. That is not the case, of course. Parameter uncertainty, stemming from having only a relatively small set of randomly generated data, is the simplest uncertainty to address. It is the primary uncertainty considered in this handbook of parameter estimation. However, the

following kinds of uncertainty can also be considered.

6.1.2.1 Uncertainty from Nonrepresentativeness of the Data Sources

One issue to consider is that the data come from settings that do not perfectly match the problem of interest. In general, this is a difficult issue.

One special case is uncertainty of the value of a parameter for one data source (such as one nuclear power plant) when data are available from many similar but not identical data sources (other nuclear power plants). This case can be formulated in terms of a hierarchical model, and analyzed by empirical Bayes or hierarchical Bayes methods, as discussed in Section 8.2 of this handbook.

Table 6.1 Comparison of Bayesian and frequentist approaches in PRA.

	Frequentist	Bayesian
Interpretation of probability	Long-term frequency after many hypothetical repetitions.	Measure of uncertainty, quantification of degree of belief.
Unknown parameter	Constant, fixed.	Assigned probability distribution, measuring current state of belief.
Data	Random (before being observed).	Random for intermediate calculations. Fixed (after being observed) for the final conclusions.
Typical estimators	Maximum likelihood estimator (MLE), confidence interval.	Bayes posterior mean, credible interval.
Interpretation of 90% interval for a parameter	If many data sets are generated, 90% of the resulting confidence intervals will contain the true parameter. We do not know if our interval is one of the unlucky ones.	We believe, and would give 9 to 1 odds in a wager, that the parameter is in the interval.
Primary uses in PRA	1. Check model assumptions. 2. Provide quick estimates, without work of determining and justifying prior distribution.	1. Incorporate evidence from various sources, as prior distribution. 2. Propagate uncertainties through fault-tree and event-tree models.

6.1.2.2 Uncertainty in the Data Counts Themselves

There can be uncertainty in the data counts themselves. For example, it may be unclear whether a certain event should be counted as a failure. Or the number of demands may not be known exactly. A Bayesian method for dealing with uncertainty in PRA data was apparently first proposed by Siu and Apostolakis (1984, 1986), and it has been used by several authors, including Mosleh (1986), Mosleh et al. (1988, Section 3.3.4.4), and Martz and Picard (1995). As outlined by Atwood and Gentillon (1996), uncertainty in classifying the data yields a number of possible data sets, each of which can be assigned a subjective probability. The general approach is to use an “average” data set, a “best estimate” of the data, and analyze it. The uncertainty in the data is ignored, lost, at that point. A better approach is to analyze each data set, and combine the results. Each analysis produces a Bayesian distribution for the unknown parameter(s), and the final result is a mixture of these distributions. This approach includes the data uncertainty in the analysis, and results in wider uncertainty intervals than the general approach. The two approaches are diagrammed in Figure 6.1.

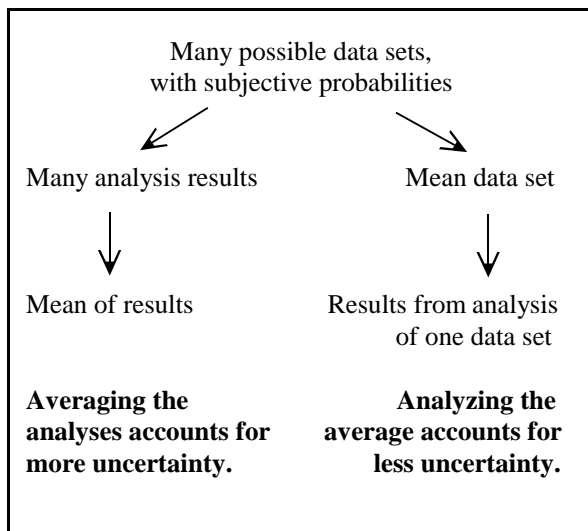


Figure 6.1 Two possible analysis paths for uncertain data.

Further treatment of this topic is beyond the scope of this handbook, but the reader can find additional guidance in the references cited above. This topic is

closely related to a statistical technique called “multiple imputation” (see Rubin 1996), in which a moderate number of data sets are randomly generated and then treated according to the left path in Figure 6.1.

6.1.2.3 Uncertainty in the Correct Model to Use

There can be uncertainty in which probability model to use. For example, there may be a slight trend, but it is borderline. Should a trend be modeled? Chapters 6 and 7 of this handbook discuss model validation extensively. However, model validation, which concludes that the model is either “adequate” or “not adequate,” is only a first step toward addressing this issue.

A more ambitious approach would quantify the degree of belief in each of a number of models, and propagate uncertainty in the models into the overall conclusions. This can use the predictions of various models as evidence in a formal Bayesian estimation procedure. See Mosleh et al. (1994) for a number of thoughtful papers on the definition and treatment of model uncertainties in the context of PRA applications. The topic is also discussed and debated in a tutorial article by Hoeting et al. (1999) with printed discussion. Bernardo and Smith (1994) also work out this approach in their Chapter 6 on “remodelling.” Drougett (1999) includes a discussion on the role of information concerning the models themselves (for example, their structure and past performance) in the estimation process.

Further consideration of such issues is beyond the scope of this handbook. The parameter uncertainties given here all assume that the model is a perfect description of the real world.

6.1.3 Chapter Contents

The rest of Chapter 6 presents statistical techniques for analyzing data for various parameters. Sections 6.2 through 6.7 cover exactly the same types of data as Sections 2.2 through 2.6, in the same order. The two kinds of failure to start in Section 2.3 are split into two sections here, 6.3 and 6.4. The three most extensive and fundamental sections are 6.2 (initiating events), 6.3 (failures on demand), and 6.7 (recovery times and other durations). The remaining sections draw on material from these three.

6.

Each section considers both parameter estimation and model validation. These two topics are considered together because checking the assumptions of the model (model validation) is a necessary part of any analysis. Separating the model validation from the parameter estimation might give the erroneous impression that it is all right to estimate parameters without checking the assumptions, or that the checks can be performed as an afterthought.

Under parameter estimation, both frequentist and Bayesian methods are presented. Under model validation, both graphical methods and formal statistical tests are given.

Much thought was given to the order of presentation: do we present the Bayesian estimates first or the frequentist estimates? In Chapter 6, the frequentist estimates are typically given first, not because they are more important or more highly recommended, but only because the frequentist point estimates are very simple, the simplest most natural estimates that someone might try. We cover them quickly before moving on to the more sophisticated Bayesian estimates. In the cases where the frequentist estimates are not simple (certain distribution models for durations), Bayesian estimation is discussed first.

6.2 Initiating Events

Initiating events here use the broad definition of the examples in Section 2.2, events that occur randomly in time and that initiate a quick response to restore the system to normal.

The event frequency is denoted λ , with units events per unit time. The data consist of x observed events in time t , where x is an integer ≥ 0 and t is some time > 0 . Note, t is considered nonrandom, and x is randomly generated. This can be expressed using the notation given in Appendix A, with upper case letters denoting random variables and lower case letters denoting numbers. Before data had been generated, the number of initiating events would have been denoted by X . For any particular number x , the probability of x initiating events in time t is

$$\Pr(X = x) = e^{-\lambda t} (\lambda t)^x / x! . \quad (6.1)$$

This formula for the Poisson distribution is a restatement of Equation 2.1, and will be used throughout this section.

The methods of parameter estimation will be illustrated by the following hypothetical data set.

Example 6.1 Initiating events with loss of heat sink.

In the last six years (during which the reactor was critical for 42800 hr.) a hypothetical PWR has had one initiating event that involved a loss of heat sink. The parameter to estimate is λ , the frequency of such events while the reactor is critical.

6.2.1 Frequentist or Classical Estimation

As explained in Section 6.1, Bayesian estimation methods are more important in PRA, but the classical estimator has a simpler form. Also, the comparison among estimators flows somewhat better if the short presentation of frequentist estimators precedes the lengthier presentation of Bayesian estimators. For these reasons, frequentist methods are given first in this section.

6.2.1.1 Point Estimate

The most commonly used frequentist estimate is the **maximum likelihood estimate** (MLE). It is found by taking the **likelihood**, given by Equation 6.1, and treating it as a function of λ . The value of λ that maximizes the likelihood is called the MLE. It can be shown (as a calculus exercise) that the maximum likelihood estimate (MLE) of λ is

$$\hat{\lambda} = x / t . \quad (6.2)$$

This formula is simple and intuitively natural, the observed number of events divided by the observed time period. This simplicity is part of the appeal of the MLE. The hat notation is used to indicate that the MLE is an estimate calculated from the data, not the true, unknown λ .

Example 6.1 has $x = 1$ and $t = 42800$ hrs. The likelihood is plotted on Figure 6.2 as a function of λ .

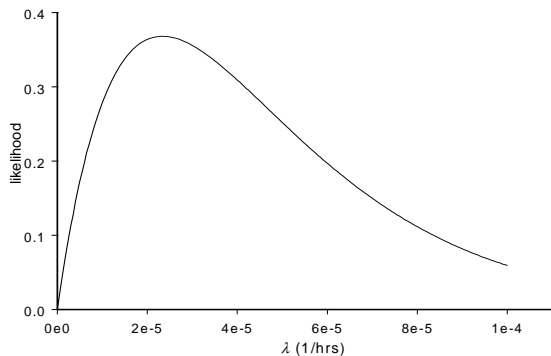


Figure 6.2 Likelihood as a function of λ , for data of Example 6.1.

The likelihood function is maximized when $\lambda = 1/42800 = 2.3\text{E-}5$. Therefore, the estimated event rate for the plant is

$$\hat{\lambda} = 1/42800 = 2.3\text{E-}5 \text{ events per critical-hour.}$$

Converting the hours to $42800/8760 = 4.89$ critical-years yields

$$\hat{\lambda} = 1/4.89 = 0.20 \text{ events per critical-year.}$$

In the above example, and in general throughout this handbook, the final answer is presented with few significant digits. This reflects the uncertainty inherent in all estimates. Indeed, sometimes not even the first significant digit is known precisely. During intermediate calculations, however, more significant digits will be shown, and used. This prevents roundoff errors from accumulating during the calculations.

It is also possible to combine, or **pool**, data from several independent processes, each having the same rate λ . In particular, suppose that the i th Poisson process is observed for time t_i , yielding the observed count x_i . The total number of event occurrences is $x = \sum x_i$, where the sum is taken over all of the processes, and the exposure time is $t = \sum t_i$. The rate λ is estimated by $\hat{\lambda} = x / t = \sum x_i / \sum t_i$. For example, if counts obtained for different years are used to estimate the rate, the estimate is the ratio of the total count to the total exposure time during these years.

6.2.1.2 Standard Deviation of the Estimator

The event count is random. In other words, if an identical plant could be observed during the same

years, a different number of events could be observed due to randomness. Similarly, the same plant might yield a different count over a different six-year period. Because the event count is random, the estimator is also random, and the estimate is simply the observed value for this plant during these years. Note the distinction in the terms: an **estimator** is a random variable, and an **estimate** is the particular value of the estimator after the data have been generated.

For a Poisson distributed random variable X , the mean and variance are the same, $E(X) = \text{var}(X) = \lambda t$, as stated in Appendix A.6.2. Consequently, the standard deviation of X is $(\lambda t)^{1/2}$, and the estimated standard deviation of the estimator $\hat{\lambda} \equiv X/t$ is

$$(\hat{\lambda} t)^{1/2} / t = (\hat{\lambda} / t)^{1/2} = x^{1/2} / t.$$

The estimated standard deviation of $\hat{\lambda}$ is also called the **standard error** for λ .

Thus, the standard error for λ in Example 6.1 is $1/4.89 = 0.20$ events per reactor-year.

A standard error is sometimes used for quick approximations when the data set is large. In that case, the MLE is approximately normal, and an approximate 95% confidence interval is given by $\text{MLE} \pm 2 \times (\text{standard error})$. This approximation holds for maximum likelihood estimation of virtually any parameter. For event frequencies, however, the following exact confidence interval can be found.

6.2.1.3 Confidence Interval for λ

Frequentist estimation is presented before Bayesian estimation because the MLE is so simple, simpler in form than the Bayes estimates. The same cannot be said for confidence intervals; the confidence-interval formulas are somewhat more complicated than the formulas for Bayesian interval estimates, and the interpretation of confidence intervals is more subtle. Readers may wish to skip directly to Section 6.2.2 on the first reading.

The confidence interval is given in many reference books, such as Johnson, Kotz, and Kemp (1992, Sec. 7.3), Bain and Engelhardt (1992, Section 11.4), or Martz and Waller (1991, Table 4.4). It is based on the chi-squared (or in symbols, χ^2) distribution, which is

6.

tabulated in Appendix C, and which can be found easily by many software packages. As used below, $\chi_p^2(d)$ is the p th quantile, or $(100p)$ th percentile, of the chi-squared distribution with d degrees of freedom. Do not misread $\chi_p^2(d)$ as involving multiplication.

For a $(1 - \alpha)$ confidence interval, or equivalently a $100(1 - \alpha)\%$ confidence interval, the lower limit is

$$\lambda_{\text{conf}, \alpha/2} = \frac{\chi_{\alpha/2}^2(2x)}{2t}$$

If $x = 0$, this formula is undefined, but then simply set $\lambda_{\text{conf}, \alpha/2} = 0$.

Similarly, the upper limit is

$$\lambda_{\text{conf}, 1-\alpha/2} = \frac{\chi_{1-\alpha/2}^2(2x+2)}{2t}$$

Notice that an upper confidence limit is defined in the case $x = 0$. It is reasonable that observing no occurrences of the event would provide some information about how large λ might be, but not about how small it might be.

The above formulas are in terms of α . Setting $\alpha = 0.1$, for example, gives the formulas for a 90% confidence interval. These formulas involves the 5th percentile of a chi-squared distribution with $2x$ degrees of freedom, and the 95th percentile of a chi-squared distribution with $(2x+2)$ degrees of freedom.

The resulting confidence interval is conservative in the sense that the actual confidence level is no smaller than the nominal level of $100(1 - \alpha)\%$, but it could be larger. This conservatism is inherent in confidence intervals based on discrete data.

In Example 6.1, 90% confidence limits are

$$\lambda_{\text{conf}, 0.05} = \frac{\chi_{0.05}^2(2)}{2 \times 4.89} = \frac{0.1026}{9.78} = 0.010$$

$$\lambda_{\text{conf}, 0.95} = \frac{\chi_{0.95}^2(4)}{2 \times 4.89} = \frac{9.488}{9.78} = 0.97$$

with units events per critical-year.

The interpretation of confidence intervals is given in

Appendix B, and deserves emphasis. In the frequentist approach, λ is fixed and the data are random. Therefore the maximum likelihood estimator and the confidence limits are all random. For most data sets the MLE, $\hat{\lambda}$, will be close to the true value of λ , and the confidence interval will contain λ . Sometimes, however, the MLE will be rather far from λ , and sometimes (less than 10% of the time) the 90% confidence interval will not contain λ . The procedure is good in the sense that most of the time it gives good answers, but the analyst never knows if the current data set is one of the unlucky ones.

To illustrate this, consider the following example with many hypothetical data sets from the same process.

Example 6.2 Confidence intervals from computer-generated data.

A computer was used to generate Poisson data, assuming an event rate $\lambda = 1.2$ events per year and assuming that 6 years were observed. Thus, the event count followed a Poisson distribution with mean $\lambda t = 7.2$. This was repeated, and 40 event counts were generated in all. These may be interpreted as counts from 40 identical plants, each observed for 6 years, or from 40 possible six-year periods at the same plant.

The first randomly generated event count was 10, the next was 5, the next was again 10, and so on. Some of the event counts were less than the long-term mean of 7.2, and some were greater. The maximum likelihood estimates of λ are plotted as dots in Figure 6.3. The corresponding 90% confidence intervals for λ are also plotted.

In Figure 6.3, the vertical dashed line shows the true value of λ , 1.2. Two of the 40 intervals (5%) are to the right of the true λ . These resulted from observing event counts of 14 and 16. One of the 40 intervals (2.5%) is to the left of the true λ . This interval was computed from an observed event count of 2.

Ideally, the error rates should both be 5%. They are not for two reasons. First, 40 is not a very large number, so the random data do not exactly follow the long-run averages. Second, confidence intervals with discrete data are inherently conservative: a 90%

confidence interval is defined so that the probability

of containing the true λ is *at least 90%*, and the error probabilities at each end are each *at most 5%*.

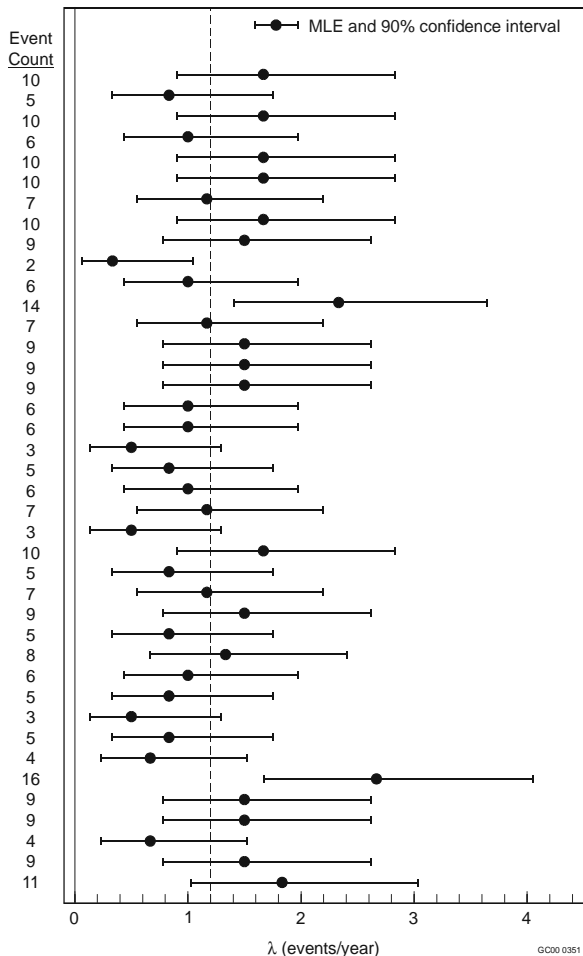


Figure 6.3 Confidence intervals from random data, all generated from the same process.

The data analyst will normally have data from just one plant for the six-year period. The resulting confidence interval will contain the true value of λ , unless the data happen to deviate greatly from the mean. Unfortunately, the analyst does not know when this has happened, only that it does not happen often.

6.2.2 Bayesian Estimation

6.2.2.1 Overview

Bayesian estimation of λ involves several steps. The prior belief about λ is quantified by a probability distribution, the **prior distribution**. This distribution will be restricted to the positive real line, because λ must be positive, and it will assign the most probability to the values of λ that are deemed most plausible. The data are then collected, and the **likelihood function** is constructed. This is given by Equation 6.1 for initiating events. It is the probability of the observed data, written as a function of λ . Finally, the **posterior distribution** is constructed, by combining the prior distribution and the likelihood function through Bayes' theorem. This theorem says that

$$f_{\text{post}}(\lambda) \propto \text{likelihood}(\lambda) \times f_{\text{prior}}(\lambda).$$

The posterior distribution shows the updated belief about the values of λ . It is a modification of the prior belief that accounts for the observed data.

Figure 6.4, adapted from a tutorial article by Siu and Kelly (1998), shows how the posterior distribution changes as the data set changes. The figure is based on a diffuse prior, and on three hypothetical data sets, with $x = 1$ event in $t = 10,000$ hours, $x = 10$ events in $t = 100,000$ hours, and $x = 50$ events in $t = 500,000$ hours, respectively. Note, each of these data sets has $\hat{\lambda} = x/t = 1.E-4$ events per hour. The figure shows the prior distribution, and the three posterior distributions corresponding to the three data sets.

For a small data set, the posterior distribution resembles the prior to some extent. As the data set becomes larger, several patterns are evident:

- The posterior distribution departs more and more from the prior distribution, because the data contribute the dominant information.
- The posterior distribution becomes more concentrated, indicating better knowledge of the parameter, less uncertainty.
- The posterior distribution becomes approximately centered around the MLE, $\hat{\lambda}$.

6.

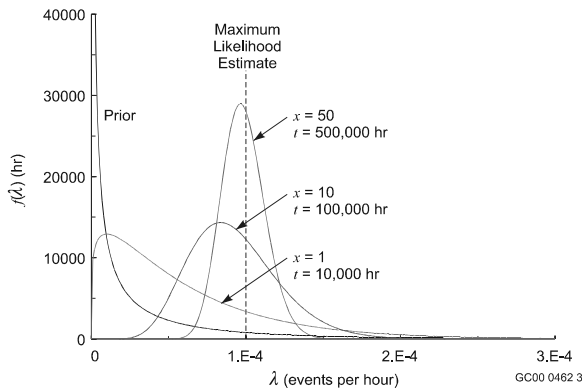


Figure 6.4 Prior distribution and posterior distributions corresponding to three hypothetical data sets.

To be consistent with the notation for random variables, upper case letters would be used for uncertain parameters that have probability distributions. Such notation is not customary in the Bayesian literature, and will not be used here. The reader must judge from context whether the letter λ denotes a particular value or the uncertain parameter with an associated distribution.

6.2.2.2 Choosing a Prior

The subsections below consider estimation of λ using various possible prior distributions. The simplest prior distribution is discrete. The posterior can be calculated easily, for example by a spreadsheet. The next simplest prior is called **conjugate**; this prior combines neatly with the likelihood to give a posterior that can be evaluated by simple formulas. Finally, the most general priors are considered; the posterior distribution in such a case can only be found by numerical integration or by random sampling.

The prior distribution should accurately reflect prior knowledge or belief about the unknown parameter. Quantifying belief is not easy, however. Raiffa and Schlaifer (1961, Sections 3.3.3-3.3.5) point out that most people can think more easily in terms of percentiles of a distribution than in terms of moments. They also give advice on looking at the situation from many directions, to make sure that the prior belief is internally consistent and has been accurately quantified. Siu and Kelly (1998, Sec. 5.1.4) present seven warnings in connection with developing a prior distribution, which are summarized here.

- Beware of zero values. If the prior says that a value of λ is impossible, no amount of data can overcome this.
- Beware of cognitive biases, caused by the way people tend to think.
- Beware of generating overly narrow prior distributions.
- Ensure that the evidence used to generate the prior distribution is relevant to the estimation problem.
- Be careful when assessing parameters that are not directly observable.
- Beware of conservatism. Realism is the ideal, not conservatism.
- Be careful when using discrete probability distributions.

For fuller discussion of these points, see Siu and Kelly.

Some priors are chosen to be "noninformative," that is, diffuse enough that they correspond to very little prior information. The **Jeffreys noninformative prior** is often used in this way. If information is available, it is more realistic to build that information into the prior, but sometimes the information is difficult to find and not worth the trouble. In such a case, the Jeffreys noninformative prior can be used. It is one of the priors discussed below.

6.2.2.3 Estimation with a Discrete Prior

To illustrate use of a discrete prior in Bayes estimation, we return to Example 6.2. In the first sample, 10 events were observed. This case is evaluated three times to illustrate several aspects of the effect of using discrete approximations and to provide a clear example for the readers to allow them to practice hand calculations of the Bayesian update process. Working out a series of Bayesian calculations with increasing quantities of evidence (such as in Figure 6.4 above) can provide a good sense of how the process works and how the posterior distribution changes with the evidence. For example, if

$$f(\lambda_i|E) = \frac{f(\lambda_i)L(E|\lambda_i)}{\sum_{i=1}^N L(E|\lambda_i)f(\lambda_i)}$$

where

$$f(\lambda_i|E) = \text{probability density function of given}$$

λ_i evidence E (posterior distribution)

$f(\lambda_i)$ = the probability density prior to having evidence E (prior distribution)

$L(E|\lambda_i)$ = the likelihood function (probability of the evidence given λ_i)

Note that the denominator, the total probability of the evidence E, is simply a normalizing constant.

Next, when the evidence is in the form of F failures over an operational time T, the likelihood function is the Poisson distribution:

$$L(E|\lambda_i) = e^{(-\lambda_i T)} \frac{(\lambda_i T)^F}{F!}$$

For this example, let us use a simple flat prior distribution over the range 0 to 6 events per year. Because of the nature of the example, we could use a distribution peaked at 1.2 events per year or, alternatively the Jeffreys noninformative prior. Over the restricted range of 1 to 6, the flat prior is essentially noninformative, but the real reason we chose it is to make the impact of the Bayesian updating process easy to see.

Given the ease of calculation with current computers, a finely discretized prior (say at 0, 0.01, 0.02,...6.00) would give the most accurate results and we will provide that calculation in a moment. First let us use a very coarse prior at 0, 0.5, 1.0, ...6.0. With only 13 bins, the reader can perform hand calculations quite easily. The results are given in Table 6.2 and in Figure 6.5.

Even with such a coarse prior, the evidence is strong and peaks at about $\lambda_i = 1.5$ per year. There is essentially no chance the value is greater than 4 or less than 0.5. We suggest that the reader make up a data set for examining the way the posterior distribution responds to growing evidence. For example, try beginning with 0 failures in year 1; then adding 2 failures in year 2; then 0 failures in year 3; etc. Also try a case that does not agree with the prior; for example 5 failures in year 1; then 7 more in year 2; then 6 in year 3.

Table 6.2 Example 6.2, First Sample (10 events in 6 years).

Event Rate	Prior Probability	Likelihood		Posterior Probability	Cumulative Probability
λ_i	p_i	L_i	$p_i \times L_i$	$P_i(\lambda E)$	ΣP_i
0.0	0.077	0.00E+00	0.00E+00	0.00E+00	0.00E+00
0.5	0.077	8.10E-04	6.23E-05	2.43E-03	2.43E-03
1.0	0.077	4.13E-02	3.18E-03	1.24E-01	1.26E-01
1.5	0.077	1.19E-01	9.12E-03	3.56E-01	4.82E-01
2.0	0.077	1.05E-01	8.06E-03	3.14E-01	7.96E-01
2.5	0.077	4.86E-02	3.74E-03	1.46E-01	9.42E-01
3.0	0.077	1.50E-02	1.15E-03	4.49E-02	9.87E-01
3.5	0.077	3.49E-03	2.68E-04	1.05E-02	9.98E-01
4.0	0.077	6.60E-04	5.07E-05	1.98E-03	1.00E+00
4.5	0.077	1.07E-04	8.20E-06	3.20E-04	1.00E+00
5.0	0.077	1.52E-05	1.17E-06	4.57E-05	1.00E+00
5.5	0.077	1.97E-06	1.51E-07	5.90E-06	1.00E+00
6.0	0.077	2.34E-07	1.80E-08	7.01E-07	1.00E+00

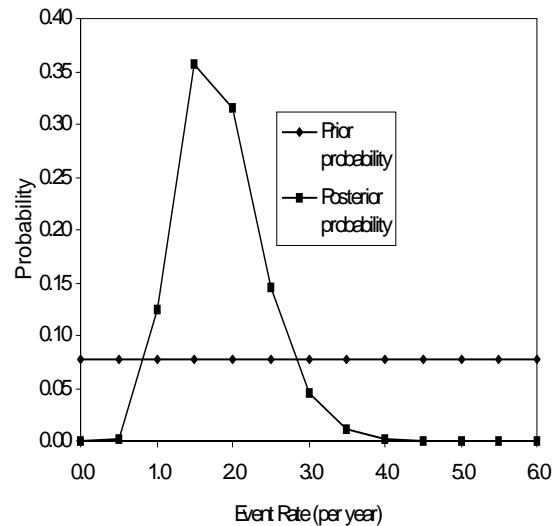


Figure 6.5 Discrete prior and posterior distributions for data in Example 6.2, coarse discretized prior.

If we repeat the calculation with a discrete prior twice as fine (i.e., 0, 0.25, 0.50, 0.75,...6.00), the prior now has 25 bins and the results are much more smooth as shown in Figure 6.6. These results are quite smooth and of course follow the previous results.

6.

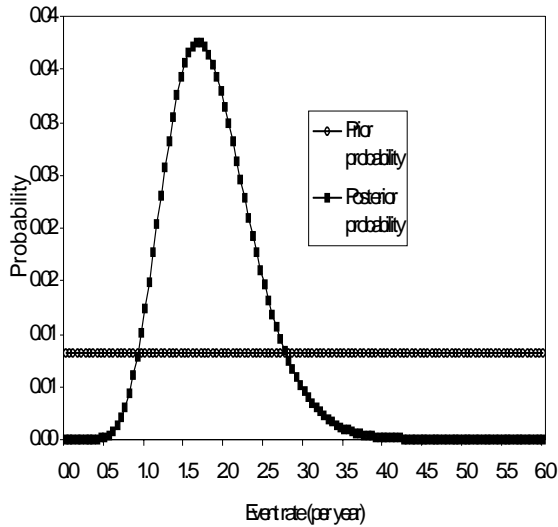


Figure 6.6 Discrete prior and posterior distributions for data in Example 6.2, finely discretized prior.

Finally, let us repeat the calculation for a discrete flat prior at 0, 0.05, 0.10, 0.15,...6.00, i.e., a 121 bin histogram. This time the results, shown in Figure 6.7, are detailed enough for us to accurately pick points off the distribution.

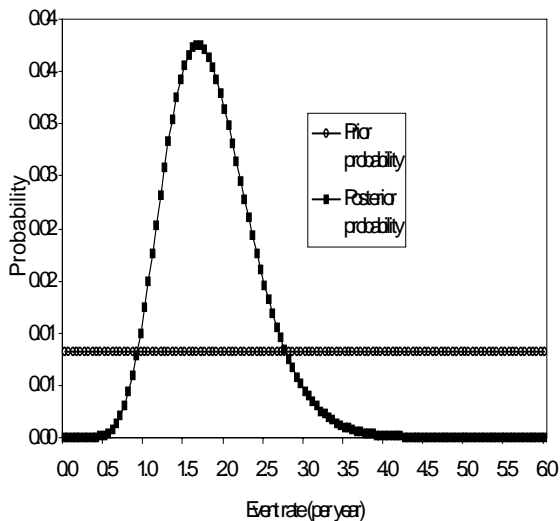


Figure 6.7 Discrete prior and posterior distributions for data in Example 6.2, very fine discretized prior.

The statistics from the spreadsheet calculation, which is identical with Table 6.2, except having 121 bins rather than 13, are provided in Table 6.3. These results are also compared with the frequentist estimates obtained from the first sample shown earlier in Figure 6.3. The Bayesian estimate, with a flat essentially noninformative prior, yields slightly more narrow 90% probability bounds than the 90% confidence interval of the frequentist estimate.

Table 6.3. Comparison of Bayesian and Frequentist estimates for the data in Example 6.2.

Estimate	5 th %tile	MLE	95 th %tile
Bayes, flat prior	1.00	1.65	2.80
Frequentist, Fig. 6.3	0.95	1.73	2.85

6.2.2.4 Estimation with a Conjugate Prior

6.2.2.4.1 Definitions

The conjugate family of prior distributions for Poisson data is the family of gamma distributions. Two parameterizations of gamma distributions are given in Appendix A.7.6. For Bayesian estimation, the following parameterization is the more convenient one:

$$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} \quad (6.3)$$

Here λ has units 1/time and β has units of time, so the product $\lambda\beta$ is unitless. For example, if λ is the frequency of events per critical-year, β has units of critical-years. The parameter β is a **scale parameter**, although purists would say that $1/\beta$ is the actual scale parameter. In any case, β corresponds to the scale of λ — if we convert λ from events per hour to events per year by multiplying it by 8760, we correspondingly divide β by 8760, converting it from hours to years. The other parameter, α , is unitless, and is called the **shape parameter**. The gamma function, $\Gamma(\alpha)$, is a standard mathematical function, defined in Appendix A.7.6; if α is a positive integer, $\Gamma(\alpha)$ equals $(\alpha-1)!$

Let λ have a gamma uncertainty distribution. In the present parameterization, the mean of the gamma distribution, or the expected value $E(\lambda)$, is α/β , and the variance, $\text{var}(\lambda)$, is α/β^2 . Note that the units are correct, units 1/time for the mean and 1/time² for the variance.

6.2.2.4.2 Update Formulas

As stated earlier and in Appendix B.5.1, the posterior distribution is related to the prior distribution by

$$f_{\text{post}}(\lambda) \propto \Pr(X = x | \lambda) f_{\text{prior}}(\lambda) \quad (6.4)$$

The probability of the data is also called the **likelihood**, in which case it is considered as a function of the parameter λ for a given x . For Poisson data, it is given by Equation 6.1. The symbol \propto denotes "is proportional to." Probability density functions generally have normalizing constants in front to make them integrate to 1.0. These constants can be complicated, but using proportionality instead of equality allows us to neglect the normalizing constants. Stripped of all the normalizing constants, the gamma p.d.f. is

$$f(\lambda) \propto \lambda^{\alpha-1} e^{-\lambda\beta}.$$

The gamma distribution and the Poisson likelihood combine in a beautifully convenient way

$$\begin{aligned} f_{\text{post}}(\lambda) &\propto e^{-\lambda t} \frac{(\lambda t)^x}{x!} \lambda^{\alpha-1} e^{-\lambda\beta} \\ &\propto \lambda^{(x+\alpha)-1} e^{-\lambda(t+\beta)} \end{aligned}$$

In the final expression, everything that does not involve λ has been absorbed into the proportionality constant. This result is "beautifully convenient," because the posterior distribution of λ is again a gamma distribution. This is the meaning of **conjugate**: if the prior distribution is a member of the family (in this case, the gamma family), the posterior distribution is a member of the same family. The update formulas are

$$\begin{aligned} \alpha_{\text{post}} &= x + \alpha_{\text{prior}} \\ \beta_{\text{post}} &= t + \beta_{\text{prior}} \end{aligned}$$

This leads to an intuitive interpretation of the prior parameters: a gamma($\alpha_{\text{prior}}, \beta_{\text{prior}}$) distribution is equivalent, at least intuitively, to having seen α_{prior} events in β_{prior} time units, prior to taking the current data.

Figure 6.4 was constructed in this way. The prior distribution was gamma(0.2, 10,000). Therefore, the posterior distributions were gamma(1.2, 20,000), gamma(10.2, 110,000), and gamma(50.2, 510,000).

When using these update formulas, be sure that t and β_{prior} have the same units. If one is expressed in hours and one in years, one of the two numbers must be converted before the two are added.

The moments of the gamma distribution were mentioned previously. The posterior mean is $\alpha_{\text{post}}/\beta_{\text{post}}$ and the posterior variance is $\alpha_{\text{post}}/(\beta_{\text{post}})^2$.

The percentiles of the gamma distribution are given by many software packages. If you use such software, be careful to check that it is using the same parameterization that is used here! Here are three ways to get the correct answer. (1) If the software uses the other parameterization, fool it by inverting your value of β . Then do a sanity check to make sure that the numbers appear reasonable. (2) A safe method is to have the software find the percentiles of the gamma($\alpha_{\text{post}}, 1$) distribution. Then manually divide these percentiles by β_{post} . This ensures that the scale parameter is treated correctly. (3) As a final alternative, the percentiles of the gamma distribution can be found from a tabulation of the chi-squared distribution, possibly interpolating the table. To do this, denote the $(100p)$ th percentile of the posterior distribution by λ_p . For example, denote the 95th percentile by $\lambda_{0.95}$. The $(100p)$ th percentile is given by

$$\lambda_p = \chi_p^2(2\alpha_{\text{post}})/(2\beta_{\text{post}})$$

where, as before, $\chi_p^2(d)$ is the p th quantile, or $(100p)$ th percentile, of a chi-squared distribution with d degrees of freedom. Note the presence of 2 in the numerator and denominator when the chi-squared distribution is used.

6.

The next section contains examples that use these update formulas with several priors.

6.2.2.5 Possible Conjugate Priors

6.2.2.5.1 Informative Priors

The prior distribution must come from sources other than the current data. It might be tempting to use the data when constructing the prior distribution, but that temptation must be resisted. Prior distributions are named "prior" for a reason: they reflect information that does not come from the current data.

Ideally, generic data provide the basis for prior belief. Consider again Example 6.1, involving initiating events with loss of heat sink. With no special knowledge about the plant, prior belief about the plant is reasonably based on the overall industry performance, so we use the generic industry distribution as the prior.

Poloski et al. (1999a) examined initiating-event data from the nuclear power industry over nine years. For PWRs, and initiating events involving loss of heat sink, they determined that the variability of λ across the industry can be described by a gamma distribution with shape parameter = 1.53 and scale parameter = 10.63 reactor-critical-years. Regrettably, Table G-1 of the report gives only a mean and a 90% interval, not the distribution and its parameters. The distribution given here is taken from the unpublished work that formed the basis of the report. The distribution is a gamma distribution, so the update formulas given above can be used in the hypothetical example of this section. The prior distribution is shown in Figure 6.8.

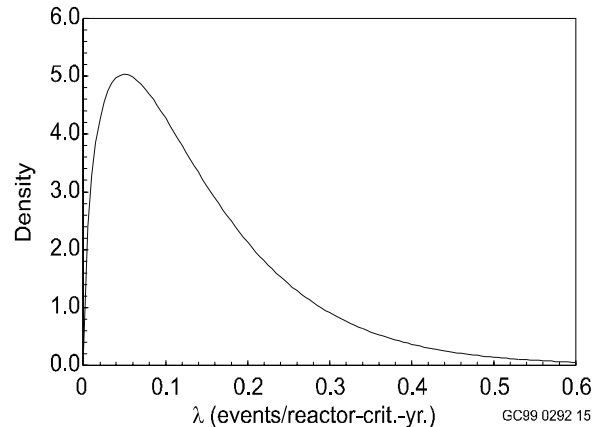


Figure 6.8 Prior density for λ , gamma(1.53, 10.63).

Now consider updating this prior with the data of Example 6.1. To make the units consistent, convert the 42800 reactor-critical-hours in the example to $42800/8760 = 4.89$ reactor-critical-years. The update formula yields

$$\alpha_{\text{post}} = x + \alpha_{\text{prior}} = 1 + 1.53 = 2.53$$

$$\beta_{\text{post}} = t + \beta_{\text{prior}} = 4.89 + 10.63 = 15.52 \text{ reactor-critical-years}$$

The mean, $\alpha_{\text{post}}/\beta_{\text{post}}$, is 0.163 events per reactor-critical-year, the variance is 0.0105 (per reactor-critical-year squared), and the standard deviation is the square root of the variance, 0.102 per reactor-critical-year.

A 90% credible interval is the interval from the 5th to the 95th percentiles of the posterior distribution. A software package finds the two percentiles of a gamma(2.53, 1.0) to be 0.5867 and 5.5817. Division by β_{post} yields the two percentiles of the posterior distribution: 0.038 and 0.36. Alternatively, one may interpolate Table C.2 of Appendix C to find the percentiles of a chi-squared distribution with 5.06 degrees of freedom, and divide these percentiles by $2\beta_{\text{post}}$. Linear interpolation gives answers that agree to three significant digits with the exact answers, but if the degrees of freedom had not been so close to an integer the linear interpolation might have introduced a small inaccuracy.

The interpretation of the above numbers is the following. The best belief is that λ is around 0.16, although it could easily be somewhat larger or smaller. Values as small as 0.038 or as large as 0.36 are possible but are approaching the limits of credibility.

Two graphical ways of presenting this information are given below. Figure 6.9 shows the posterior density. The areas to the left of the 5th percentile and to the right of the 95th percentile are shaded. The 90% credible interval is the interval in the middle, with probability 90%. Figure 6.10 shows the same information using the cumulative distribution. The 5th and 95th percentiles are the values of λ where the cumulative distribution is 0.05 and 0.95, respectively. These percentiles are the same values as shown in the plot of the density.

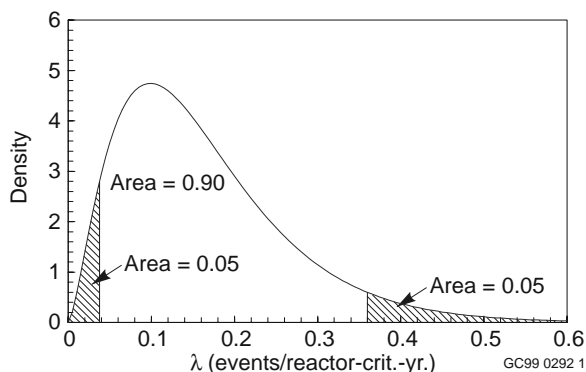


Figure 6.9 Posterior density of λ , gamma(2.53, 15.52), for Example 6.1 with industry prior. The 5th and 95th percentiles are shown.

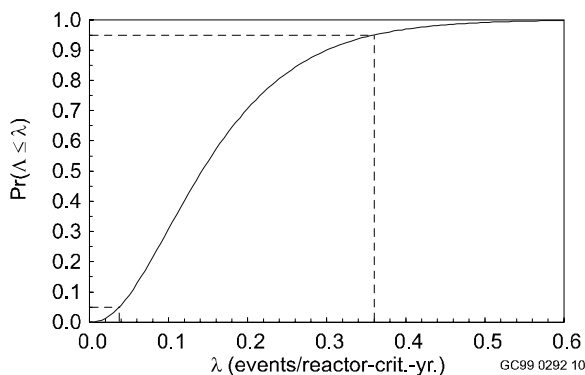


Figure 6.10 Posterior cumulative distribution of λ for Example 6.1 with industry prior. The 5th and 95th percentiles are shown.

Several points deserve mention.

- The above interval puts equal probability in the two tails outside the credible interval.

Other credible intervals are possible, such as a one-sided interval that puts all the error in one tail, or an interval that includes the highest posterior density, with possibly unequal probabilities in the two tails.

- For PRA applications, however, the right tail is typically of concern for risk, corresponding to high initiating event frequency (or, in other sections of this chapter, high probability of failure on demand, high unavailability, or long time to recovery). The interval given above holds the error probability for the right tail equal to 0.05. This number is customary in much statistical practice, and has therefore been used in many studies for the NRC. As for the left tail, it is easy to put a positive lower end on the credible interval, even though values of λ near zero are not a concern for risk. Therefore, the above 90% interval, corresponding to 5% posterior probability in each tail, is commonly presented in PRA studies.
- Actually, however, the interval presents only a portion of the information in the posterior distribution. The full distribution is used in a PRA.

6.2.2.5.2 Noninformative Prior

The **Jeffreys noninformative prior** is intended to convey little prior belief or information, thus allowing the data to speak for themselves. This is useful when no informed consensus exists about the true value of the unknown parameter. It is also useful when the prior distribution may be challenged by people with various agendas. Some authors use the term **reference prior** instead of "noninformative prior," suggesting that the prior is a standard default, a prior that allows consistency and comparability from one study to another.

With Poisson data, the Jeffreys noninformative prior is obtained if the shape parameter of a gamma distribution is taken to be $\alpha = 1/2$ and the parameter β is taken to be zero. (See, for example, Box and Tiao 1973.) Ignoring the normalizing constant at the front of Equation 6.1 yields a function that is proportional to $\lambda^{-1/2}$, shown in Figure 6.11. Although this function is interpreted as a density function, it is an **improper**

6.

distribution because its integral from 0 to ∞ is infinite.

Suppose that the data consist of x events in time t . Formal application of the update formulas yields

$$\alpha_{\text{post}} = x + 1/2$$

$$\beta_{\text{post}} = t + 0.$$

That is, the Bayes posterior distribution for λ is $\text{gamma}(x + 1/2, t)$.

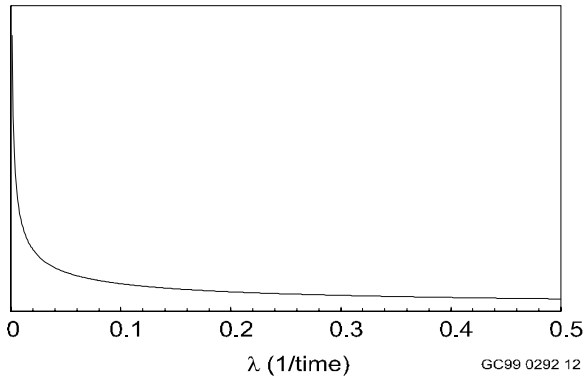


Figure 6.11 Jeffreys noninformative prior distribution for an event frequency.

It is interesting to compare the interval using the Jeffreys prior with the corresponding confidence interval. The 90% posterior credible interval is

$$\lambda_{0.05} = \chi^2_{0.05}(2x + 1)/2t$$

$$\lambda_{0.95} = \chi^2_{0.95}(2x + 1)/2t$$

These may be compared with the 90% confidence interval:

$$\lambda_{\text{conf}, 0.05} = \chi^2_{0.05}(2x)/2t$$

$$\lambda_{\text{conf}, 0.95} = \chi^2_{0.95}(2x + 2)/2t$$

The confidence intervals differ from the Bayes credible intervals only in the degrees of freedom, and there only slightly. This is the primary sense in which the Jeffreys prior is "noninformative." The lower and upper *confidence* limits have degrees of freedom $2x$ and $2x + 2$, respectively. The two *Bayesian* limits each use the average, $2x + 1$. The confidence interval is wider than the Jeffreys credible interval, a reflection of the conservatism of confidence limits with discrete data. How-

ever the similarity between the confidence limits and the Jeffreys limits shows that the result using the Jeffreys prior will resemble the result using frequentist methods, that is, using no prior information at all.

Consider again Example 6.1, with 1 event in 4.89 critical-years, and use the Jeffreys noninformative prior. The resulting posterior distribution has

$$\alpha_{\text{post}} = 1.5$$

$$\beta_{\text{post}} = 4.89 \text{ critical-years}.$$

The mean of this distribution is $1.5/4.89 = 0.31$ events per critical-year. A 90% Bayes credible interval can be obtained from a chi-squared table without any need for interpolation, because the degrees of freedom parameter is $2 \times 1 + 1$, an integer. The 5th and 95th percentiles of the chi-squared distribution are 0.3518 and 7.815. Division by 2×4.89 yields the percentiles of the posterior distribution, 0.036 and 0.80.

This posterior distribution has a larger mean and larger percentiles than the posterior distribution in Section 6.2.2.5.1. The data set is the same, but the different prior distribution results in a different posterior distribution. The results will be compared in Section 6.2.2.5.4.

6.2.2.5.3 Constrained Noninformative Prior

This prior is a compromise between an informative prior and the Jeffreys noninformative prior. The mean of the constrained noninformative prior uses prior belief, but the dispersion is defined to correspond to little information. These priors are described by Atwood (1996) and by references given there. Constrained noninformative priors have not been widely used, but they are mentioned here for the sake of completeness.

For Poisson data, the constrained noninformative prior is a gamma distribution, with the mean given by prior belief and the shape parameter = $1/2$. That is,

$$\alpha_{\text{prior}} = 1/2$$

$$\beta_{\text{prior}} \text{ satisfies } \alpha_{\text{prior}}/\beta_{\text{prior}} = \text{prior mean}.$$

To illustrate the computations, consider again the Example 6.1, with 1 event in 4.89 reactor-critical-years. Suppose we knew that in the industry overall such events occur with an average frequency of

0.144 events per reactor-critical-year. (This is consistent with the informative prior given above in Section 6.2.2.5.1.) Suppose further that we were unable or unwilling to make any statement about the dispersion around this mean — the full information used to construct the informative prior was not available, or the plant under consideration was atypical in some way, so that a more diffuse prior was appropriate.

The constrained noninformative prior with mean 0.144 has $\alpha_{\text{prior}} = \frac{1}{2}$ and $\beta_{\text{prior}} = 3.47$ critical-years. The resulting posterior distribution has

$$\alpha_{\text{post}} = x + \frac{1}{2} = 1.5$$

$$\beta_{\text{post}} = t + 3.47 = 8.36$$

The mean is 0.18 events per critical-year, and the 90% credible interval is (0.021, 0.47). This notation means the interval from 0.021 to 0.47.

6.2.2.5.4 Example Comparisons Using Above Priors

In general, the following statements can be made.

- The Jeffreys noninformative prior results in a posterior credible interval that is numerically similar to a confidence interval, but slightly shorter.
- If the prior mean exists, the posterior mean is

between the prior mean and the MLE.

- If two prior distributions have the same mean, the more concentrated (less diffuse) prior distribution will yield the more concentrated posterior distribution, and will pull the posterior mean closer to the prior mean.

These statements are now illustrated by example. The estimates found in the above sections for Example 6.2 and the various priors are compared in Table 6.4 and in Figure 6.12.

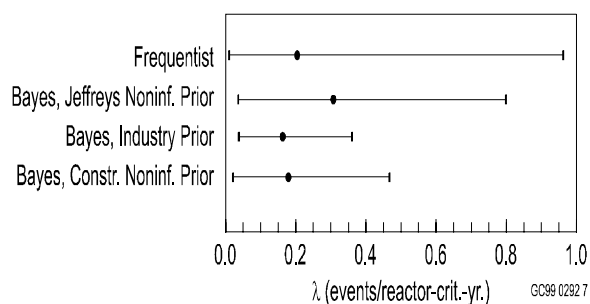


Figure 6.12 Comparison of four point estimates and interval estimates for λ .

Table 6.4 Comparison of estimates with 1 event in 4.89 reactor-critical-years.

Method	Prior mean	Posterior parameters	Point estimate (MLE or posterior mean)	90% interval (confidence interval or posterior credible interval)
Frequentist	NA	NA	0.20	(0.010, 0.97)
Bayes with Jeffreys noninformative prior, gamma(0.5, 0)	undefined	$\alpha = 1.5$ $\beta = 4.89$	0.31	(0.036, 0.80)
Bayes with (informative) industry prior, gamma(1.53, 10.63)	0.144	$\alpha = 2.53$ $\beta = 15.52$	0.16	(0.038, 0.36)
Bayes with constrained noninformative prior, gamma(0.5, 3.47)	0.144	$\alpha = 1.5$ $\beta = 8.36$	0.18	(0.021, 0.47)

6.

In Table 6.4 and in Figure 6.12, the Jeffreys prior and the frequentist approach are listed next to each other because they give numerically similar results. The Jeffreys prior yields a posterior credible interval that resembles the frequentist confidence interval. It is a little shorter, but it is neither to the right nor to the left. This agrees with the earlier discussion of the Jeffreys prior.

In each Bayesian case, the posterior mean falls between the prior mean (if defined) and the MLE, 0.20. The prior distribution has more influence when the prior distribution is more tightly concentrated around the mean. The concentration is measured by the shape parameter α_{prior} , because $1/\alpha$ equals the relative variance (= variance/mean²). Therefore the larger α the smaller the relative variance. The industry prior and the constrained noninformative prior have the same mean, but the industry prior has the larger α , that is, the smaller variance. As a consequence, in both cases the posterior mean is between the MLE, 0.204, and the prior mean, 0.144, but the posterior mean based on the industry prior is closer to 0.144, because that prior has a smaller variance. Because the prior mean is smaller than the MLE, the bottom two lines give smaller posterior estimates than do the top two lines. Also, the prior distribution with the most information (largest α) yields the most concentrated posterior distribution, and the shortest 90% interval.

In some situations, no conjugate prior is satisfactory. For example, a gamma distribution is very unrealistic if the shape parameter is very small. As a rule of thumb, the lower percentiles of the distribution are unrealistic if α is much smaller than 0.5. Such a posterior distribution arises with Poisson data when the prior distribution is very skewed (α very small) and the data contain zero events. Then the posterior distribution also is very skewed, and the posterior 5th percentile may be many orders of magnitude below the posterior mean. The subject-matter experts must look at the percentiles and decide if they are believable. If not, a more appropriate prior should be chosen. It will not be conjugate. This is the subject of the next subsection.

6.2.2.6 Estimation with a Continuous Nonconjugate Prior

Discrete priors and conjugate priors were updated above with simple formulas. What remains are the

continuous nonconjugate priors. Three approaches for updating them are given here.

6.2.2.6.1 Direct Numerical Integration

If software is available for performing numerical integration, the following approach can be used. Find the form of the posterior distribution, using Equation 6.4. Suppose, for example, that the prior distribution for λ is lognormal, with μ and σ^2 denoting the mean and variance of the normal distribution of $\ln\lambda$. As stated in Appendix A.7.3, the lognormal density is proportional to

$$f_{\text{LN}}(\lambda) \propto \frac{1}{\lambda} e^{-\frac{1}{2}\left(\frac{\ln\lambda-\mu}{\sigma}\right)^2}$$

Substitute this and Equation 6.1 into Equation 6.4, to obtain the form of the posterior density:

$$Cf_{\text{post}}(\lambda) = e^{-\lambda x} \lambda^x \frac{1}{\lambda} e^{-\frac{1}{2}\left(\frac{\ln\lambda-\mu}{\sigma}\right)^2}$$

All terms that do not involve λ have been absorbed into the normalizing constant, C . The normalizing constant can be evaluated by numerically integrating Cf_{post} from 0 to ∞ . Unless x is unrealistically large, the function does not need to be integrated in practice out beyond, say, $\ln\lambda = \mu + 5\sigma$. C equals the integral of Cf_{post} , because the integral of f_{post} must equal 1. Once C has been evaluated, the mean and percentiles of f_{post} can be found numerically.

Numerical integration, using a technique such as the trapezoidal rule or Simpson's rule, can be programmed easily, even in a spreadsheet. The ideas are found in some calculus texts, and in books on numerical methods such as Press et al. (1992).

6.2.2.6.2 Simple Random Sampling

A second approach, which does not directly involve numerical integration, is to generate a large random sample from the posterior distribution, and use the sample to approximate the properties of the distribution. Some people think of this as numerical integration via random sampling. Surprisingly, the random

sample can be generated without explicitly finding the form of the posterior distribution, as explained by Smith and Gelfand (1992).

The algorithm, called the **rejection method** for sampling from a distribution, is given here in its general form, and applied immediately to sampling from the posterior distribution. In general, suppose that it is possible to sample some parameter θ from a continuous distribution g , but that sampling from a different distribution f is desired. Suppose also that a positive constant M can be found such that $f(\theta)/g(\theta) \leq M$ for all θ . The algorithm is:

- (1) Generate θ from $g(\theta)$.
- (2) Generate u from a uniform distribution, $0 \leq u \leq 1$.
- (3) If $u \leq f(\theta)/Mg(\theta)$ accept θ in the sample. Otherwise discard it.

Repeat Steps (1) through (3) until enough values of θ have been accepted to form a sample of the desired size.

This algorithm is the basis for many random-number generation routines in software packages. It is applied as follows to the generation of a sample from the posterior distribution for λ . The equations are worked out here, and the algorithm for the posterior distribution is restated at the end.

Let f be the posterior density and let g be the prior density. Then Equation 6.4 states that the ratio $f(\lambda)/g(\lambda)$ is proportional to the likelihood, which is maximized, by definition, when λ equals the maximum likelihood estimate, x/t . That is, the ratio of interest is

$$f(\lambda)/g(\lambda) = Ce^{-\lambda t}(\lambda t)^x$$

for some constant C . This is maximized when λ equals x/t . Therefore, define $M = \max[f(\lambda)/g(\lambda)] = Ce^{-x}x^x$. The condition in Step (3) above is equivalent to

$$u \leq [Ce^{-\lambda t}(\lambda t)^x] / [Ce^{-x}x^x] = [e^{-\lambda t}(\lambda t)^x] / [e^{-x}x^x].$$

The constant cancels in the numerator and denominator, so we do not need to evaluate it! It would have been possible to work with $m = M/C$, and the calcula-

tions would have been simpler. This rewritten form of the algorithm, for Poisson data, is given here.

If $x > 0$, define $m = e^{-x}x^x$. If $x = 0$, define $m = 1$.

The steps of the algorithm are:

- (1) Generate a random λ from the prior distribution.
- (2) Generate u from a uniform distribution, $0 \leq u \leq 1$.
- (3) If $u \leq e^{-\lambda t}(\lambda t)^x/m$, accept λ in the sample. Otherwise discard λ .

Repeat Steps (1) through (3) until a sample of the desired size is found.

Intuitively, this algorithm generates possible values of λ from the prior distribution, and discards most of those that are not very consistent with the data. The result is a sample from the posterior distribution.

6.2.2.6.3 More Complicated Random Sampling

All-purpose Bayesian update programs can be used for the present simple problem. For example, the program BUGS¹ (Bayesian inference Using Gibbs Sampling) performs **Markov chain Monte Carlo (MCMC)** sampling. This package is intended for complicated settings, such as those described in Chapters 7 and 8. Using it here is like using the proverbial cannon to kill a mosquito. Nevertheless, the program is free, and very flexible, and can be used here. It is available for download at

<http://www.mrc-bsu.cam.ac.uk/bugs/>

and is described more fully in Sections 7.2.3 and 8.2.3.3.3 of this handbook.

6.2.2.7 Examples Involving Nonconjugate Priors

These techniques will be illustrated with the following example, from Appendix J-4 of Poloski et al. (1999a).

¹ Mention of specific products and/or manufacturers in this document implies neither endorsement or preference, nor disapproval by the U.S. Government or any of its agencies of the use of a specific product for any purpose.

6.

Example 6.3 Small-break LOCAs.

No small-break loss-of-coolant accidents (SBLOCAs) have occurred in 2102 reactor-calendar-years at U.S. nuclear power plants. The WASH-1400 (NRC 1975) distribution for the frequency of this event was lognormal with median $1E-3$ and error factor 10.

6.2.2.7.1 Example with Lognormal Prior

Poloski et al. (1999a) use the WASH-1400 distribution as a prior, and update it with the 2102 years of data.

The resulting posterior distribution was sampled 100,000 times using the method described in Section 6.2.2.6.2 above, and the mean was found. Then the values were arranged in increasing order, and the percentiles of the sample were found. All this took less than 15 seconds in 1999 on a 166 MHz computer. Based on the mean and percentiles of the sample, the mean of the posterior distribution is $3.5E-4$, and the 90% posterior credible interval is ($4.5E-5$, $9.8E-4$).

To illustrate the method of Section 6.2.2.6.3, the distribution was also sampled using BUGS. Figure 6.13 shows the script used for running BUGS.

```
model
{
  mu <- lambda*rxys
  x ~ dpois(mu)
  lambda ~ dlnorm(-6.908, 0.5104)
}
list(rxys=2102, x=0)
```

Figure 6.13 Script for analyzing Example 6.3 using BUGS.

The section in curly brackets defines the model. Note that $<-$, intended to look like a left-pointing arrow, is used to define quantities in terms of other quantities, and \sim is used to generate a random quantity from a distribution. Thus, X is a Poisson random variable with mean μ , with $\mu = \lambda \times rxys$. The prior distribution of λ is lognormal. The parameters given in the script arise as follows. BUGS parameterizes the normal in terms of the mean and inverse of the variance, for reasons explained in Section 6.7.1.2.1. It parameterizes the lognormal distribution

using the parameters of the underlying normal. It is shown below that a lognormal with median $1E-3$ and error factor 10 corresponds to an underlying normal with mean -6.980 and standard deviation 1.3997 . Therefore the inverse of the variance is $1/1.3997^2 = 0.5104$.

The line beginning "list" defines the data, 0 events in 2102 reactor years. BUGS also requires an initial value for λ , but generated it randomly.

When BUGS generated 100,000 samples, the mean, 5th percentile, and 95th percentile of λ were $3.5E-4$, $4.5E-5$, and $9.8E-4$, just as found above.

6.2.2.7.2 Example with "Moment-Matching" Conjugate Prior

Conjugate priors have appeal: Some people find algebraic formulas tidier and more convenient than brute-force computer calculations. Also, when a PRA program requests a distribution for a parameter, it is usually easier to enter a distributional form and a couple of parameters than to enter a simulated distribution.

Therefore, a nonconjugate prior is sometimes replaced by a conjugate prior having the same mean and variance. This method is carried out here with the above example.

Begin by finding the gamma prior with the same moments as the above lognormal prior. As explained in Appendix A.7.3, the median, error factor, and moments of the lognormal distribution are related to μ and σ of the underlying normal distribution of $\ln\lambda$ as follows.

$$\begin{aligned} \text{median}(\lambda) &= \exp(\mu) \\ \text{EF}(\lambda) &= \exp(1.645\sigma) \\ \text{mean}(\lambda) &= \exp(\mu + \sigma^2/2) \\ \text{var}(\lambda) &= [\text{median}(\lambda)]^2 \cdot \exp(\sigma^2) \cdot [\exp(\sigma^2) - 1] \end{aligned}$$

The lognormal prior has median $1.0E-3$, and error factor 10. Solving the first two equations yields

$$\begin{aligned} \mu &= -6.907755 \\ \sigma &= 1.399748 \end{aligned}$$

Substituting these values into the second two equations yields

$$\text{mean}(\lambda) = 2.6635E-3$$

$$\text{var}(\lambda) = 4.3235\text{E}-5 .$$

Now the gamma distribution must be found with this mean and this variance. The formulas for the moments of a gamma distribution were given in Section 6.2.2.4.1 and in Appendix A.7.6:

$$\begin{aligned} \text{mean} &= \alpha/\beta \\ \text{variance} &= \alpha/\beta^2 . \end{aligned}$$

Therefore,
 $\alpha = \text{mean}^2/\text{variance} = 0.164$
 $\beta = \text{mean}/\text{variance} = 61.6 \text{ reactor-years}.$

Warning flags should go up, because α is considerably smaller than 0.5. Nevertheless, we carry out the example using this gamma distribution as the prior. The update formulas yield

$$\begin{aligned} \alpha_{\text{post}} &= 0 + 0.164 = 0.164 \\ \beta_{\text{post}} &= 2102 + 61.6 = 2164 \text{ reactor-years} . \end{aligned}$$

The posterior mean is $7.6\text{E}-5$, and a 90% credible interval is $(3.4\text{E}-12, 4.1\text{E}-4)$, all with units events per reactor-year.

6.2.2.7.3 Comparison of Example Analyses

The two posterior distributions do not agree closely, as will be discussed below. If the shape parameter α of the gamma prior had been larger, the two prior distributions would have had more similar percentiles, and the two posterior distributions likewise would have agreed better. As it is, however, the two analyses are summarized in Table 6.5.

Table 6.5 Posterior distributions from two analyses.

Prior	Mean	90% Interval
Lognormal	$3.5\text{E}-4$	$(4.5\text{E}-5, 9.8\text{E}-4)$
Gamma	$7.6\text{E}-5$	$(3.4\text{E}-12, 4.1\text{E}-4)$

The most notable difference between the two posterior distributions is in the lower endpoints, the 5th percentiles, which differ by many orders of magnitude. This is explained, to some extent, by graphical comparisons. Figures 6.14 and 6.15 show the prior cumulative distributions. When plotted on an ordinary scale in Figure 6.14, the two prior distributions look fairly similar, although the gamma distribution

seems to put more probability near zero. The differences become much more obvious when the two prior distributions are plotted on a logarithmic scale in Figure 6.15. These differences between the two prior distributions are present in spite of the fact that the two priors have equal means and equal variances.

The two resulting posterior distributions are also quite different in the lower tail, as shown in Figure 6.16, and this difference is especially clear when the distributions are plotted on a log scale, as shown in Figure 6.17.

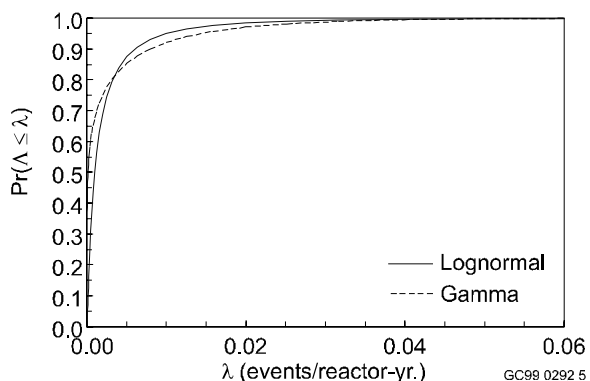


Figure 6.14 Two prior distributions having same means and variances.

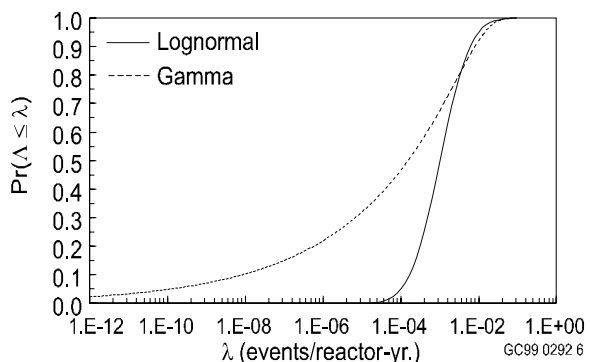


Figure 6.15 Same prior distributions as in previous figure, with λ plotted on a logarithmic scale.

6.

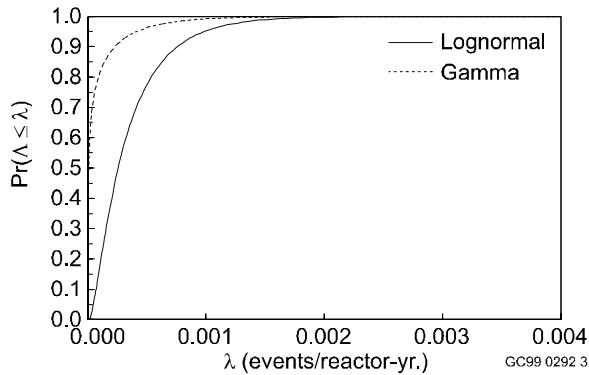


Figure 6.16 Two posterior distributions, from priors in previous figures.

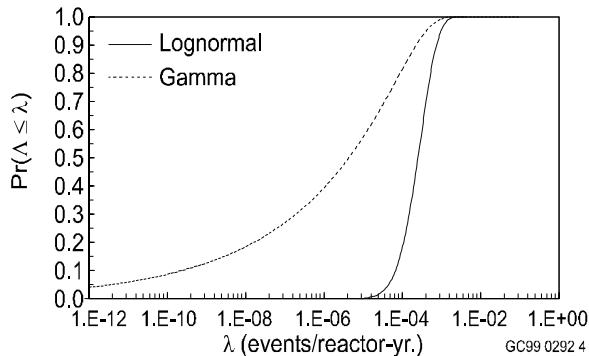


Figure 6.17 Same posterior distributions as in previous figure, with λ plotted on logarithmic scale.

Incidentally, these illustrations use cumulative distributions instead of densities, for an important reason. Cumulative distributions simply show probabilities, and so can be plotted with the horizontal scale either linear or logarithmic. Alternatively, the density of $\ln(\lambda)$ could be plotted against $\ln(\lambda)$, but the density of $\ln(\lambda)$ is not the same function as the density of λ , as explained in Appendix A.4.7.

6.2.3 Model Validation

Model validation should go hand in hand with parameter estimation. Philosophically, it would seem natural first to confirm the form of the model and second to estimate the parameters of that model. However, typically one can perform goodness-of-fit tests and other validations of a model only after the model has been fully specified, that is, only after the form of the model has been assumed *and* the corresponding parameters have been estimated. Because parameter-estima-

tion is built into most model-validation procedures, it was presented first.

It is often wise not to stop the analysis with just estimating the parameters. Foolish results have been presented by analysts who estimated the parameters but did not thoroughly check that the assumptions of the model were correct. This section presents ways to check the model assumptions. Much of this material is taken from an INEEL report by Engelhardt (1994).

The Poisson process was introduced in Section 2.2.2. The three assumptions were listed there: constant event occurrence rate, no simultaneous events, and independent time periods. These assumptions are considered here.

The assumption of constant rate is considered in the next two sections, first where the alternative possibility is that different data sources may have different values of λ but in no particular order, and then where the alternative possibility is that a time trend exists. Both graphical methods and formal statistical hypothesis tests are given for addressing the issues. The assumption of no exactly simultaneous events is then discussed from the viewpoint of examining the data for common-cause events. Finally the assumption of independent time intervals is considered, and some statistical tests of the assumption are given.

When Bayesian methods are used, one must also examine whether the data and the prior distribution are consistent. It makes little sense to update a prior with data, if the data make it clear that the prior belief was just plain wrong. That topic constitutes the final subsection of the present section.

6.2.3.1 Poolability of Data Subsets

Assumption 1 in Section 2.2.2 implies that there is one rate λ for the entire process. The correctness of such an assumption can be investigated by analyzing subsets of the data and comparing the estimates of λ for the various subsets.

Example 2.2 described LOSP events during shutdown. For this section, consider a portion of that example. The entire data set could be used, but to keep the example from being too cumbersome we arbitrarily

restrict it to five plants at three sites, all located in one state.

An obvious question concerns the possibility of different rates for different plants. A general term used in this handbook will be **data subsets**. In Example 6.4, five subsets are shown, corresponding to plants. In other examples the subsets could correspond to years, or systems, or any other way of splitting the data. For initiating events, each subset corresponds to one **cell** in the table, with an event count and an exposure time.

Sometimes, data subsets can be split or combined in reasonable ways. For example, if the subsets were time periods, the data could be partitioned into decades or years or months. The finer the division of the cells, the more sparse the data become within the cells. Too fine a partition allows random variation to dominate within

Example 6.4 Shutdown LOSP events at five plants, 1980-96.

During 1980-1996, five plants experienced eight LOSP events while in shutdown. These were events from plant-centered causes rather than external causes. The data are given here.

Plant code	Events	Plant shutdown years
CR3	5	5.224
SL1	0	3.871
SL2	0	2.064
TP3	2	5.763
TP4	1	5.586
Totals	8	22.508

each cell, but too coarse a partition may hide variation that is present within individual cells. In the present simple example, the most reasonable partition is into plants. Analysis of more complicated data sets may require examination of many partitionings.

First, a graphical technique is given, to help the analyst understand what the data set shows. Then, a formal statistical procedure is presented, to help quantify the

strength of the evidence for patterns seen in the graphical investigation.

6.2.3.1.1 Graphical Technique

To explore the relations between cells, identify the cells on one axis. Then, for each cell, plot a point estimate of λ and an interval estimate of λ against the other axis. Patterns such as trends, outliers, or large scatter are then made visible.

In Example 6.4, the cells are plants. The data set from each plant was analyzed separately, using the tools of Section 6.2.1. The graph in Figure 6.18 shows the maximum likelihood estimate and a confidence interval for each plant, plotted side by side. For this handbook, the plot was produced with a graphics software package, although a hand-drawn sketch would be adequate to show the results.

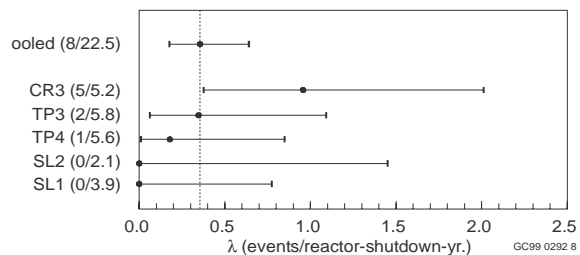


Figure 6.18 MLEs and 90% confidence intervals for λ , based on each plant's data and based on pooled data from all the plants.

The confidence interval for the pooled data is also shown. Take care, however: this interval is only valid if all the plants have the same λ , which is what must be decided! Nevertheless, the interval and point estimate for the pooled data give a useful reference for comparisons with the individual plants. For this reason a vertical dotted line is drawn through the mean of the pooled data.

Note that the plants are displayed not in alphabetical order, which is a meaningless order for the event rate, but in order of decreasing $\hat{\lambda}$. (When two plants have the same MLE, as do SL1 and SL2, the upper confidence limit is used to determine the order.) Experience has shown that such a descending order assists the eye in making comparisons. Cleveland (1985, Chap. 3.3) discusses this and other ways of ordering data.

6.

CR3 appears somewhat high compared to the others — although there is considerable overlap of the intervals, the lower confidence limit for CR3 is just barely higher than the MLE for the utility as a whole. Of course, the picture might give a different impression if slightly different intervals were used: 95% confidence intervals instead of 90% confidence intervals, or Bayes intervals with the Jeffreys noninformative prior instead of confidence intervals. From the graph alone, it is difficult to say whether the data can be pooled.

A graph like this should not be used to draw naive conclusions without also using a formal statistical test. For example, if many confidence intervals are plotted, based on data sets generated by the same λ , a few will be far from the others because of randomness alone. This was seen in Figure 6.3, where all the variation was due to randomness of the data, and some intervals did not overlap some others at all. Thus, an outlying interval does not prove that the λ s are unequal. This same statement is true if other intervals are used, such as Bayes credible intervals based on the noninformative prior. The issue is the random variability of data, not the kind of interval constructed.

Conversely, if there are only a few intervals, intervals that just barely overlap can give strong evidence for a difference in the λ s.

To quantify the strength of the evidence against poolability, a formal statistical procedure is given in the next subsection. The graph gives an indication of what the test might show, and helps in the interpretation of the test results. If the statistical test turns out to find a statistically significant difference between plants, it is natural then to ask what kind of difference is present. Figure 6.17 shows that most of the plants appear similar, with only one possible outlier. An unusually long interval, such as that seen in Figure 6.18 for SL2, is generally associated with a smaller exposure time. The picture provides insight even though it does not give a quantitative statistical test.

6.2.3.1.2 Statistical Test

The Chi-Squared Test. To study whether the rate is the same for different cells, use a **chi-squared test**. Many statistics texts, such as Bain and Engelhardt (1992, Chapter 13), discuss this test, and many soft-

ware packages perform the chi-squared test. It is presented here in enough detail so that the reader could perform the calculations by hand if necessary, because it is instructive to see how the test works.

Let the null hypothesis be
 H_0 : λ is the same in all the data subsets.

In the present application the data subsets are the five plants. The method is to see what kind of data would be expected when λ really is constant, and then to see how much the observed counts differ from the expected counts. If the difference is small, the counts are consistent with the hypothesis H_0 that the rate is constant. If, instead, the difference is large, the counts show strong evidence against H_0 .

If H_0 is true, that is, if λ is really the same for all the plants, then the estimate (MLE) of λ is $\hat{\lambda} = x/t$. The estimate of the expected count is built from this quantity. Let x_j be the number of observed events in the j th cell, the j th plant in the present example. Assuming the hypothesis of a single rate λ , an estimate of the expected count for the j th cell is simply $e_j = \hat{\lambda}t_j$.

In Example 6.4, the estimate of λ is $8/22.508 = 0.355$ events per shutdown-year. Therefore, the expected count for CR3 is the estimate of λ times the exposure time for CR3, $0.335 \times 5.224 = 1.857$ events. Table 6.6 is an extension of the original table given in Example 6.4, showing the quantities needed for the calculation.

Table 6.6 Quantities for calculation of chi-squared test.

Cell code	x_j	t_j	e_j
CR3	5	5.224	1.857
SL1	0	3.871	1.376
SL2	0	2.064	0.734
TP3	2	5.763	2.048
TP4	1	5.586	1.985
Totals	8	22.508	8.000

The total of the expected counts agrees with the total of the observed counts, except possibly for small round-off error.

The test for equality of rates that is considered here is based on the following calculated quantity,

$$X^2 = \sum_j (x_j - e_j)^2 / e_j,$$

sometimes called the **Pearson chi-squared statistic**, after its inventor, Karl Pearson, or simply the **chi-squared statistic**. The notation became standard long before the custom developed of using upper-case letters for random variables and lower-case letters for numbers. In the discussion below, the context must reveal whether X^2 refers to the random variable or the observed value.

Observe that X^2 is large if the x_j s (observed counts) differ greatly from the e_j s (expected values when H_0 is true). Conversely, X^2 is small if the observed values are close to the expected values. This statement is made more precise as follows. When H_0 is true and the total count is large, the distribution of X^2 has a distribution that is approximately chi-squared with $c - 1$ degrees of freedom, where c is the number of cells. If the calculated value of X^2 is large, compared to the chi-squared distribution, there is strong evidence that H_0 is false; the larger the X^2 value, the stronger the evidence.

For the data of Table 6.4, $X^2 = 7.92$, which is the 0.906th quantile of the chi-squared distribution with 4 degrees of freedom. The next subsection discusses the interpretation of this.

Interpretation of Test Results. Suppose, for any example with 5 cells, that X^2 were 9.8. A table of the chi-squared distribution shows that 9.488 is the 0.95th quantile of the chi-squared distribution with 4 degrees of freedom, and 11.14 is the 0.975th quantile. After comparing X^2 to these values, we would conclude that the evidence is strong against H_0 , but not overwhelming. The full statement is

- If H_0 is true, that is, if all the cells have the same λ , the chance of seeing such a large X^2 is less than 0.05 but more than 0.025.

Common abbreviated ways of saying this are

- We reject H_0 at the 5% **significance level**, but not at the 2.5% significance level.
- The difference between cells is **statistically significant at the 0.05 level**, but not at the 0.025 level.
- The **p-value** is between 0.05 and 0.025.

There will be some false alarms. Even if λ is exactly the same for all the cells, sometimes X^2 will be large, just from randomness. It will be greater than the 95th percentile for 5% of the data sets, and it will be greater than the 99th percentile for 1% of the data sets. If we observed such a value for X^2 , we would probably decide that the data could not be pooled. In that case, we would have believed a false alarm and made the incorrect decision. Just as with confidence intervals, we cannot be sure that this data set is not one of the rare unlucky ones. But following the averages leads us to the correct decision most of the time.

If, instead, X^2 were 4.1, it would be near the 60th percentile of the chi-squared distribution, and therefore be in the range of values that would be expected under H_0 . We would say the observed counts are consistent with the hypothesis H_0 , or H_0 cannot be rejected, or the evidence against H_0 is weak. We would not conclude that H_0 is true, because it probably is not exactly true to the tenth decimal place, but the conclusion would be that H_0 cannot be rejected by the data.

In fact, for the data of Table 6.6, X^2 equals 7.92, which is the 0.906th quantile of the chi-squared distribution with 4 degrees of freedom. That means: if all five plants have the same event rate, there is a 9.4% probability of seeing such a large value of X^2 . The evidence against H_0 is not convincingly strong. CR3 might be suspected of having a higher event rate, but the evidence is not strong enough to prove this.

The traditional cut-off is 5%. The difference between cells is called **statistically significant**, with no qualifying phrase, if it is significant at the 0.05 level. This is tradition only, but it is very widely followed.

In actual data analysis, do not stop with the decision that a difference is or is not statistically significant. Do not even stop by reporting the p-value. That may be acceptable if the p-value is very small (much less than 0.05) or very large (much larger than 0.05). In many cases, however, statistical significance is far from the

6.

whole story. Engineering significance is just as important.

To illustrate this, consider a possible follow-up to the above statistical analysis of Example 6.4. As mentioned, the statistical evidence against poolability is not strong, but some might consider it borderline. Therefore, a thorough analysis would ask questions such as:

- Are there engineering reasons for expecting CR3 to have a different event rate than the other plants do, either because of the hardware or because of procedures during shutdown? (Be warned that it is easy to find justifications in hindsight, after seeing the data. It might be wise to hide the data and ask these questions of a different knowledgeable person.)
- What are the consequences for the PRA analysis if the data are pooled or if, instead, CR3 is treated separately from the other plants? Does the decision to pool or not make any practical difference?

Required Sample Size. The above considerations are valid if the total count is "large," or more precisely, if the e_j s are "large." If the e_j s are small, the chi-squared distribution is not a good approximation to the distribution of X^2 . Thus, the user must ask how large a count is necessary for the chi-squared approximation to be adequate. An overly conservative rule is that each expected cell-count, e_j , should be 5.0 or larger. Despite its conservatism, this rule is still widely used, and cited in the statistical literature and by some software packages.

A readable discussion of chi-squared tests by Moore (1986, p.71) is applicable here. Citing the work of Roscoe and Byars (1971), the following recommendations are made:

- (1) With equiprobable cells, the average expected frequency should be at least 1 when testing at the 0.05 level. In other words, use the chi-squared approximation at the 5% level when $x/c \geq 1$, where x is the number of events and c is the number of cells. At the 1% level, the chi-squared approximation is recommended if $x/c \geq 2$.
- (2) When the cells are not approximately equiprobable, the average expected frequencies in (1) should

be doubled. Thus, the recommendation is that at the 5% level $x/c \geq 2$, and at the 1% level $x/c \geq 4$.

Note that in rules (1) and (2) above, the recommendation is based on the average rather than the minimum expected cell-count. As noted in another study by Koehler and Larntz (1980), any rule such as (2) may be defeated by a sufficiently skewed assignment of cell probabilities.

Roscoe and Byars also recommend when $c = 2$ that the chi-squared test should be replaced by the test based on the exact binomial distribution of X_1 conditional on the total event count. For example, if the two cells had the same exposure times, we would expect that half of the events would be generated in each cell. More generally, if

- the two cells have exposure times t_1 and t_2
- a total of x events are observed
- λ is the same for both cells

then, conditional on x , X_1 has a binomial(n, p) distribution, with $p = t_1/(t_1 + t_2)$. Exact binomial tests are discussed by Bain and Engelhardt (1992, p.405).

Example 6.4 has $x = 8$ and $c = 5$. The cells are not equiprobable, that is, e_j is not the same for all cells, because the plants did not all have the same exposure time. Nevertheless, the expected cell counts differ from each other by at most a factor of two. This is not a large departure from equiprobability, as differences of an order of magnitude would be. Because $x/c = 1.6$, and the calculated significance level is about 10%, the sample size is large enough for the chi-squared approximation to be adequate. The conclusions reached earlier still stand. If, on the other hand, the sample size had been considerably smaller, one would have to say that the p-value is *approximately* given by the chi-squared distribution, but that the exact p-value has not been found.

If the expected cell-counts are so small that the chi-squared approximation is not recommended, the analyst can pool data in some "adjacent cells," thereby increasing the expected cell-counts.

In the Example 6.4, suppose that there were engineering reasons for thinking that the event rate is similar at units at a single site. Then the sister units might be pooled, transforming the original table of Example 6.4 into Table 6.7 here.

We repeat, this pooling of cells is not required with the actual data, but it could be useful if (a) the cell counts were smaller and (b) there were engineering reasons for believing that the pooled cells are relatively homogeneous, that is, the event rates are similar for both units at a site, more similar than the event rates at different sites.

Table 6.7 Shutdown LOSP events at three sites, 1980-96.

Site code	Events	Plant shutdown years
CR	5	5.224
SL	0	5.935
TP	3	11.349

Generally speaking, a chi-squared test based on a larger number of cells will have better power for detecting when rates are not equal, but this also makes it more difficult to satisfy guidelines on expected cell-counts for the chi-squared approximation. Thus, it is sometimes necessary to make a compromise between expected cell counts and the number of cells.

Options involving the exact distribution of X^2 are also possible. The most widely known commercial software for calculating the exact p-value is StatExact (1999).

6.2.3.2 No Time Trend

The chi-squared method given above does not use any ordering of the cells. Even if the test were for differences in years, say, the test would not use the natural ordering by calendar year or by plant age. When there is a meaningful order to the data subsets, it may be useful to perform additional analyses. The analysis given above is valid, but an additional possible analysis, making use of time order, is considered now.

The methods will be illustrated with Example 6.5.

6.2.3.2.1 Graphical Techniques

Confidence-Interval Plot. First, the same kind of plot that was used in the previous subsection can be used here. The time axis is divided into cells, or **bins** in the terminology of some authors. For example, if the time

span is divided into calendar years, the counts and reactor-critical-years for Example 6.5 are given in Table 6.8.

Example 6.5 Unplanned HPCI demands.

Grant et al. (1995, Table B-5) list 63 unplanned demands for the HPCI system to start at 23 BWRs during 1987-1993. The demand dates are given in columns below, in format MM/DD/YY.

01/05/87	08/03/87	03/05/89	08/16/90	08/25/91
01/07/87	08/16/87	03/25/89	08/19/90	09/11/91
01/26/87	08/29/87	08/26/89	09/02/90	12/17/91
02/18/87	01/10/88	09/03/89	09/27/90	02/02/92
02/24/87	04/30/88	11/05/89	10/12/90	06/25/92
03/11/87	05/27/88	11/25/89	10/17/90	08/27/92
04/03/87	08/05/88	12/20/89	11/26/90	09/30/92
04/16/87	08/25/88	01/12/90	01/18/91	10/15/92
04/22/87	08/26/88	01/28/90	01/25/91	11/18/92
07/23/87	09/04/88	03/19/90	02/27/91	04/20/93
07/26/87	11/01/88	03/19/90	04/23/91	07/30/93
07/30/87	11/16/88	06/20/90	07/18/91	
08/03/87	12/17/88	07/27/90	07/31/91	

Table 6.8 HPCI demands and reactor-critical-years.

Calendar year	HPCI demands	Reactor-critical-years
1987	16	14.63
1988	10	14.15
1989	7	15.75
1990	13	17.77
1991	9	17.11
1992	6	17.19
1993	2	17.34

This table has the same form as in Example 6.4, showing cells with events and exposure times. The relevant exposure time is reactor-critical-years, because the HPCI system uses a turbine-driven pump, which can only be demanded when the reactor is producing steam. The counts come from the tabulated events of Example 6.5, and the critical-years can be constructed from information in Poloski et al. (1999a). The variation in critical-years results from the facts that several reactors were shut down for extended periods, and one reactor did not receive its low power license until 1989.

This leads to a plot similar to Figure 6.18, showing the estimated value of the demand frequency, λ , and

6.

a confidence interval for each year. This is shown in Figure 6.19.

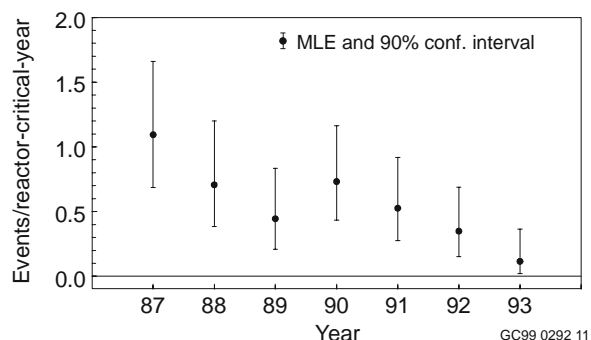


Figure 6.19 MLEs and 90% confidence intervals for λ , each based on data from one calendar year.

Figure 6.19 seems to indicate a decreasing trend in the frequency of HPCI demands. However, the picture does not reveal whether the apparent trend is perhaps merely the result of random scatter. To answer that question, a formal statistical test is necessary, quantifying the strength of the evidence. Such tests will be given in Section 6.2.3.2.2.

Cumulative Plot. Figure 6.19 required a choice of how to divide the time axis into cells. A different plot, given next, does not require any such choice, if the dates of the events are recorded. Plot the cumulative event count at the n event dates.

Figure 6.20 shows this for Example 6.5. The events are arranged in chronological order, and the cumulative count of events is plotted against the event times.

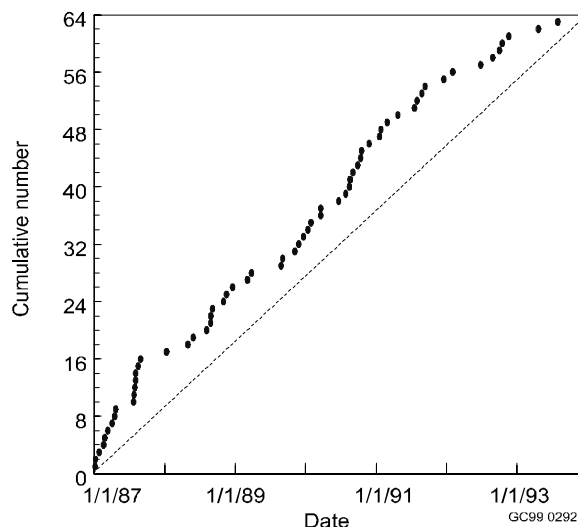


Figure 6.20 Cumulative number of HPCI demands, by date.

The **slope** of a string of plotted points is defined as the vertical change in the string divided by the horizontal change, $\Delta y/\Delta x$. This is the familiar definition of slope from mathematics courses. In the plot given here, the horizontal distance between two points is elapsed time, and the vertical distance is the total number of events that occurred during that time period. Therefore,

$$\text{slope} = (\text{number of events})/(\text{elapsed time}),$$

so the slope is a graphical estimator of the event frequency, λ . A constant slope, or a straight line, indicates a constant λ . Changes in slope indicate changes in λ : if the slope becomes steeper, λ is increasing, and if the slope becomes less steep, λ is decreasing.

In Example 6.4 the time axis represents calendar years. Because the relevant frequency is events per reactor-critical-year, it would be better to plot the time axis in terms of total reactor-critical-years from the start of 1987. However, it is somewhat difficult to calculate the reactor-critical-years preceding any particular event, or equivalently, the reactor-critical-years between successive events. Therefore, simple calendar years are used. This is adequate if the number of reactors operating at any time is fairly constant, because then the rate per reactor-critical-year remains roughly proportional to the rate per

industry-calendar year. In the present case, as shown by Table 6.8, later calendar-years correspond to a few more critical-years than do early calendar-years.

The slope in Figure 6.20 is steepest on the left, and gradually lessens, so that the plot is rising fastest on the left and more gently on the right. More HPCI demands are packed into a time interval on the left than into a time interval of the same length on the right. This indicates that the frequency of unplanned HPCI demands was decreasing during the time period of the study. Thus, this figure leads to the same general conclusion as does Figure 6.19. Figure 6.20 shows more detail, with the individual events plotted, but it is less accurate in this example because we have not gone to the work of plotting events versus reactor-critical time.

It is important that the horizontal axis cover the entire data-collection period and not stop at the final event. In Figure 6.20, the lack of events during the last half of 1993 contributes to the overall curvature of the plot.

If the frequency is constant, the plot should follow a roughly straight line. For comparison, it is useful to show a straight diagonal line, going from height 0 at the start of the data collection period to height $n + 1$ at the end of the data collection period, where n is the number of data points.

In Figure 6.20, the diagonal line is shown as a dotted line, rising from height 0 on the left to height $n + 1 = 64$ on the right.

As mentioned above, the early calendar years contain fewer reactor-critical-years than do the later calendar years. Therefore, the time axis in Figure 6.20 would reflect reactor-critical-years more accurately if the left end of the axis were compressed slightly or the right end were stretched slightly. The effect would be to increase the curvature of the plot, making it rise more quickly on the left and more slowly on the right.

A cumulative plot contains random bounces and clusters, so it is not clear whether the observed pattern is more than the result of randomness. As always, a formal statistical test will be needed to measure the strength of the evidence against the hypothesis of constant event frequency.

6.2.3.2.2 Statistical Tests for a Trend in λ

The Chi-Squared Test. This is the same test as given in Section 6.2.3.1.2, only now the cells are years or similar divisions of time.

In Example 6.5, the p-value is 0.009, meaning that a random data set with constant λ would show this much variability with probability only 0.9%. Two points are worth noting.

- The chi-squared test makes no use of the order of the cells. It would give exactly the same conclusion if the intervals in Figure 6.19 were scrambled in a random order instead of generally decreasing from left to right.
- The calculated p-value is accurate enough to use, by the guidelines of Section 6.2.3.1.2, because the number of events is 63, and the number of cells is 7, so $x/c = 63/7 = 9$. Even splitting the cells into 6-month periods or smaller periods would be justified.

Chapter 7 will take Figure 6.19, fit a trend, and perform an additional test based on the fit; see Sections 7.2.3 and 7.2.4. Therefore, the chi-squared test is not discussed further here.

The Laplace Test. This test does not use the binning of times into cells, but instead uses the exact dates. In the example, there are 63 occurrences of events during a seven-year period. In general, consider a time interval $[0, L]$, and suppose that during this period n events occur at successive random times T_1, T_2, \dots, T_n . Although the number of occurrences, n , is random when the plants are observed for a fixed length of time L , we condition on the value of n , and so treat it as fixed. Consider the null hypothesis

H_0 : λ is constant over time .

Consider the alternative hypothesis

H_1 : λ is either an increasing or a decreasing function of time.

This hypothesis says that the events tend to occur more at one end of the interval. A test that is often used is based on the mean of the failure times, $\bar{T} = \sum_i T_i / n$. The intuitive basis for the test is the following. If λ is

6.

constant, about half of the events should occur before time $L/2$ and half afterwards, and the average event time should be close to $L/2$. On the other hand, if λ is decreasing, more events are expected early and fewer later, so the average event time should be smaller than $L/2$. Similarly, if λ is increasing, the average event time is expected to be larger than $L/2$. Therefore, the test rejects H_0 if \bar{T} is far from $L/2$. Positive values of the difference $\bar{T} - L/2$ indicate an increasing trend, and negative values indicate a decreasing trend.

When H_0 is true, \bar{T} has expected value $L/2$ and variance $L^2/(12n)$. The resulting test statistic is

$$U = \frac{\bar{T} - L/2}{L/\sqrt{12n}} .$$

The statistic U is approximately standard normal for $n \geq 3$. A test of H_0 at significance level 0.05 versus an increasing alternative,

H_1 : λ is increasing in time ,

would reject H_0 if $U \geq 1.645$. A 0.05 level test versus a decreasing alternative,

H_1 : λ is decreasing in time ,

would reject H_0 if $U \leq -1.645$. Of course, ± 1.645 are the 0.95th and 0.05th quantiles, respectively, of the standard normal distribution. A two-sided test, that is, a test against the original two-sided alternative hypothesis, at the 0.10 level would reject H_0 if $|U| \geq 1.645$.

This test, generally known as the "Laplace" test, is discussed by Cox and Lewis (1978, p. 47). The Laplace test is known to be good for detecting a wide variety of monotonic trends, and consequently it is recommended as a general tool for testing against such alternatives.

Let us apply the Laplace test to the HPCI-demand data of Example 6.4. First, the dates must be converted to times. The first event time is 0.011 years after January 1, 1987, the final event is 6.581 years after the starting date, and the other times are calculated similarly. Here, a "year" is interpreted as a 365-day year. The total number of 365-day years is $L = 7.00$. The mean of the event times can be

calculated to be 2.73. Therefore, the calculated value of U is

$$\frac{2.73 - 3.5}{7.00 / \sqrt{12 \times 63}} = -3.02 .$$

This is statistically very significant. The value 3.02 is the 0.001th quantile of the standard normal distribution. Thus, the evidence is very strong against a constant demand rate, in favor instead of a decreasing demand rate. Even against the two-sided hypothesis

H_1 : λ is increasing or decreasing in time ,

the p-value is $\Pr(|U| > 3.02) = 0.002$.

In the example, the Laplace test statistic was calculated in terms of calendar time instead of reactor-critical-time. As remarked earlier, using reactor-critical-time would increase the curvature of the plot in Figure 6.20. A similar argument shows that using reactor-critical-time in computing U would increase the strength of the evidence against the hypothesis of a constant demand rate. However, the computations would be very tedious. That is an advantage of the chi-squared test, because it is typically easier to find the exact relevant exposure time for blocks of time, such as years, than for each individual event.

In the example, the result of the Laplace test agrees with the result from the chi-squared test, but is more conclusive. The chi-squared test gave a p-value of 0.009, meaning that if H_0 is true, the cells would appear so different from each other with probability only 0.009. The Laplace test gives a p-value of 0.002.

The chi-squared and Laplace tests differ because they are concerned with different alternatives to H_0 . The chi-squared test is concerned with any variation from cell to cell (from year to year in the example). If the event rate goes up and down erratically, that is just as much evidence against H_0 as if the event rate decreases monotonically. The Laplace test, on the other hand, is focused on the alternative of a trend. It has more power for detecting trends, but no power at all for detecting erratic changes upward and downward.

Other tests exist in this setting. See Ascher and Feingold (1984, page 80) and Engelhardt (1994, p. 19) for details.

6.2.3.3 No Multiple Failures

The second assumption of the Poisson process is that there are no exactly simultaneous failures. In practice this means that common-cause failures do not occur. In most situations, common-cause failures will occur from time to time. This was seen in some of the examples discussed in Section 2.2. However, if common-cause events are relatively infrequent, their effect on the validity of the Poisson model can normally be ignored.

No statistical methods are given here to examine whether common-cause events can occur. Instead, the analyst should think of the engineering reasons why common-cause events might be rare or frequent, and the data should be examined to discover how frequent common-cause events are in practice.

In Example 6.5, HPCI demands, it is reasonable that common-cause events could occur only at multiple units at a single site. There was one such pair of events in the data, with HPCI demands at Hatch 1 and Hatch 2, both on 08/03/87. Examination of the LERs reveals that the demands occurred from different causes. They happened at different times, and so were not exactly simultaneous. The conclusion is that common causes may induce exactly simultaneous events, but they are infrequent.

If common-cause events are relatively frequent, so that they cannot be ignored, it might be necessary to perform two analyses, one of the "independent", or not-common-cause, events, and one of the common-cause occurrences. For each type of event, the event frequency, λ , could be estimated. Then an additional analysis would be necessary to characterize the number of separate events for each common-cause occurrence.

6.2.3.4 Independence of Disjoint Time Periods

This section is probably less important than the others, and of interest only to truly dedicated readers. Others should skip directly to Section 6.2.3.5.

The final assumption of the Poisson model is that event occurrences in disjoint time periods are statistically independent. This should first be addressed by careful thinking, similar to that in the examples of Section 2.2.

However, the following statistical approach may also be useful.

One possible type of dependence would be if events tended to cluster in time: large between-event times tended to occur in succession, or similarly small ones tended to occur in succession. For example, suppose that a repair is done incorrectly several times in succession, leading to small times between failures. The occurrence of a failure on one day would increase the probability of a failure in the next short time period, violating the Poisson assumption. After the problem is diagnosed, the personnel receive training in proper repair procedures, thereafter resulting in larger times between failures.

To illustrate the ideas, an example with no trend is needed. The shutdown LOSP events introduced in Section 2.2 can be used as such an example. The data are restricted here to the years 1991-1996, primarily to reduce any effect of the overall downward trend in total shutdown time. Atwood et al. (1998) report 24 plant-centered LOSP events during shutdown in 1991-1996. They are given as Example 6.6.

The null hypothesis is that the successive times between events are independent and exponentially distributed. We consider the alternative hypotheses that

- the times are not exponentially distributed, possibly with more short times between events than expected from an exponential distribution, or
- successive times are correlated, that is that short times tend to be followed by short times and long times by long times.

Example 6.6 Dates of shutdown LOSP events and days between them.

The consecutive dates of shutdown LOSP events are shown in columns below. After each date is the time since the preceding event, in days. For the first event, the time since the start of the study period is shown. Also, the time is shown from the last event to the end of the study period, a 25th "between-event time."

6.

03/07/91	66	04/02/92	10	09/27/94	129
03/13/91	6	04/06/92	4	11/18/94	52
03/20/91	7	04/28/92	22	02/27/95	101
04/02/91	13	04/08/93	345	10/21/95	236
06/22/91	81	05/19/93	41	01/20/96	91
07/24/91	32	06/22/93	34	05/23/96	124
10/20/91	88	06/26/93	4	—	223
01/29/92	101	10/12/93	108		
03/23/92	54	05/21/94	221		

Section 6.7.2.3 discusses ways to investigate whether data come from a particular distribution. Therefore, the issue of the exponential distribution is deferred to that section. The issue of serial correlation motivates the following procedure. Let y_i be the i th time between events, and let x_i be the $(i-1)$ time between events, $x_i = y_{i-1}$. We look to see if x_i and y_i are correlated.

In the above example, the first few (x, y) pairs are $(66, 6)$, $(6, 7)$, and $(7, 13)$, and the final pair is $(124, 223)$.

6.2.3.4.1 Graphical Method

As just mentioned, the issue of whether the distribution is exponential is deferred to Section 6.7.2.3. Consider here the question of serial correlation. A scatter plot of x versus y will indicate whether the values are correlated. However, with skewed data the large values tend to be visually dominant, distorting the overall message of the plot. One could try an ad hoc transformation, such as the logarithmic transformation, but a more universally applicable approach is to use the **ranks** of the variables. That is, sort the n times in increasing order, and assign rank 1 to the smallest time and rank n to the largest time.

In the example, the two shortest times are each equal to 4 days. Each is assigned the average of ranks 1 and 2, namely 1.5. The next largest time is 6 days, which is assigned rank 3, and so forth. The 17th and 18th times are each 101 days, so those two are each assigned rank 17.5. Selected values of x , y and their ranks are shown in Table 6.9. For compactness, not all of the values are printed.

Table 6.9 Calculations for analyzing LOSP dates.

x	rank(x)	y	rank(y)
—	—	66	13
66	13	6	3
6	3	7	4
7	4	13	6
13	6	81	14
81	14	32	8
32	8	88	15
88	15	101	17.5
101	17.5	54	12
54	12	10	5
...
52	11	101	17.5
101	17.5	236	24
236	24	91	16
91	16	124	20
124	20	223	23
223	23	—	—

Figure 6.21 shows a scatter plot of rank(x) versus rank(y). The plot seems to show very little pattern, indicating little or no correlation from one time to the next. The barely perceptible trend from lower left to upper right (“southwest to northeast”) is probably not meaningful, but a hypothesis test will need to be performed to confirm or refute that judgment.

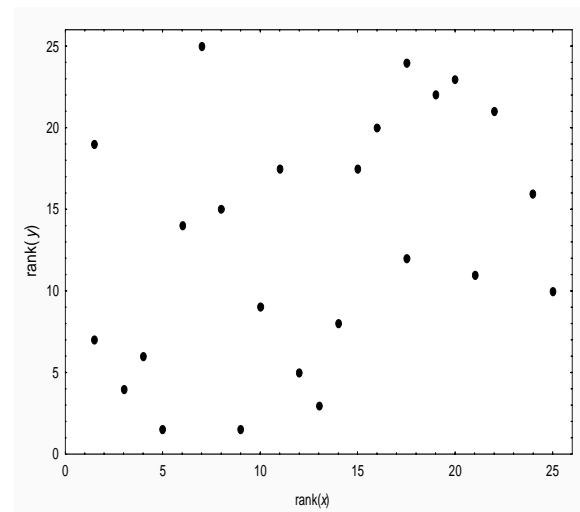


Figure 6.21 Scatter plot of rank(x) versus rank(y).

6.2.3.4.2 Statistical Tests

This section considers whether the between-event times are serially correlated. The question of whether they

are exponentially distributed is discussed in Section 6.7.2.3, under the topic of goodness-of-fit tests.

To test for correlation, it is not appropriate to assume normality of the data. Instead, a nonparametric test should be used, that is, a test that does not assume any particular distributional form. A test statistic that is commonly produced by statistical software is Kendall's tau (τ). Tau is defined in Conover (1999), Hollander and Wolfe (1999), and other books on nonparametric statistics.

Based on the data of Table 6.9, the hypothesis of no correlation between X and Y was tested. Kendall's tau gave a p-value of 0.08. This calculation indicates that the very slight trend seen in Figure 6.21 is not quite statistically significant.

Recall, from the discussion of Section 6.2.3.1.2, that a small p-value is not the end of an analysis. The p-value for this example is small, indicating that the trend in Figure 6.21 is rather unlikely under the assumption of no correlation. If we are concerned about this fact, we must seek possible engineering mechanisms for the trend. The data are times between LOSP events in the industry as a whole. Therefore, the most plausible explanation is the overall industry trend of fewer shutdown LOSP events. This trend would produce a tendency for the short times to occur together (primarily near the start of the data collection period) and the long times to occur together (primarily near the end of the data period).

6.2.3.5 Consistency of Data and Prior

As an example, if the prior distribution has mean $E_{\text{prior}}(\lambda)$, but the observed data show x/t very different from the prior mean, the analyst might wonder if the data and the prior are consistent, or if, instead, the prior distribution was misinformed. To investigate this, one could ask what the prior probability is of getting the observed data. Actually, any individual x may have small probability, so a slightly more complicated question is appropriate.

Suppose first that x/t is in the left tail of the prior distribution. The relevant quantity is the prior probability of observing x or fewer events. This is

$$\Pr(X \leq x) = \int \Pr(X \leq x | \lambda) f_{\text{prior}}(\lambda) d\lambda \quad (6.5)$$

where

$$\Pr(X \leq x | \lambda) = \sum_{k=0}^x e^{-\lambda t} (\lambda t)^k / k! . \quad (6.6)$$

If the prior distribution is gamma($\alpha_{\text{prior}}, \beta_{\text{prior}}$), it can be shown that the probability in question equals

$$\Pr(X \leq x) = \sum_{k=0}^x \frac{\Gamma(\alpha + k)}{k! \Gamma(\alpha)} (t / \beta)^k (1 + t / \beta)^{-(\alpha+k)} \quad (6.7)$$

where $\Gamma(s)$ is the gamma function, a generalization of the factorial function as described in Appendix A.7.6. The distribution defined by Equation 6.7 is named the **gamma-Poisson** or **negative binomial distribution**. The above probability can be evaluated with the aid of software. If the prior distribution is not a gamma distribution, Equation 6.5 does not have a direct analytical expression.

One method of approximating the integral in Equation 6.5 is by Monte Carlo sampling. Generate a large number of values of λ from the prior distribution. For each value of λ , let y be the value of Equation 6.6, which can be calculated directly. The average of the y values is an approximation of the integral in Equation 6.5. Another method of approximating the Equation 6.5 is by numerical integration.

If the probability given by Equation 6.5 is small, the observed data is not consistent with the prior belief — the prior belief mistakenly expected λ to be larger than it apparently is.

Similarly, if x/t is in the right tail of the prior distribution, the relevant quantity is the prior probability that $X \geq x$. When the prior is a gamma distribution, the desired probability is the analogue of Equation 6.7, with the limits of the summation going from x to ∞ . In any case, the desired probability can be approximated by Monte Carlo sampling. If that probability is small, the prior distribution mistakenly expected λ to be smaller than it apparently is.

In Example 6.3, we ask whether the observed zero failures in 2102 reactor-calendar-years is consistent with the WASH-1400 prior, lognormal with median 1E-3 per year and error factor 10. To investigate

6.

this, 100,000 random values of λ were generated from the lognormal prior. (The details are given below.) For each λ , $\Pr(X \leq 0) = \exp(-2102\lambda)$ was found. The mean of these probabilities was 0.245. This is a sample mean, and it estimates the true probability. It is not small, and therefore gives no reason to question the applicability of the prior.

One must ask whether the sample was large enough. The software that calculated the sample mean also calculated the standard error to be 0.0009. Recall from Section 6.2.1.2 that in general a 95% confidence interval can be written as the estimate plus or minus $2 \times$ (standard error). In this case, this interval becomes 0.245 ± 0.002 . We conclude that the true mean equals 0.245 except perhaps for random error in the third digit. This shows that the sample size was more than large enough to give an answer to the accuracy required.

The recipe for generating λ from a lognormal distribution is as follows:
Generate z from a standard normal distribution, using commercial software.
Define $\text{loglam} = \mu + \sigma z$, where μ and σ were found in Section 6.2.2.7.2.
Define $\text{lambda} = \exp(\text{loglam})$.

6.3 Failures to Change State: Failure on Demand

This section has a similar flavor to Section 6.2, but the details are different. It applies to data satisfying the assumptions of Section 2.3.2.1. The probability of a failure on demand is denoted p , a unitless quantity. The data consist of x failures in n demands, with $0 \leq x \leq n$. Before the data are generated, the number of failures is random, denoted X . For any particular number x , the probability of x failures in n demands is

$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad (6.8)$$

where the binomial coefficient is defined as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

The methods will be illustrated by the following hypothetical data set.

Example 6.7 AFW turbine-train failure to start.

In the last 8 demands of the turbine train of the auxiliary feedwater (AFW) system at a PWR, the train failed to start 1 time. Let p denote the probability of failure to start for this train.

As in Section 6.2, frequentist methods are presented first, followed by Bayesian methods. This choice is made because the frequentist point estimate is so very simple, not because frequentist estimation is preferable to Bayesian estimation. Indeed, in PRA p is normally estimated in a Bayesian way.

6.3.1 Frequentist, or Classical, Estimation

6.3.1.1 Point Estimate

The most commonly used frequentist estimate is the **maximum likelihood estimate** (MLE). It is found by taking the **likelihood**, given by Equation 6.8, and treating it as a function of p . The value of p that maximizes the likelihood is called the MLE. It can be shown, by setting a derivative to zero, that the maximum likelihood estimate (MLE) of p is $\hat{p} = x/n$. This is intuitively appealing, the observed number of failures divided by the observed number of demands.

Figure 6.22 shows the likelihood as a function of p , for the data of Example 6.7. The figure shows that the likelihood is maximized at $p = 1/8$, just as stated by the formula.

If several subsets of data, such as data corresponding to several plants, several types of demand, or several years, are assumed to have the same p , data from the various sources may be combined, or **pooled**, for an overall estimate. Denoting the number of failures and demands in data subset j by x_j and n_j , respectively, let $x = \sum_j x_j$ and $n = \sum_j n_j$. The MLE is x/n .

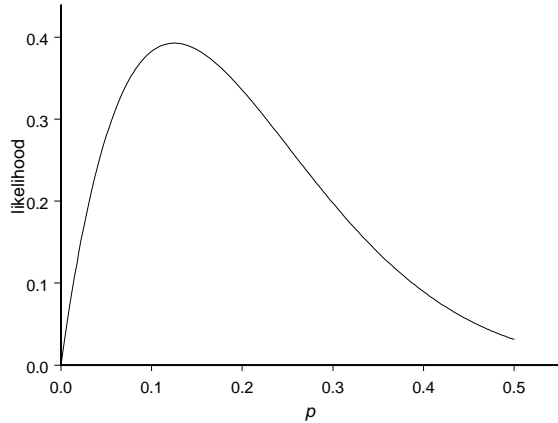


Figure 6.22 Likelihood as a function of p , for the data of Example 6.7.

As mentioned in Sec. 6.2.1.1, final answers will be shown in this handbook with few significant digits, to avoid giving the impression that the final answer reflects precise knowledge of the parameter. Intermediate values will show more significant digits, to prevent roundoff errors from accumulating.

6.3.1.2 Standard Deviation of Estimator

The number of failures is random. One number was observed, but if the demands were repeated a different number of failures might be observed. Therefore, the estimator is random, and the calculated estimate is the value it happened to take this time. Considering the data as random, one could write $\hat{P} = X / n$. This notation is consistent with the use of upper case letters for random variables, although it is customary in the literature to write \hat{p} for both the random variable and the calculated value. The standard deviation of the estimator is $[p(1-p)/n]^{1/2}$. Substitution of the estimate \hat{p} for p yields an estimate of the standard deviation,

$$[\hat{p}(1 - \hat{p}) / n]^{1/2} .$$

The estimated standard deviation of an estimator is also called the **standard error** of the estimate. The handy rule given in Section 6.2.1.2 applies here as well:

$$\text{MLE} \pm 2 \times (\text{standard error})$$

is an approximate 95% confidence interval for p , when the number of demands, n , is large. However, an exact confidence interval is given below.

In Example 6.7, the standard error for p is

$$[0.125 \times (1 - 0.125) / 8]^{1/2} = 0.12.$$

6.3.1.3 Confidence Interval for p

Readers who are only interested in Bayesian estimation may wish to skip this section on the first reading.

The interpretation of confidence intervals is given in Appendix B and in Section 6.2.1.3. It is so important that it is repeated once more here. In the frequentist approach, p is fixed and the data are random. Therefore the maximum likelihood estimator and the confidence limits are all random. For most data sets the MLE, \hat{p} , will be close to the true value of p , and the confidence interval will contain p . Sometimes, however, the MLE will be rather far from p , and sometimes (less than 10% of the time) the 90% confidence interval will not contain p . The procedure is good in the sense that most of the time it gives good answers, but the analyst never knows if the current data set is one of the unlucky ones. A figure like Figure 6.3 could be constructed for p , to illustrate that many data sets could be generated from the same p , yielding many confidence intervals, *most of which* contain the true value of p .

The following material is drawn from Johnson et al. (1992, Section 3.8.3). A confidence interval for p is most naturally expressed in terms of quantiles of a beta distribution. Appendix A.7.8 presents the basic facts about the beta distribution. As mentioned there, the beta family of distributions includes many distributions that are defined on the range from 0 to 1, including the uniform distribution, bell-shaped distributions, and U-shaped distributions. The beta distribution is also discussed more fully in the section below on Bayesian estimation.

Denote the lower and upper ends of a $100(1 - \alpha)\%$ confidence interval by $p_{\text{conf}, \alpha/2}$ and $p_{\text{conf}, 1-\alpha/2}$, respectively. It can be shown that the lower limit is

$$p_{\text{conf}, \alpha/2} = \text{beta}_{\alpha/2}(x, n - x + 1)$$

6.

and the upper limit is

$$p_{\text{conf}, 1-\alpha/2} = \text{beta}_{1-\alpha/2}(x+1, n-x)$$

where $\text{beta}_q(\alpha, \beta)$ denotes the q quantile, or $100 \times q$ percentile, of the $\text{beta}(\alpha, \beta)$ distribution. For example, a 90% confidence interval for p is given by $\text{beta}_{0.05}(x, n-x+1)$ and $\text{beta}_{0.95}(x+1, n-x)$. If $x=0$, the beta distribution for the lower limit is not defined; in that case, set $p_{\text{conf}, \alpha/2} = 0$. Similarly, if $x=n$, the beta distribution for the upper limit is not defined; in that case, set $p_{\text{conf}, 1-\alpha/2} = 1$. In any case, note carefully that the parameters of the beta distribution are not quite the same for the lower and upper endpoints.

Appendix C tabulates selected percentiles of the beta distribution. However, interpolation may be required. Some software packages, including commonly used spreadsheets such as Microsoft Excel (2001) and Quattro Pro (2001), calculate the percentiles of the beta distribution. Finally, Appendix A.7.8 gives a last-resort method, which allows beta percentiles to be calculated by rather complicated formulas involving tabulated percentiles of the F distribution.

In the Example 6.7, with 1 AFW train failure in 8 demands, suppose that a 90% interval is to be found. Then $\alpha = 0.10$, and $1-\alpha/2 = 0.95$. For the lower limit, $\text{beta}_{0.05}(1, 8-1+1) = 6.39\text{E}-3$, from Table C.5. Thus $p_{\text{conf}, 0.05} = 0.0064$.

For the upper limit, $\text{beta}_{0.95}(1+1, 8-1) = 4.71\text{E}-1$, also from Table C.5. Thus $p_{\text{conf}, 0.95} = 0.47$.

6.3.2 Bayesian Estimation

Just as for λ in Sec. 6.2.2, Bayesian estimation of p involves several steps. The prior belief about p is quantified by a probability distribution, the **prior distribution**. This distribution will be restricted to the range $[0,1]$, because p must lie between 0 and 1, and it will assign the most probability to the values of p that are deemed most plausible. The data are then collected, and the **likelihood function** is constructed. This is given by Equation 6.8 for failures on demand. It is the probability of the observed data, written as a function of p . Finally, the **posterior distribution** is constructed, by combining the prior distribution and the likelihood function through Bayes' theorem. The

posterior distribution shows the updated belief about the values of p . It is a modification of the prior belief that accounts for the observed data.

Figure 6.4, showing the effect of various data sets on the posterior distribution, is worth studying. Although that figure refers to λ , an analogous figure applies to p .

As mentioned for λ , lower case p will be used to denote the uncertain parameter with its associated probability distribution and also individual values.

The subsections below consider estimation of p using various possible prior distributions. The simplest prior distribution is discrete. The posterior can be calculated easily, for example by a spreadsheet. The next simplest prior is called **conjugate**; this prior combines neatly with the likelihood to give a posterior that can be evaluated by simple formulas. Finally, the most general priors are considered; the posterior distribution in such a case can only be found by numerical integration or by random sampling.

Section 6.2.2.2 discusses how to choose a prior, and gives references for further reading. It applies to estimation of p as much as to estimating of λ and should be read in connection with the material given below.

6.3.2.1 Estimation with a Discrete Prior

This illustration uses a discrete prior in Bayes estimation for Example 6.7. Two things are different from the example presented in 6.2.2.3. First Example 6.7 deals with evidence in the form of F failures in D demands (rather than in time T). Second, this time we apply an informed prior. Of course Bayes theorem remains:

$$f(\lambda_i | E) = \frac{f(\lambda_i)L(E|\lambda_i)}{\sum_{i=1}^N L(E|\lambda_i)f(\lambda_i)}$$

where

$f(\lambda_i | E)$ = probability density function of λ_i given evidence E (posterior distribution)

$f(\lambda_i)$ = the probability density prior to having evidence E (prior distribution)

$L(E|\lambda_i)$ = the likelihood function (probability of the evidence given λ_i)

Again the denominator, the total probability of the evidence E, is simply a normalizing constant.

Now, when the evidence is in the form of F failures in D demands, the likelihood function is the Poisson distribution:

$$L(E|\lambda_i) = \frac{e^{-\lambda_i T} (\lambda_i T)^F}{F!}$$

For this example, assume that a prior distribution was developed by plant equipment experts based on population variability data from similar systems, but adapted to account for untested new design aspects of this system. The prior has a most likely value of 0.1, falls linearly to 0.3, then tails off to 0 at 0.8. On the low end it tails off toward 0.001. The results are shown in Figure 6.23.

Note that the posterior follows the shape of the prior very closely. This is because the data are consistent with the peak area of the prior, but are not yet strong enough to appreciably reduce the uncertainty in the prior. What happens to this posterior as additional data accumulate? Figure 6.24 shows that the results with 10 distributions, for data in Example 6.7, 10 times the times as much data confirms the initial results. The data as in Figure 23.

Table 6.10 compares the results of the Bayesian analyses with the original data and with ten times as much data.

Table 6.10 Comparison of results for Example 6.7.

Estimate	5 th %tile	MLE	95 th %tile
Bayes, original data	0.05	0.10	0.28
Bayes, ten times more confirmatory data	0.07	0.10	0.16

6.3.2.2 Estimation with a Conjugate Prior

6.3.2.2.1 Definitions

By far the most convenient form for the prior distribution of p is a beta($\alpha_{\text{prior}}, \beta_{\text{prior}}$) distribution. The beta

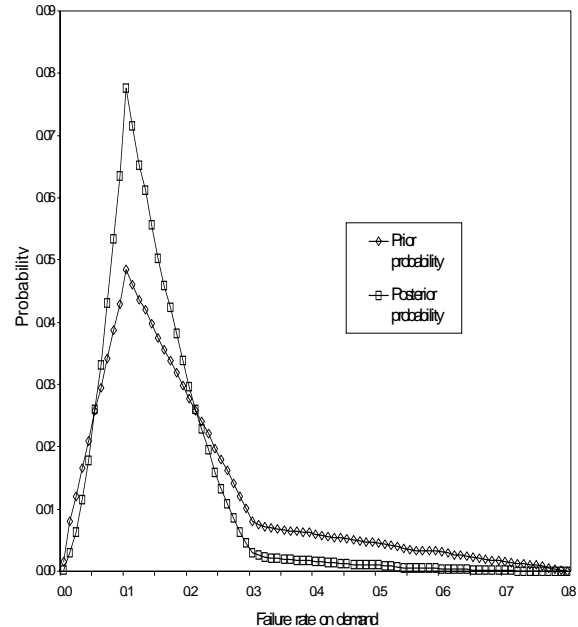


Figure 6.23 Discrete prior and posterior distributions for the data in Example 6.73.

distributions are the conjugate family for binomial data.

6.

The properties of the beta distribution are therefore summarized here, as well as in Appendix A.7.8.

If p has a beta(α, β) distribution, the density is

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} .$$

For most applications the gamma functions in the front can be ignored — they only form a normalizing constant, to ensure that the density integrates to 1. The important feature of the density is that

$$f(p) \propto p^{\alpha-1} (1-p)^{\beta-1} \quad (6.9)$$

where, as always, the symbol \propto denotes "is proportional to." The parameters of the distribution, α and β , must both be positive. The mean and variance of the distribution are

$$\mu = \alpha/(\alpha+\beta) \quad (6.10)$$

$$\begin{aligned} \text{variance} &= \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} \\ &= \mu(1-\mu)/(\alpha+\beta+1). \end{aligned} \quad (6.11)$$

The shape of the beta density depends on the size of the two parameters. If $\alpha < 1$, the exponent of p is negative in Equation 6.9, and therefore the density is unbounded as $p \rightarrow 0$. Likewise, if $\beta < 1$, the density is unbounded as $p \rightarrow 1$. If both $\alpha > 1$ and $\beta > 1$, the density is roughly bell shaped, with a single mode. Appendix A.7.8 shows graphs of some beta densities. Equation 6.11 shows that as the sum $\alpha + \beta$ becomes large, the variance becomes small, and the distribution becomes more tightly concentrated around the mean.

As will be seen below, if the prior distribution is a beta distribution, so is the posterior distribution. Therefore, the above statements apply to both the prior and the posterior distributions.

Appendix C tabulates selected percentiles of beta distributions. Also, the percentiles of a beta distribution can be found by many software packages, including some spreadsheets. Also, the percentiles can be obtained from algebraic formulas involving percentiles of the F distribution, as explained in Appendix A.7.8.

6.3.2.2 Update Formulas

The beta family is **conjugate** to binomial data. That is, updating a beta prior distribution with the data produces a posterior distribution that is also a beta distribution. This follows immediately from the derivation of the posterior distribution. As stated in Appendix B, the posterior distribution is related to the prior distribution by

$$f_{\text{post}}(p) \propto \Pr(X = x|p) f_{\text{prior}}(p) . \quad (6.12)$$

This is the analogue of Equation 6.4, replacing λ by p . As mentioned in Sec. 6.3.1.1, the probability of the data is also called the "likelihood." It is given by Equation 6.8. Stripped of all the normalizing constants, the beta p.d.f. is given by Equation 6.9.

Therefore, the beta distribution and the binomial likelihood combine as:

$$\begin{aligned} f_{\text{post}}(p) &\propto p^x (1-p)^{n-x} p^{\alpha-1} (1-p)^{\beta-1} \\ &\propto p^{x+\alpha-1} (1-p)^{n-x+\beta-1} . \end{aligned}$$

In the final expression, everything that does not involve p has been absorbed into the proportionality constant. This shows that the posterior distribution is of the form beta($\alpha_{\text{post}}, \beta_{\text{post}}$), with

$$\begin{aligned} \alpha_{\text{post}} &= \alpha_{\text{prior}} + x \\ \beta_{\text{post}} &= \beta_{\text{prior}} + (n - x) . \end{aligned}$$

The mean and variance of the prior and posterior distributions are given by Equations 6.10 and 6.11, using either the prior or posterior α and β .

These update formulas give intuitive meaning to the beta parameters: α_{prior} corresponds to a prior number of failures and β_{prior} to a prior number of successes. Assuming a beta($\alpha_{\text{prior}}, \beta_{\text{prior}}$) distribution is equivalent to having observed α_{prior} failures and β_{prior} successes before the current data were observed.

6.3.2.3 Possible Conjugate Priors

A concentrated distribution (small variance, large value of $\alpha_{\text{prior}} + \beta_{\text{prior}}$) represents much presumed prior know-

ledge. A diffuse prior (large variance, small value of $\alpha_{\text{prior}} + \beta_{\text{prior}}$) represents very little prior knowledge of p .

6.3.2.3.1 Informative Prior

The warning given in Section 6.2.2.5.1 applies here as well: the prior distribution must be based on information other than the data. If possible, relevant information from the industry should be used.

The calculations are now illustrated with Example 6.7, 1 failure to start in 8 demands of the AFW turbine train. Poloski et al. (1998) examined 9 years of data from many plants, and found a beta(4.2, 153.1) distribution for the probability of the AFW train failure to start.

Application of the update formulas yields

$$\alpha_{\text{post}} = \alpha_{\text{prior}} + x = 4.2 + 1 = 5.2$$

$$\beta_{\text{post}} = \beta_{\text{prior}} + (n - x) = 153.1 + (8 - 1) = 160.1$$

The mean of this distribution is

$$5.2 / (5.2 + 160.1) = 0.031,$$

the variance is

$$0.031 \times (1 - 0.031) / (5.2 + 160.1 + 1) = 1.89 \times 10^{-4},$$

and the standard deviation is the square root of the variance, 0.014. The 5th and 95th percentiles of the posterior beta(α , β) distribution are found from Table C.5, except the tabulated β values do not go above 100. A footnote to that table gives an approximation that is valid for $\beta \gg \alpha$. That formula applies, because $160.1 \gg 5.2$. According to the formula the q quantile is approximated by

$$\chi^2_q(2 \times 5.2) / [2 \times 160.1 + \chi^2_q(2 \times 5.2)].$$

Therefore the 5th percentile of the beta distribution is approximately

$$\chi^2_{0.05}(10.4) / [320.2 + \chi^2_{0.05}(10.4)] = 4.19 / [320.2 + 4.19] = 0.013$$

and the 95th percentile is approximately

$$\chi^2_{0.95}(10.4) / [320.2 + \chi^2_{0.95}(10.4)] = 18.86 / [320.2 + 18.86] = 0.056$$

All these quantities are unitless.

The prior density, posterior density, and posterior c.d.f. of p are shown in Figures 6.25 through 6.27.

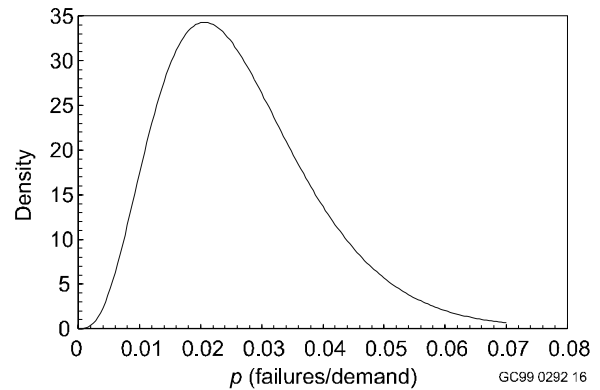


Figure 6.25 Prior density for p , beta(4.2, 153.1).

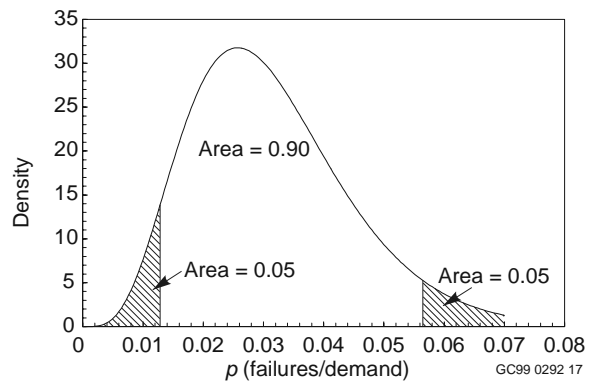


Figure 6.26 Posterior density for p , beta(5.2, 160.1). The 5th and 95th percentiles are shown.

The posterior density is slightly to the right of the prior density. It is to the right because the data, 1 failure in 8 demands, show worse performance than the industry history. The posterior density is only slightly different from the prior density because the data set is small compared to the industry experience (8 demands in the data and an effective 157.3 demands for the industry).

6.

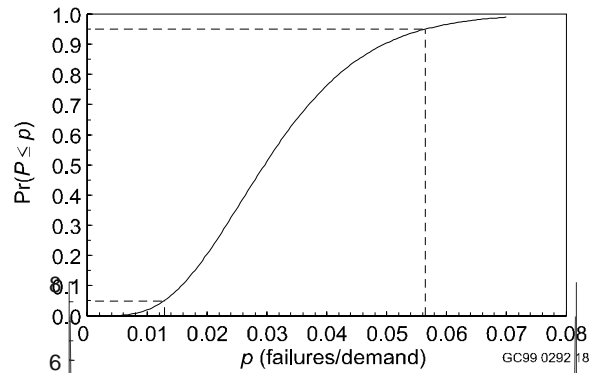


Figure 6.27 Posterior cumulative distribution of p . The 5th and 95th percentiles are shown.

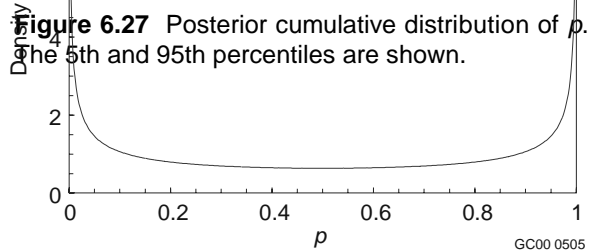


Figure 6.28 Jeffreys noninformative prior distribution for p .

The 5th and 95th percentiles are shown for the posterior distribution, both in the plot of the density and in the plot of the cumulative distribution.

6.3.2.3.2 Noninformative Prior

The Jeffreys noninformative prior is $\text{beta}(1/2, 1/2)$; see Box and Tiao (1973), Sections 1.3.4-1.3.5. This density is shown in Figure 6.28. It is not the uniform distribution, which is a $\text{beta}(1, 1)$ distribution, but instead rises sharply at the two ends of the interval (0, 1). Although the uniform distribution is sometimes used to model no prior information, there are theoretical reasons for preferring the Jeffreys noninformative prior. These reasons are given by Box and Tiao, and are suggested by the comparison with confidence intervals presented below. The uniform distribution would correspond intuitively to having seen 1 failure in 2 demands, which turns out to be too informative. The Jeffreys noninformative prior corresponds to having seen $1/2$ a failure in 1 demand.

The Bayes posterior distribution for p , based on the Jeffreys noninformative prior, is $\text{beta}(x + 1/2, n - x +$

$1/2)$. The mean of the distribution is $(x + 1/2)/(n + 1)$. Selected percentiles are tabulated in Appendix C.

The posterior distribution given here is very similar to the distributions used in the formulas for confidence intervals in Section 6.3.1.3. The only difference is in the parameters. The parameters here are averages of the parameters used in the confidence intervals. For example, the first parameter for the lower confidence limit is x , and the first parameter for the upper confidence limit is $x+1$. The Bayesian limits, on the other hand, use the same parameters for the entire posterior distribution, and the first parameter is $x + 1/2$, the average of the corresponding values for the confidence limits.

In Example 6.7, failure to start of the turbine-driven AFW train, the posterior distribution is $\text{beta}(1.5, 7.5)$. The posterior mean is $1.5/(1.5 + 7.5) = 0.17$. The posterior 90% interval is (0.023, 0.40). As is always the case with discrete data, the confidence interval is conservative, and so is wider than the Jeffreys credible interval. However, the two intervals are similar to each other, being neither to the right nor to the left of the other. Tabular and graphical comparisons are given later.

6.3.2.3.3 Constrained Noninformative Prior

This prior distribution is a compromise between an informative prior and the Jeffreys noninformative prior. As was the case in Section 6.2.2.5.3, the prior mean, denoted here as p_0 , is based on prior belief, but the dispersion is defined to correspond to little information. The priors are described by Atwood (1996) and by references given there.

For binomial data, the constrained noninformative prior distribution is not as neat as for Poisson data. The exact constrained noninformative prior has the form

$$f_{\text{prior}}(p) \propto e^{bp} p^{-1/2} (1 - p)^{-1/2}, \quad (6.13)$$

where b is a number whose value depends on the assumed value of the mean, p_0 . The parameter b is positive when $p_0 > 0.5$ and is negative when $p_0 < 0.5$. Thus, in typical PRA analysis b is negative. Atwood (1996) gives a table of values, a portion of which is reproduced in Appendix C as Table C.8. The table gives the parameter b of the distribution for selected

values of p_0 . In addition, it gives a beta distribution that has the same mean and variance as the constrained noninformative prior.

The beta approximation is illustrated here, and the exact constrained noninformative distribution is treated more fully in the section below on nonconjugate priors.

Return again to Example 6.7, the AFW turbine train failure to start. Let us use the mean of the industry prior found above, $4.2/157.3 = 0.0267$. However, suppose that the full information for the industry prior is not available, or that the system under consideration is considered atypical so that the industry prior is not fully relevant. Therefore, the beta-approximation of the constrained noninformative prior will be used.

Interpolation of Table C.8 at $p_0 = 0.0267$ yields $\alpha = 0.4585$. Solving $\beta = \alpha(1 - p_0)/p_0$ gives $\beta = 16.7138$. The resulting posterior distribution has parameters 1.4585 and 23.7138. Interpolation of Table C.5 gives a 90% interval of (0.0068, 0.15).

6.3.2.3.4 Example Comparison of Above Methods

Just as in Section 6.2, the following general statements can be made.

- The Jeffreys noninformative prior results in a posterior credible interval that is numerically similar to a confidence interval.
- If the prior mean exists, the posterior mean is between the prior mean and the MLE.
- If two prior distributions have about the same mean, the more concentrated (less diffuse) prior distribution will yield the more concentrated posterior distribution, and will pull the posterior mean closer to the prior mean.

Table 6.11 and Figure 6.29 summarize the results of analyzing the AFW-failure-to-start data in the four ways given above.

6.

Table 6.11 Comparison of estimates with 1 failure in 8 demands.

Method	Prior mean	Posterior parameters	Point estimate (MLE or posterior mean)	90% interval (confidence interval or posterior credible interval)
Frequentist	NA	NA	0.125	(0.0064, 0.47)
Bayes with Jeffreys noninformative prior, beta(0.5, 0.5)	0.5	$\alpha = 1.5$ $\beta = 7.5$	0.17	(0.022, 0.40)
Bayes with industry prior, beta(4.2, 153.1)	0.027	$\alpha = 5.2$ $\beta = 160.1$	0.031	(0.013, 0.056)
Bayes with approx. constrained noninform. prior, beta(0.4585, 16.7138)	0.027	$\alpha = 1.4585$ $\beta = 23.7138$	0.058	(0.0068, 0.15)

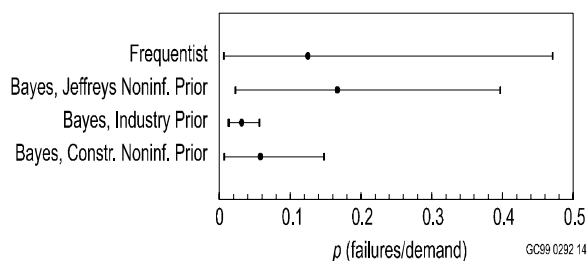


Figure 6.29 Comparison of four point estimates and interval estimates for p .

As in Section 6.2.2.5.4, the Jeffreys prior and the frequentist approach are listed next to each other because they give numerically similar results. The Jeffreys prior yields a posterior credible interval that is strictly contained in the confidence interval, neither to the right nor to the left.

In each Bayesian case, the posterior mean falls between the prior mean and the MLE, 0.125. The prior distribution has more influence when the prior distribution is more tightly concentrated around the mean. One measure of the concentration (at least when the means are similar) is the sum $\alpha_{\text{prior}} + \beta_{\text{prior}}$, because it corresponds to the total number of prior demands, and it is in the denominator of the variance in Equation 6.11. In the present example, when the prior distributions in Table 6.11 are ordered by increasing values of $\alpha_{\text{prior}} + \beta_{\text{prior}}$, the order is the noninformative prior, then the approximate constrained noninformative prior, and finally the industry

prior. The three 90% intervals for the corresponding posterior distributions have decreasing length in the same order.

6.3.2.4 Estimation with a Continuous Nonconjugate Prior

Just as for λ , continuous nonconjugate priors for p cannot be updated with simple algebra. Instead, the posterior distribution must be characterized by numerical integration or by random sampling. Three methods are mentioned here, and the analyst may choose whatever seems easiest.

6.3.2.4.1 Direct Numerical Integration

To use numerical integration, use Equation 6.12 and write the posterior distribution as the product of the likelihood and the prior distribution:

$$Cf_{\text{post}}(p) = p^x(1 - p)^n - x f_{\text{prior}}(p) \quad (6.14)$$

Here C is a constant of proportionality. All the normalizing constants in f_{prior} and in the likelihood may be absorbed into C , leaving only the parts that depend on p on the right-hand side of the equation. Integrate $Cf_{\text{post}}(p)$ from 0 to 1. This integral equals C , because the integral of f_{post} must equal 1. Divide both sides of Equation 6.14 by the just-found constant C , to obtain the function f_{post} . Use numerical integration to find the moments and percentiles of this distribution. Some

suggested methods of numerical integration are mentioned in Section 6.2.2.6.

6.3.2.4.2 Simple Random Sampling

To use random sampling, follow the rejection algorithm given in Section 6.2.2.6. The general algorithm, given in Section 6.2.2.6, can be restated for binomial data as follows. Define

$$m = (x/n)^x(1 - x/n)^{n-x}$$

if $0 < x < n$. If $x = 0$ or $x = n$, define $m = 1$. The steps of the algorithm are:

- (1) Generate a random p from the prior distribution.
- (2) Generate u from a uniform distribution, $0 \leq u \leq 1$.
- (3) If $u \leq p^x(1 - p)^{n-x}/m$, accept p in the sample. Otherwise discard p .

Repeat Steps (1) through (3) until a sample of the desired size is found.

6.3.2.4.3 More Complicated Random Sampling

All-purpose Bayesian update programs can be used for the present simple problem, just as in Section 6.2. The powerful program BUGS is mentioned in Section 6.2.2.6.3, and described more fully in Sections 7.2.3 and 8.2.3.3.3. It can be used here, although it is intended for much more complicated problems.

6.3.2.5 Examples with Nonconjugate Priors

Several possible nonconjugate prior distributions are discussed here.

6.3.2.5.1 Lognormal Distribution

The lognormal distribution is by far the most commonly used nonconjugate distribution. The parameter p has a lognormal distribution if $\ln(p)$ is normally distributed with some mean μ and variance σ^2 .

Facts about the lognormal distribution are given in Appendix A.7.3. One important fact is that the range of the lognormal distribution is from 0 to ∞ . Thus, the

distribution of p cannot be exactly lognormal, because p cannot be greater than 1. When using a lognormal prior, one must immediately calculate the prior $\Pr(p > 1)$. If this probability is very small, the error can be neglected. (When generating values p from the lognormal distribution, either throw away any values greater than 1 or set them equal to 1. In either case, such values hardly ever occur and do not affect the analysis greatly.) On the other hand, if the prior $\Pr(p > 1)$ is too large to be negligible, then the lognormal distribution cannot possibly be used. Even if the software accepts the lognormal distribution, and hides the problem by somehow handling the values that are greater than 1, the actual distribution used is not lognormal. It is truncated lognormal, or lognormal with a spike at 1, with a different mean and different percentiles from the initially input lognormal distribution. The analyst's two options are to recognize and account for this, or to use a different prior distribution.

To use the above sampling algorithm with a lognormal prior, p must be generated from a lognormal distribution. The easiest way to do this is first to generate z from a standard normal distribution, that is, a normal distribution with mean = 0 and variance = 1. Many software packages offer this option. Then let $y = \mu + \sigma z$, so that y has been generated from a normal(μ , σ^2) distribution. Finally, let $p = e^y$. It follows that p has been randomly generated from the specified lognormal distribution.

6.3.2.5.2 Logistic-Normal Distribution

This distribution is explained in Appendix A.7.9. The parameter p has a logistic-normal distribution if $\ln[p/(1 - p)]$ is normally distributed with some mean μ and variance σ^2 . The function $\ln [p/(1 - p)]$ is called the **logit** function of p . It is an analogue of the logarithm function for quantities that must lie between 0 and 1. Using this terminology, p has a logistic-normal distribution if $\text{logit}(p)$ is normally distributed.

Properties of the logistic-normal distribution are given in Appendix A.7.9, and summarized here. Let $y = \ln[p/(1 - p)]$. Then $p = e^y / (1 + e^y)$. This is the inverse of the logit function. As p increases from 0 to 1, y increases from $-\infty$ to $+\infty$.

6.

Note, unlike a lognormally distributed p , a logistic-normally distributed p must be between 0 and 1. Therefore the logistic-normal distribution could be used routinely by those who like the lognormal distribution, but do not know what to do when the lognormal distribution assigns p a value that is greater than 1.

The relation between p and $y = \text{logit}(p)$ gives a way to quantify prior belief about p in terms of a logistic-normal distribution. Decide on two values, such as lower and upper plausible bounds on p or a median and plausible upper bound, equate them to percentiles of p , translate those percentiles to the corresponding two percentiles of the normal random variable Y , and solve those two equations for μ and σ .

To generate a random value from a logistic-normal distribution, first generate y from a normal(μ, σ^2) distribution, exactly as in the section above on the lognormal distribution. Then let $p = e^y / (1 + e^y)$. This p has been randomly generated from the specified logistic-normal distribution.

6.3.2.5.3 Exact Constrained Noninformative Distribution

The prior distribution has the form of Equation 6.13, and the posterior distribution is

$$f_{\text{post}}(p) = C_1 e^{bp} p^{x-1/2} (1-p)^{n-x-1/2},$$

where C_1 is a normalizing constant to make the density integrate to 1.0. Except for the normalizing constant, this is e^{bp} times a beta($x+1/2, n-x+1/2$) distribution. Numerical integration is straightforward, and will not be explained here. To generate a sample from the posterior distribution, the rejection method algorithm originally given in Sec. 6.2.2.6 takes the following form.

Write the beta($x+1/2, n-x+1/2$) density as

$$f_{\text{beta}}(p) = C_2 p^{x-1/2} (1-p)^{n-x-1/2}.$$

Typically, the desired mean of p is less than 0.5; if it is not, reverse the roles of p and $1-p$. The algorithm first defines M to be the maximum possible value of the ratio $f_{\text{post}}(p) / f_{\text{beta}}(p)$. Because $b < 0$ in Table C.8, we

have $e^{bp} \leq 1$, making M equal to C_1/C_2 . Therefore, the condition in Step (3) of the algorithm reduces to

$$u \leq e^{bp}.$$

Therefore, the algorithm simplifies to the following.

- (1) Generate a random p from the beta($x+1/2, n-x+1/2$) distribution. Ways to do this are discussed below.
- (2) Generate u from a uniform distribution, $0 \leq u \leq 1$.
- (3) If $u \leq e^{bp}$, accept p in the sample. Otherwise discard p .

Repeat Steps (1) through (3) until a sample of the desired size is found.

Not all standard software packages give the option of generating random numbers from a beta distribution, although many more allow random number generation from a gamma distribution or from a chi squared distribution. When working with such software, let y_1 be randomly generated from a gamma($x+1/2, 1$) distribution and let y_2 be randomly generated from a gamma($n-x+1/2, 1$) distribution. Alternatively, let y_1 be randomly generated from a chi-squared($2x+1$) distribution and let y_2 be randomly generated from a chi-squared($2n-2x+1$) distribution. In either case, define $p = y_1 / (y_1 + y_2)$. Then p has been generated from the specified beta($x+1/2, n-x+1/2$) distribution. (See Chapter 25 of Johnson et al. 1995.)

6.3.2.5.4 Maximum Entropy Prior

The maximum entropy prior and the constrained noninformative prior were developed with the same goal, to produce a diffuse distribution with a specified plausible mean. The diffuseness of the maximum entropy distribution is obtained by maximizing the entropy, defined as

$$-E[\ln f(p)] = -\int [\ln f(p)] f(p) dp.$$

When p is restricted to the range from 0 to 1, it can be shown that the density f maximizing the entropy is uniform,

$$f(p) = 1 \quad \text{for } 0 \leq p \leq 1$$

and $f(p) = 0$ elsewhere. More interesting is the case when the mean of the distribution is required to equal some prespecified value p_0 . In this case the maximum entropy distribution has the form of a truncated exponential distribution,

$$f(p) = Ce^{bp} \quad \text{for } 0 \leq p \leq 1$$

and $f(p) = 0$ elsewhere. In this form, b is negative when $p_0 < 0.5$ and b is positive when $p_0 > 0.5$. The value of b corresponding to a particular mean must be found by numerical iteration. Some authors write e^{-bp} instead of e^{bp} ; this simply reverses the sign of the parameter b .

The maximum entropy distribution and the uniform distribution are related — if the constraint on the mean is removed, the maximum entropy distribution equals the uniform distribution. In this sense, the maximum entropy distribution is a generalization of the uniform distribution. The constrained noninformative distribution is the same sort of generalization of the Jeffreys noninformative distribution — if the constraint is removed, the constrained noninformative prior becomes the Jeffreys noninformative prior. Atwood (1996) reviews the reasons why the Jeffreys prior is superior to the uniform prior, and uses the same reasoning to argue that the constrained noninformative prior is superior to the maximum entropy prior.

In practice, however, it may make little difference which distribution is used. Both distributions are intended to be used when little prior knowledge is available, and quantifying "little prior knowledge" is not something that can be done precisely.

Sampling from the posterior distribution is similar to the other sampling procedures given above, so most of the details are not given. The only point deserving discussion is how to generate a random sample from the maximum entropy prior. The most convenient method is the **inverse c.d.f. algorithm**. This algorithm is simple in cases when the c.d.f. and its inverse can be calculated easily.

The idea is this. Let the random variable P have c.d.f. F . Let F^{-1} be the inverse function, defined by $u = F(p)$

if and only if $p = F^{-1}(u)$. Let U be defined as $F(P)$. What is the distribution of U ? The c.d.f. of U is found by a little mathematical trickery,

$$\begin{aligned} \Pr(U \leq u) &= \Pr[F(P) \leq u] \\ &= \Pr[P \leq F^{-1}(u)] \\ &= F[F^{-1}(u)] \quad \text{because } F \text{ is the c.d.f. of } P \\ &= u . \end{aligned}$$

Therefore, U has a uniform distribution. The letter U was not chosen by accident, but in anticipation of the uniform distribution.

To generate a random value p from the distribution F , generate a random u from the uniform(0, 1) distribution, something that many software packages allow. Then define $p = F^{-1}(u)$. This is the inverse c.d.f. method of random number generation.

To apply this to the maximum entropy distribution, first integrate the maximum entropy density to yield the c.d.f.

$$F(p) = (1 - e^{bp}) / (1 - e^b) .$$

Generate u from a uniform(0, 1) distribution, and set

$$u = (1 - e^{bp}) / (1 - e^b) .$$

Solve this equation for p ,

$$p = -\ln[1 - (1 - e^b)u] / b .$$

Then p has been randomly generated from the maximum entropy distribution. Repeat this with new values of u until enough values of p have been obtained.

6.3.2.5.5 Example Calculation

These techniques will be illustrated with the Example 6.7, 1 failure to start in 8 demands of the AFW turbine train. Two prior distributions will be assumed, the lognormal prior used by the Accident Sequence Evaluation Program (ASEP), as presented by Drouin et al. (1987), and a logistic-normal distribution having the same 50th and 95th percentiles.

The ASEP distribution for turbine-driven pump failure to start is lognormal with mean $3E-2$ per demand and error factor 10. The three relevant equations from Appendix A.7.3 are

6.

$$\begin{aligned} EF(p) &= \exp(1.645\sigma) \\ \text{mean}(p) &= \exp(\mu + \sigma^2/2) \\ p_q &= \exp(\mu + \sigma z_q) \end{aligned}$$

where the subscript q denotes the q th quantile, and z_q is the q th quantile of the standard normal distribution.

Solving the first equation yields $\sigma = 1.3997$. Substitution of this into the second equation yields $\mu = -4.4862$.

The percentiles are not needed yet, but the third equation gives the median, $p_{0.50} = \exp(\mu) = 0.01126$, and the 95th percentile, $p_{0.95} = \exp(\mu + 1.645\sigma) = 0.1126$. (The relation of these two percentiles can also be derived from the fact that the error factor equals 10.)

The prior $\Pr(p > 1)$ is $6.75E-4$, a very small number. In the calculations of this section, the lognormal distribution is truncated at 1.0. That is, integrals are renormalized to make the integral of the density from 0 to 1 equal to exactly 1.0. If random sampling is performed, any sampled values that are greater than 1 are discarded.

The prior and posterior densities of p are shown in Figure 6.30. The densities were calculated using software for numerical integration.

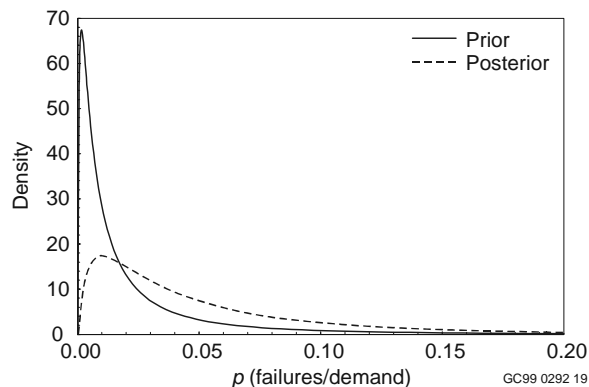


Figure 6.30 Lognormal prior density and posterior density for p .

As a second example, consider the logistic-normal prior distribution having the same 50th and 95th percentiles as the above lognormal prior. These percentiles are 0.01126 and 0.1126. To find the parameters of the underlying normal distribution, set $Y = \ln[p/(1 - p)]$. By the properties of the logistic-

normal distribution given in Appendix A.7.9, the 50th and 95th percentiles of Y are

$$\begin{aligned} y_{0.50} &= \ln[0.01126/(1 - 0.01126)] = -4.475 \\ y_{0.95} &= \ln[0.1126/(1 - 0.1126)] = -2.064 \end{aligned}$$

Because Y has a normal(μ, σ^2) distribution, it follows that

$$\begin{aligned} \mu &= -4.475 \\ \mu + 1.645\sigma &= -2.064 \end{aligned}$$

so $\sigma = 1.466$.

Monte Carlo simulation shows that the truncated-lognormal and logistic-normal prior densities are virtually the same, with means, medians, 5th and 95th percentiles agreeing to two significant digits. As a consequence, the posterior distributions from the two priors are also nearly the same, although the means and percentiles may differ slightly in the second significant digit.

Numerical integration was used, but BUGS could have been used. As an illustration, the script for using BUGS is given in Figure 6.31.

```

model
{
  y ~ dnorm(-4.475, 0.4653)
  p <- exp(y)/(1 + exp(y))
  x ~ dbin(p, 8)
}
list(x = 1)

```

Figure 6.31 Script for analyzing Example 6.7 with BUGS.

This script assigns a logistic-normal prior distribution to p . If a lognormal prior is used instead, BUGS returns an error message during the simulation, presumably because it has generated a value of p greater than 1. The script assigns Y a normal distribution with mean -4.475 . The second parameter is $1/\sigma^2$, because that is how BUGS parameterizes a normal distribution. The entered value, 0.4653, equals $1/1.466^2$. The script then gives X a binomial(8, p) distribution. Finally, the line beginning "list" contains the data, the single observed value 1 in this example. BUGS also wants an initial value for p , but it is willing to generate it randomly.

For the present example, the difference between the lognormal and logistic-normal priors is very small, having no effect on the posterior. The difference between the two priors can be important if the probability of failure is larger and/or the uncertainty is larger. That can be the case with some human errors, with hardware failures in unusually stressful situations, and with recovery from failure if recovery is modeled as an event separate from the original failure. For example, the NUREG 1150 PRA for Surry (Bertucio and Julius 1990) uses the lognormal distribution for most failure probabilities. However, some failure probabilities are large, considerably larger than $3E-2$. In nearly all of those cases, the PRA does not use a lognormal distribution. Instead, the maximum entropy distribution is the PRA's distribution of choice. Other possible distributions, which were not widely known in the PRA community in 1990, would have been the constrained noninformative distribution or a logistic-normal distribution.

6.3.3 Model Validation

All the methods of this section are analogues of methods considered for failure rates, but the details are somewhat different. Some repetition is inevitable, but the examples in this section are chosen to complement the examples of Section 6.2.3, not to duplicate them. For a more complete appreciation of the model validation techniques, both this section and Section 6.2.3 should be read.

The comments at the start of Section 6.2.3 apply equally to this section, and must not be ignored. In particular, an analyst who estimates parameters should check the assumptions of the model.

The first assumption of the binomial model, given in Section 2.3.2, is that the probability of failure is the same on any demand. This assumption will be examined against two possible alternative assumptions: (1) different subsets of the data have different values of p , but in no special order, and (2) a time trend exists. The second assumption of the binomial model is that the outcome on one demand is statistically independent of the outcome on a different demand. This will be examined against the alternatives of common-cause failures and of clustering in time of the failures.

Finally, the consistency of the prior distribution and the data will be considered.

One might also worry about whether n is really constant. If n is not constant, we may treat it as constant by conditioning on n , as explained in Section 2.3.2.4.2.

6.3.3.1 Poolability of Data Sources

The method will be illustrated by data from diesel generator failures to start shown in Example 6.8.

Table C.1 of Grant et al. (1996) gives the data for the first two rows, at plants reporting under Regulatory Guide RG-1.108 during 1987-1993. The failures were those reported in LERs. The number of failures on monthly tests at those plants comes from the unpublished database used for that report, and the number of monthly demands was estimated in a very crude way for use in this example.

Example 6.8 EDG failures to start on demand.

Emergency diesel generator (EDG) failures to start on demand were recorded for three kinds of demands: unplanned demands, the tests performed once per operating cycle (approximately every 18 months), and the monthly tests. The counts are given below.

Type of demand	Failures to start	Number of demands
Unplanned	2	181
Cyclic test	17	1364
Monthly test	56	15000

6.3.3.1.1 Graphical Technique

To explore the relations between subsets of the data, mark the subsets on one axis. For each of these subsets of the data, plot an estimate of p and a confidence interval for p against the other axis. Patterns such as trends, outliers, or large scatter are then made visible.

In Example 6.8, the subsets are types of demand. The data set from each demand type is analyzed separately, and the graph shows an estimate and

6.

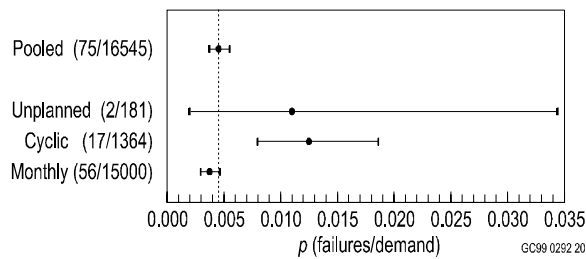


Figure 6.32 MLEs and 90% confidence intervals for p , for three types of demand and for the pooled data.

a confidence interval for each year, plotted side by side. This is shown in Figure 6.32. The plot was produced with a graphics package, although a hand-drawn plot would be adequate to show the results.

The plot shows that the unplanned demands and the cyclic tests appear to have similar values of p , but the monthly tests appear to have a lower value. Several reasons for the difference could be conjectured: the monthly tests may be less stressful, or the failures may not all be reported in LERs, or the estimated number of demands may be badly incorrect.

Figure 6.18, which is the corresponding plot in Section 6.2.3.1.1, has the cells (plants, in that example) arranged in order of decreasing $\hat{\lambda}$. Figure 6.32 does not order the cells by decreasing \hat{p} , because the number of cells is small, only three, and because the cells already have a natural order. The analyst must decide what order makes the most sense and is easiest for the user to interpret. Cleveland (1985, Chap. 3.3) discusses ways of ordering data.

The interval for the pooled data is also shown, not because the data justify pooling, but simply as a reference for comparison. A dotted reference line is drawn through the point estimate based on the pooled data. If only a few data subsets need to be compared, as in Figure 6.32, these embellishments are unnecessary. With many subsets, however, the eye tends to get lost without the reference line. The reference line has the added advantage of focusing the eye on the confidence intervals rather than the point estimates.

The graph is only a picture. Pictures like these are useful, but cannot be used in an easy way to draw naive conclusions about differences between data subsets. The warnings given in Section 6.2.3.1.1 deserve repetition:

- If many confidence intervals are plotted, all based on data with the same p , a few will be far from the others because of randomness alone. An outlying interval does not prove that the p s are unequal.
- This same statement is true if other intervals are used, such as Bayes credible intervals based on the noninformative prior. The issue is the random variability of data, not the kind of interval constructed.
- If there are few intervals, on the other hand, intervals that just barely overlap can give strong evidence for a difference in the p s.

To quantify the strength of the evidence seen in the picture, a formal statistical procedure is given in the next subsection. The picture gives a preview, and helps in the interpretation of the formal statistical quantification. In the present example, if the statistical test finds a statistically significant difference between data subsets, it is natural to then ask what kind of difference exists. The picture shows that p seems to be similar for the unplanned demands and for the cyclic tests, but smaller for the monthly tests. In this way, the picture provides insight even though it does not provide a quantitative statistical test.

6.3.3.1.2 Statistical Tests

Simple Contingency Tables ($2 \times J$). The natural format for the data is a "contingency table." An introductory reference to this subject is Everitt (1992), and many general statistics texts also have a chapter on the topic. In a two-way table, two attributes of the events are used to define rows and columns, and the numbers in the table are counts. In the present example, two attributes of any event are the type of demand and whether it is a failure or success. One way to build a contingency table is to let the first row show system failures and the second row system successes. Then let the columns correspond to the demand types. The table entries are the counts of the events for each cell, shown in Table 6.12 for Example 6.8.

The essence of this table is a 2×3 table, because the basic data counts occupy two rows and three columns.

The row totals, column totals, and grand total are shown in the right and bottom margins. A general two-way contingency table has I rows and J columns. (Although this discussion considers only $2 \times J$ tables, it does no harm to give the general formulas, keeping in mind that the examples of this section have $I = 2$.) The count in the i th row and j th column is denoted n_{ij} , for i any number from 1 to I and j from 1 to J . The total count in row i is denoted n_{i+} and the total count in column j is denoted n_{+j} . The grand total is denoted n_{++} .

Table 6.12 Contingency table for Example 6.8.

	Unplanned	Cyclic	Monthly	Total
Failure	2	17	56	75
Success	179	1347	14944	16470
Total	181	1364	15000	16545

For example, Table 6.12 has $n_{1,3} = 56$ and $n_{2,1} = 179$. It has $n_{2+} = 16470$ and $n_{+2} = 1364$. The grand total, n_{++} , equals 16545 in the example.

Let the null hypothesis be
 H_0 : p is the same for all the data subsets.

The alternative hypothesis is
 H_1 : p is not the same for all the data subsets.

In the example, the data subsets are the three demand types. The analyst must investigate whether H_0 is true. The method used is to see what kind of data would be expected when p really is the same, and then to see how much the observed counts differ from the expected. If the differences are small, the counts are consistent with the hypothesis H_0 . If, instead, the differences are large, the counts show strong evidence against H_0 .

If H_0 is true, that is, if p is really the same for all the demand types, the natural estimate of p is

$$\hat{p} = n_{1+} / n_{++} .$$

Then for column j , one would have expected $n_{+j}\hat{p}$ failures on average. This reasoning leads to the formula for the expected count in cell ij :

$$e_{ij} = n_{i+}n_{+j} / n_{++} .$$

In Table 6.12, for unplanned demands one would have expected $181 \times (75/16545) = 0.82$ failures on average, for cyclic tests $1364 \times (75/16545) = 6.19$ failures, and so forth.

The difference between the observed count and the expected count for any cell is $n_{ij} - e_{ij}$. There are many cells, and therefore many ways of combining the differences to yield an overall number. One useful way is to construct

$$X^2 = \sum_i \sum_j (n_{ij} - e_{ij})^2 / e_{ij} .$$

X^2 is called the chi-squared statistic, or sometimes the Pearson chi-squared statistic. Note, X^2 as defined here is slightly different from the chi-squared statistic for constant event rate in Section 6.2.3.1.2. In that section, the cells had one index, whereas in this section the cells have two indices, and the expected counts are calculated differently. Other than that, the statistics are the same. Table 6.13 expands Table 6.12 to show the quantities needed to calculate X^2 . The observed counts and the expected counts have the same totals, except for roundoff.

Table 6.13 Counts, expected counts, and contributions to X^2 , for Example 6.8.

	Unplanned	Cyclic	Monthly	Total
Failure	2	17	56	75
	0.82	6.19	68.00	
	1.70	18.92	2.12	
Success	179	1347	14944	16470
	180.18	1357.80	14932	
	0.01	0.09	0.01	
Total	181	1364	15000	16545

For example, there were 2 failures on unplanned demands. The expected number of failures on unplanned demands, if H_0 is true, is $181 \times 75/16545 = 0.82$. And the contribution of that cell to X^2 is

6.

$$(2 - 0.82)^2/0.82 = 1.70 .$$

When H_0 is true and the total count is large, the distribution of X^2 has a distribution that is approximately chi-squared with $(I-1) \times (J-1)$ degrees of freedom. In Table 6.12, the number of degrees of freedom is $(2-1) \times (3-1) = 2$. If X^2 is large, compared to the chi-squared distribution, the evidence is strong that H_0 is false; the larger X^2 , the stronger the evidence.

Interpretation of Test Results. Based on any 2×3 contingency table, such as Table 6.12, suppose that X^2 were 6.4. A table of the chi-squared distribution shows that 5.991 is the 95th percentile of the chi-squared distribution with 2 degrees of freedom, and 7.378 is the 97.5th percentile. After comparing X^2 to these values, an analyst would conclude that the evidence is strong against H_0 , but not overwhelming. Quantitatively, the analyst would "reject H_0 at the 5% significance level, but not at the 2.5% significance level." This is sometimes phrased as "the p-value is between 0.05 and 0.025." See the bulleted list in Section 6.2.3.1.2, in the interpretation following Table 6.6, for other phrases that are sometimes used.

If instead X^2 were 1.5, it would lie between the 50th and the 60th percentiles of the chi-squared distribution, and therefore would be in the range of values that would be expected under H_0 . The analyst could say "the observed counts are consistent with the hypothesis H_0 ," or " H_0 cannot be rejected," or "the evidence against H_0 is very weak." The analyst would not conclude that H_0 is true, because it probably is not exactly true to the tenth decimal place, but would conclude that it cannot be rejected by the data.

In fact, in Example 6.8 X^2 equals 22.8, as found by totaling the six contributions in Table 6.13. This number is far beyond the 99.5th percentile of the chi-squared distribution, so the evidence is overwhelming against H_0 . Such an analysis contributed to the decision of Grant et al. not to consider monthly tests in their report.

This example was chosen to illustrate that subsets of the data can correspond not only to different locations or different hardware (for example, different plants or systems), but also to different conditions, in this case different types of demands. In reality, the data analyst should consider various kinds of subsets; in this

example, with data coming from many plants, the analyst should consider possible between-plant differences. The plots and chi-squared tests are exactly the same as given above.

This brings up a difficulty with the present example that has been carefully hidden until now. The hypothesis H_0 is that all the subsets of the data have the same p . A hidden hypothesis, never even proposed for testing, is that within each data subset every demand has the same p . In fact, this turns out not to be the case. Based on only the unplanned demands and cyclic tests, Grant et al. report that the difference between plants is statistically significant — the evidence is strong that p differs from plant to plant. This means that the above analysis must be refined to account for possible differences between plants. Such variation is discussed in Section 8.2 of this handbook.

Thus, the data set has two sources of variation, differences between demand types and also differences between plants. In such a situation, consideration of only one variable at a time can throw off the results if the data set is "unbalanced," for example, if the worst few plants also happen to have the most unplanned demands and the fewest monthly demands. If such between-plant differences are contaminating the EDG data in Example 6.8, the observed difference might not reflect anything about the nature of the demands, but only that the plants with EDG problems were underrepresented on the monthly tests. Example 6.9 shows hypothetical data under such a scenario.

Example 6.9 Hypothetical unbalanced data.

Suppose that the industry consists of "bad" plants and "good" plants. The bad plants have a relatively high probability of failure to start, and also have relatively many unplanned demands. Suppose that the tests perfectly mimic unplanned demands, so that at either kind of plant p is the same on an unplanned demand and on a test. Data from such an industry might be given in the table below. The tables entries show failures/demands.

	Unplanned	Tests
Bad plants	4/20 = 0.2	4/20 = 0.2
Good plants	1/50 = 0.02	8/400 = 0.02
Totals	5/70 = 0.07	12/420 = 0.03

If only the good plants are considered, or if only the bad plants are considered, the data of Example 6.9 show no difference between unplanned demands and tests. The estimated p is the same for unplanned demands and for tests, 0.2 from the bad plants' data and 0.02 from the good plants' data. However, if the data from good plants and bad plants are combined, the unplanned demands appear to have a much higher failure probability than do the tests, 0.07 versus 0.03. This erroneous conclusion is a result of ignoring differences in the data, the existence of two kinds of plants, when the data are unbalanced because the bad plants have a much higher percentage of unplanned demands. Such a situation is known as **Simpson's paradox**.

In fact, this scenario cannot be greatly influencing the data in Example 6.8, because most of the demands are periodic. Therefore, every plant must have approximately the same fraction of monthly tests and of cyclic tests. In conclusion, although between-plant variation must be considered, it is hard to imagine that it affects the outcome in Example 6.8.

As mentioned in Section 6.2.3.1.2, a full data analysis must not stop with the calculation of a p-value. In the present example, with a very large number of demands, it may be that the statistically significant difference is not very important from an engineering viewpoint. In other words, a large data set can detect differences in

the second decimal place, differences that are not worth worrying about in practice.

This concern is addressed in the example by Figure 6.32, which shows that the probability of FTS is about 1/3 as large on monthly tests as on other demands, at least according to the reported data. Therefore, the difference is substantial in engineering terms, and the engineering portion of the data analysis can investigate reasons for the difference.

Required Sample Size. The above approach is valid if the values of n_{ij} are "large." If they are small, X^2 has a discrete distribution, and so cannot have a chi-squared distribution. As a rather extreme example, if n_{++} , the total number of demands, were equal to 4 in the framework of Example 6.8, there would only be a few ways that the four demands (and the number of failures, at least zero and at most four) could be arranged among the three demand types. Therefore X^2 could only take a few possible values. Therefore, the user must ask how large a count is necessary for the chi-squared approximation to be adequate. An overly conservative rule is that all the expected cell counts, e_{ij} , be 5.0 or larger. Despite its conservatism, this rule is still widely used, and cited in the outputs of some current statistics packages. For a $2 \times J$ table, Everitt (1992, Sec. 3.3), citing work by Lewontin and Felsenstein (1965), states that the chi-squared approximation is adequate if all the values of e_{ij} are 1.0 or greater, and that in "the majority of cases" it is sufficient for the e_{ij} values to be 0.5 or greater. For a 2×2 table, however, it is generally best not to use the chi-squared approximation at all, but to use the p-value from "Fisher's exact two-sided test," discussed below.

If the expected cell counts are so small that the chi-squared approximation appears untrustworthy, the analyst has two choices. (a) Pool some columns, thereby combining cells and increasing the expected cell counts. For example, in an investigation of differences between years, with few failures, it might be necessary to combine adjacent years so that the expected number of failures in each time-bin is at least 0.5. (b) Some statistical software packages can compute the "exact distribution" of X^2 in some cases (typically for small tables). Conditional on the n_{i+} values and n_{+j} values, this exact distribution is the finite set of values that X^2 can possibly take, together with their associated probabilities. If the analyst is willing

6.

to base the decision on this conditional distribution, the exact distribution can be used. The commercial package StatXact performs such calculations using modern fast algorithms, even for large tables, subject only to the available memory in the machine. In the special case of a 2x2 contingency table, many software packages compute this p-value, calling it the p-value from "Fisher's exact two-sided test." In general, the p-value from Fisher's exact test is preferable to the p-value from the chi-squared approximation, and should be used whenever the software produces it. This, and other considerations for a 2x2 table, are discussed by Everitt (1992) and Atwood (1994).

In Table 6.13, the smallest expected count is $e_{11} = 0.82$. All the other expected counts are larger than 1.0. This indicates that the sample size is large enough.

6.3.3.2 No Time Trend

This section uses the unplanned HPCI demands from Example 6.5, with the failures indicated. To make a data set with a moderate number of failures, all types of failures are counted together, including failure to start, failure to run, failure of the injection valve to reopen after operating successfully earlier in the mission, and unavailability because of maintenance. For the example, no credit is taken for failures that were recovered. The data are given as Example 6.10.

Example 6.10 Dates of HPCI failures and unplanned demands, 1987-1993.

The HPCI demands of Example 6.5 are listed here with an asterisk marking demands on which some kind of failure occurred. The demands dates are given in columns, in format MM/DD/YY.

01/05/87*	08/03/87*	03/05/89	08/16/90*	08/25/91
01/07/87	08/16/87	03/25/89	08/19/90	09/11/91
01/26/87	08/29/87	08/26/89	09/02/90	12/17/91
02/18/87	01/10/88	09/03/89	09/27/90	02/02/92
02/24/87	04/30/88	11/05/89*	10/12/90	06/25/92
03/11/87*	05/27/88	11/25/89	10/17/90	08/27/92
04/03/87	08/05/88	12/20/89	11/26/90	09/30/92
04/16/87	08/25/88	01/12/90*	01/18/91*	10/15/92
04/22/87	08/26/88	01/28/90	01/25/91	11/18/92
07/23/87	09/04/88*	03/19/90*	02/27/91	04/20/93
07/26/87	11/01/88	03/19/90	04/23/91	07/30/93
07/30/87	11/16/88*	06/20/90	07/18/91*	
08/03/87*	12/17/88	07/27/90	07/31/91	

6.3.3.2.1 Graphical Techniques

Just as elsewhere in this chapter, the time axis can be divided into bins, and the data can be analyzed separately for each bin and compared graphically.

For Example 6.10, defining the bins to be years leads to Table 6.14.

Table 6.14 HPCI failures on demand, by year.

Calendar year	Failures	Demands
1987	4	16
1988	2	10
1989	1	7
1990	3	13
1991	2	9
1992	0	6
1993	0	2

This leads to a plot similar to Figures 6.18 and 6.19, shown in Figure 6.33. The plot with the example data shows no evidence of a trend.

A plot that does not require a choice of how to construct bins is given in Figure 6.34, the analogue of Figure 6.20. It can be constructed when the demands can be ordered sequentially, as is the case for Example 6.10. In this plot, the cumulative number of failures is plotted against the cumulative number of demands. To help the eye judge curvature, a straight line is drawn, connecting the origin with the dot at the upper right.

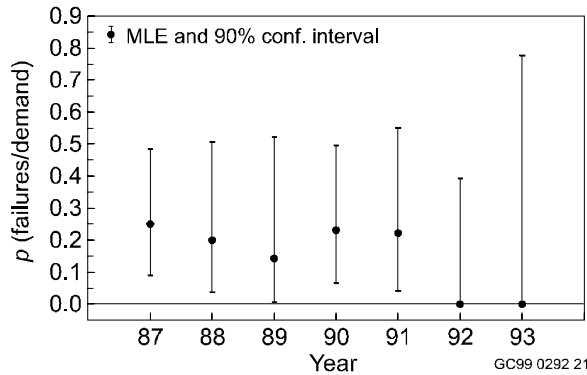


Figure 6.33 Point and interval estimates of p , each based on one year's data.

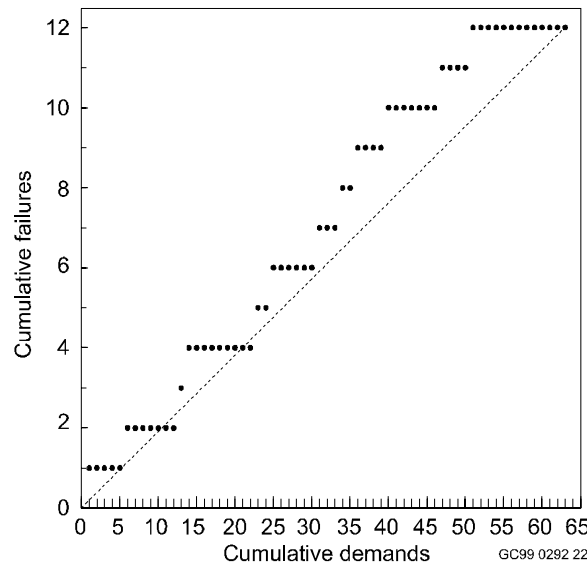


Figure 6.34 Cumulative number of failures versus cumulative number of demands.

The **slope** of any part of the graph is the vertical distance divided by the horizontal distance, $\Delta y/\Delta x$. In the present figure the horizontal distance is the number of demands that have occurred, and the vertical distance is the corresponding number of failures. Therefore,

$$\text{slope} = (\text{number of failures})/(\text{number of demands}),$$

so the slope is a visual estimator of p . A roughly constant slope, that is, a roughly straight line, indicates a constant p . A changing slope indicates changes in p .

In Figure 6.34, the slope is relatively constant, indicating that p does not seem to change with time.

This agrees with Figure 6.33. It is not clear whether the slight departure from the diagonal line in the right half of the figure is more than can be attributed to random variation. Such questions must be addressed by statistical tests, given below.

The details of the diagonal line probably do not matter. The line shown is the maximum likelihood estimate of the expected height of the plot at any horizontal point, assuming constant p . Other lines, slightly different, could also be justified.

6.3.3.2 Statistical Tests for a Trend in p

In this section, the null hypothesis remains

H_0 : p is the same for all the data subsets.

but the alternative is now

H_1 : p is either increasing or decreasing over time .

The Chi-Squared Test. This is the same test as given in Section 6.3.3.1.2, only now the data subsets are years or similar bins of time.

The data of Table 6.14 can be written as a 2×7 contingency table. The smallest expected cell count corresponds to failures in 1993, with the expected count = $2 \times 12/63 = 0.4$. This is too small to justify calculating a p-value from the chi-squared distribution. The problem can be remedied by pooling the two adjacent years with the smallest numbers of demands, 1992 and 1993. (Note, the decision of which subsets to pool is based on the numbers of demands only, not on whether or not those demands resulted in failures. Pooling based on demand counts is legitimate. Pooling based on the failure counts is not.)

When this 2×6 contingency is analyzed by the chi-squared test, the p-value is 0.77, indicating no evidence at all of differences between years. This is no surprise.

The Wilcoxon-Mann-Whitney Test. This test is similar in spirit to the Laplace test for a trend in λ . The null hypothesis is that p is the same for all demands. Suppose that the individual demands are in a known sequence. Against the alternative hypothesis that the failures tend to occur more at one end of the sequence — that is, p is either an increasing or a decreasing

6.

function of the sequence number — use the Wilcoxon-Mann-Whitney test, described in texts that cover nonparametric statistics. Two good sources of standard nonparametric methods are Conover (1999) and Hollander and Wolfe (1999). Hollander and Wolfe call this test the **Wilcoxon rank sum test**.

The test is based on the sum of the ranks of the failures. For example, in the sequence of failures and successes

failure, success, failure, failure, success

the three failures have ranks 1, 3, and 4, and the sum of their ranks is 8. Let W denote the sum of the ranks of x failures in n trials. If x and $n - x$ are both large and if the probability of a failure is the same for the entire sequence, W is approximately normal with mean $\mu_w = x(n+1)/2$ and variance $\sigma_w^2 = x(n-x)(n+1)/12$. If $Z = (W - \mu_w)/\sigma_w$ is in either tail of the distribution, the null hypothesis should be rejected. If x or $n - x$ is small, statistics books give tables, or statistical computer packages calculate the exact tail probability.

The data of Example 6.10 show 12 failures in 63 demands. The first failure was on the first demand (01/05/87), so that failure has rank 1. The next was on the sixth demand, so that failure has rank 6. Two demands occurred on 03/19/90, the 36th and 37th demands. One of the two demands resulted in failure, so that failure was assigned rank 36.5, as is usual in case of ties. The sum of the ranks of the failures is 321.5, and Z can be calculated to equal -1.09 . This is the 13.8th percentile of the normal distribution. Because Z is not in either tail, H_0 is not rejected.

6.3.3.3 Independence of Outcomes

This section is less important than the others. Some readers may wish to skip directly to Section 6.3.3.4.

The second assumption for binomial data is that the outcomes of different demands be independent — a success or failure on one demand does not influence the probability of failure on a subsequent demand.

Outcomes can be dependent in many ways, and some of them must be addressed by careful thinking rather than by statistical data analysis. The analyst or the study team should consider possible common-cause mechanisms, and examine the data to see if many

common-cause failures occurred. If common-cause failures form a noticeable fraction of all the failures, the analyst should probably divide the independent failures and the common-cause failures into separate data sets, and separately estimate the probabilities of each kind of failure.

If demands occur in sequence, it is natural to consider serial dependence, in which the occurrence of a failure on one demand influences the probability of a failure on the next demand. Some people believe that hits in baseball occur this way, that a slump or streak can persist because of a batter's attitude, which is influenced by how successful he has been recently. In the context of hardware failures, suppose that failures are sometimes diagnosed incorrectly, and therefore repaired incorrectly. Immediately after any failure, the probability of failure on the next demand is higher, because the first failure cause may not have been truly corrected. In such a case, the failures would tend to cluster rather than being uniformly scattered among the successes. A cumulative plot such as that in Figure 6.32 can be inspected for such clusters.

If the question of independence is restricted to successive outcomes — outcome $i - 1$ versus outcome i — the data can be analyzed by a 2×2 contingency table. Let y_i be the outcome on demand i , either success or failure. Let x_i be the outcome on demand $i - 1$. The possible values of successive outcomes (x_i, y_i) are (S, S), (S, F), (F, S), and (F, F).

To put this in more familiar language, let p denote the probability of a failure, and consider two kinds of demands, those when the previous outcome (x) was a failure and those when the previous outcome was a success. The null hypothesis is

H_0 : p is the same on both kinds of demands .

Perform the usual chi-squared test of H_0 based on a contingency table.

Example 6.10 results in the following contingency table.

Table 6.15 Contingency table for successive outcomes in Example 6.10

	x = F	x = S	Total
y = F	1	10	11
y = S	11	40	51
Total	12	50	62

Although the chi-squared approximation should be acceptable, it is preferable to use Fisher's exact test for a 2x2 table. The p-value reported by SAS for Fisher's exact test is 0.67. This large p-value shows that the data are very consistent with the hypothesis of independence of successive outcomes. Because the data come from the entire industry, independence is entirely reasonable.

6.3.3.4 Consistency of Data and Prior

If the prior distribution has mean $E_{\text{prior}}(p)$, but the observed data show x/n very different from the prior mean, the analyst must ask if the data and the prior are inconsistent, if the prior distribution was misinformed. The investigation is similar to that in Section 6.2.3.5.

Suppose first that x/n is in the left tail of the prior distribution. The relevant quantity is the prior probability of observing x or fewer events. This is

$$\Pr(X \leq x) = \int \Pr(X \leq x | p) f_{\text{prior}}(p) dp \quad (6.15)$$

where

$$\Pr(X \leq x | p) = \sum_{k=0}^x \binom{n}{k} p^k (1-p)^{n-k} \quad (6.16)$$

If the prior distribution is beta(α , β), it can be shown that Equation 6.15 equals

$$\Pr(X \leq x) = \sum_{k=0}^x \binom{n}{k} \frac{\Gamma(\alpha + k) \Gamma(\beta + n - k)}{\Gamma(\alpha) \Gamma(\beta)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + n)}$$

where $\Gamma(s)$ is the gamma function, a generalization of the factorial function as described in Appendix A.7.6. The name of this distribution is beta-binomial. This probability can be evaluated with the aid of software. If the prior probability is any distribution other than a

beta distribution, Equation 6.15 does not have a direct analytical expression.

Just as in Sec. 6.2.3.5, one method of approximating the integral in Equation 6.15 is by Monte Carlo sampling. Generate a large number of values of p from the prior distribution. For each value of p , let y be the value of Equation 6.16, which can be calculated directly. The average of the y values is an approximation of the integral in Equation 6.15. Another method of approximating the Equation 6.15 is by numerical integration.

If the probability given by Equation 6.15 is small, the observed data are not consistent with the prior belief — the prior belief mistakenly expected p to be larger than it apparently is.

Similarly, if x/n is in the right tail of the prior distribution of the prior distribution, the relevant quantity is the prior $\Pr(X \geq x)$. It is the analogue of Equation 6.15 with the limits of the summation in Equation 6.16 going from x to n . If that probability is small, the prior distribution mistakenly expected p to be smaller than it apparently is.

Again consider Example 6.7, one AFW failure to start in eight demands, and consider the industry prior, beta(4.2, 153.1). One easy approach is Monte Carlo simulation. Therefore, values of p were generated from the beta distribution, using the technique mentioned at the end of Section 6.3.2.5.3. That is, y_1 was generated from a gamma(4.2, 1) distribution, y_2 was generated from a gamma(153.1, 1) distribution, and p was set to $y_1/(y_1 + y_2)$.

The industry-prior mean of p is 0.027. Because the observed number of failures, 1, is larger than the prior expected number, $8 \times 0.027 = 0.21$, we ask whether such a large failure count is consistent with the prior. The probability in question is $\Pr(X \geq 1)$. For each randomly generated p , $\Pr(X \geq 1 | p)$ was found, equal to $1 - \Pr(X = 0 | p) = 1 - (1 - p)^8$. The average of these probabilities, calculated for 100,000 random values of p , was 0.192, with a standard error of 0.0003. This means that the true probability is 0.192, with negligible random error. Because this probability is not small, the data appear consistent with the prior distribution.

6.

6.4 Failure to Change State: Standby Failure

As explained in Sec. 2.3.3, this type of failure is modeled as a failure condition that occurs at an unknown time between the most recent previous inspection, test, or demand and the present one.

Each demand corresponds to a standby time. The only thing that can be observed is whether the system is failed or not at the end of the standby period. From Equation 2.3, the probability that the system is failed at time t is

$$p = 1 - e^{-\lambda t} . \quad (6.17)$$

Suppose that x failures are observed on n demands. For any one of the failures, denote the corresponding standby time by t_i , $i = 1, \dots, x$. For any one of the successes, denote the corresponding standby time by s_j , $j = 1, \dots, n - x$. All these numbers are observable in principle. Therefore, the likelihood is proportional to

$$\prod_{j=1}^{n-x} e^{-\lambda s_j} \prod_{i=1}^x (1 - e^{-\lambda t_i}) . \quad (6.18)$$

Here the capital pi denotes a product, the analogue of using capital sigma for a sum. This likelihood will be treated in three distinct ways below. First, a simple special case will be considered. Second, an approximation of the likelihood will be developed and used. Finally, a way to use the exact likelihood in Bayesian analysis will be given.

First, consider a simple special case, when all the standby times are equal, say to some number t . This can happen if all the demands are test demands at equally spaced intervals. In this case, the probability of failure on demand is the same for each demand, the quantity p given by Equation 6.17. Therefore, the number of failures in n demands is binomial(n, p). The analysis methods of Section 6.3 can all be used – Bayesian or frequentist estimation of p and all the methods of model validation. At the very end of the analysis, the conclusions in terms of p should be translated into conclusions in terms of λ , by solving Equation 6.17 for

$$\lambda = -\ln(1 - p)/t .$$

This equation for λ can be approximated as

$$\lambda \approx p/t$$

if p is small (say, < 0.1).

This last equation shows that the MLE of λ is approximated by $\hat{p}/t = x/nt$. Here x is the number of failures and nt is the total standby time. This total standby time is approximately the total calendar time, so a simple estimate of λ is the number of failures divided by the total calendar time.

The above simple approach assumes that all the standby times are equal. If the standby times are *approximately* equal, or *nearly all* equal, it is very appealing to use the above technique, calling it an adequate approximation. If, instead, the standby times differ greatly, one of the two approaches given below can be used.

The exact likelihood given in Equation 6.18 can be approximated as follows. It is well known that

$$1 - \exp(-\lambda t_i) \approx \lambda t_i .$$

This is the first order Taylor-series approximation, and is valid when λt_i is small. The error is on the order of $(\lambda t_i)^2$. A second-order approximation is less well known, but it is not hard to show that

$$1 - \exp(-\lambda t_i) \approx \lambda t_i \exp(-\lambda t_i/2) .$$

That is, the two quantities have the same second order Taylor expansions, and they differ only by a term of order $(\lambda t_i)^3$. Therefore, the likelihood in Equation 6.18 is approximately equal to

$$\exp\left(-\sum_{j=1}^{n-x} \lambda s_j\right) \left(\prod_{i=1}^x \lambda t_i\right) \exp\left(-\sum_{i=1}^x \lambda t_i / 2\right) .$$

This is proportional to

$$e^{-\lambda t} \lambda^x$$

where

$$t = \sum_{j=1}^{n-x} s_j + \left(\sum_{i=1}^x t_i / 2 \right).$$

Compare this approximation of the likelihood with Equation 6.1, and see that the approximate likelihood here is proportional to the likelihood of x Poisson events in time t , where t equals the total standby time for the successes plus half the standby time for the failures.

Therefore, all the likelihood-based methods for Poisson data are approximately valid, treating the data as showing x failures in time t . The likelihood-based methods consist of maximum-likelihood estimation and all the Bayesian techniques.

The graphical methods for model validation from Section 6.2 are also probably valid, because they do not require a rigorous justification. The above argument also suggests that the chi-squared test of poolability in Section 6.2 can be used with the present data, because the chi-squared test is only an approximation in any case. However, no simulations to confirm this have been carried out for this handbook.

Finally, we give a third approach, an exact Bayesian method that can be used if the standby times have been recorded, based on Equation 6.18. Figure 6.35 gives a portion of a script for analyzing this type of data with BUGS, based on the exact likelihood. (See Figures 6.13 and 6.31 for similar scripts in other situations.)

```

model
{ for (i in 1:n) {
  p[i] <- 1 - exp(-lambda*t[i])
  x[i] ~ dbern(p[i])
}
lambda ~ dgamma(0.5, 0.00001)
}

```

Figure 6.35 Script for analyzing standby failure data exactly.

In this script, p_i is defined as $1 - \exp(-\lambda t_i)$. The random variable X_i is assigned a **Bernoulli**(p_i) distribution. This means that X_i equals 1 with probability p_i and equals 0 with probability $1 - p_i$. It is the same as a binomial distribution with $n = 1$. Finally, λ is assigned a prior distribution. In Figure 6.35, the prior

distribution is chosen to be close to the Jeffreys noninformative prior for Poisson data, but any proper prior distribution could be used. BUGS requires a proper distribution, so the second parameter of the gamma distribution cannot be exactly zero. An additional required portion of the script, giving the data, is not shown in Figure 6.35.

6.5 Failures to Run during Mission

6.5.1 Estimates and Tests

This type of data can be analyzed using almost exactly the same tools as for event rates in Sec. 6.2. Certain tools carry over exactly, and others are approximately correct.

6.5.1.1 Likelihood-Based Methods: MLEs and Bayesian Methods

Suppose that n systems are run for their missions. (Equivalently, we might assume that a system is run for n missions.) Suppose that x of the runs result in failure, at times t_1, \dots, t_x . The remaining $n - x$ runs are completed successfully, and the systems are turned off at times s_1, \dots, s_{n-x} . Observe the notation: t for a failure time and s for a completed mission time. The likelihood is the product of the densities of times to failure, for the systems that fail, times the probability of no failure, for the systems that did not fail:

$$\prod_i f(t_i) \prod_j \Pr(\text{no failure by } s_j)$$

Under the model introduced in Section 2.4, the failure rate is assumed to be constant, λ , the same for all the systems. Therefore, the time to failure has an exponential distribution. As stated in Appendix A.7.4, the density of an exponential(λ) distribution is

$$f(t) = \lambda e^{-\lambda t}$$

and the cumulative distribution function (c.d.f.) is

$$F(t) = 1 - e^{-\lambda t}.$$

In particular, the probability of no failure by time s is $1 - F(s)$. Substitution of these values into the general expression for the likelihood results in

6.

$$\begin{aligned} \prod_i [\lambda \exp(-\lambda t_i)] \prod_j \exp(-\lambda s_j) \\ = \lambda^x \exp[-\lambda(\sum t_i + \sum s_j)] \\ = \lambda^x \exp(-\lambda t), \end{aligned}$$

where t is defined as $\sum t_i + \sum s_j$, the total running time.

Except for a normalizer that does not depend on λ , this is the Poisson probability of x failures in time t ,

$$\exp(-\lambda t) \lambda^x t^x / x! .$$

Recall that Section 6.2 dealt with x failures in time t . Therefore, any statistical analysis that requires only a multiple of the likelihood is the same in Section 6.2 and here. In particular, the maximum likelihood estimate of λ is x/t . The gamma distributions form the family of conjugate priors, and any Bayesian analysis is carried out the same way for the data here and the data in Section 6.2.

The subtle difference is that $\sum t_i$ is randomly generated here, so t is randomly generated (although if most of the systems do not fail during their missions, the random portion of t is relatively small.) Also, the likelihood here is not a probability, but a combination of densities and probabilities — that explains the missing normalizer in the likelihood. These differences between this section and Section 6.2 result in small differences in the confidence intervals and the tests for poolability.

6.5.1.2 Confidence Intervals

Engelhardt (1995) recommends the following method when all the mission times equal the same value, s . The probability of a system failure before time s is

$$p = F(s) = 1 - \exp(-\lambda s). \quad (6.19)$$

Based on x failures in n trials, find a confidence interval for p , using the methods of Sec. 6.3. Translate this into a confidence interval for λ , using Equation 6.19

$$\begin{aligned} \lambda_{\text{conf}, 0.05} &= -\ln(1 - p_{\text{conf}, 0.05})/s \\ \lambda_{\text{conf}, 0.95} &= -\ln(1 - p_{\text{conf}, 0.95})/s . \end{aligned}$$

This method does not use all of the information in the data, because it ignores the times of any failures, using only the fact that there was a failure at some time

before the mission time s . However, if failures are few the loss of information is small.

Similarly, to perform tests when all the mission times are the same, for example to test whether two data subsets can be pooled, one can work with p , defined by Equation 6.19, and use the tests given in Sec. 6.3. The translation to λ needs to be made only at the very end of the analysis.

When the mission times are not all equal, no exact confidence interval method exists. However, Bayesian intervals can still be found, and are suggested

6.5.1.3 Jeffreys Noninformative Prior

The Jeffreys prior can be worked out exactly, following the process given in Appendix B.5.3.1. If $\lambda \times$ (typical mission time) is small (say, < 0.1), then the Jeffreys prior is approximately the same as in Section 6.2, an improper distribution proportional to $\lambda^{-1/2}$.

6.5.1.4 Tests for Poolability

The above arguments suggest that it is adequate to ignore the random element of t , and use the methods of Sec. 6.2, when estimating λ . For testing whether subsets of the data can be pooled, the same arguments suggest that the chi-squared test of Sec. 6.2 can be used. The chi-squared distribution is only an asymptotic approximation in any case, and can probably be used even when t has a small amount of randomness, although no simulations to confirm this have been carried out for this handbook.

The rest of this section considers a diagnostic plot that was not introduced earlier.

6.5.2 Hazard Function Plot

One plot that is especially useful for failures to run is the hazard function plot. It is used to investigate whether λ is constant during the entire mission. As explained in Appendix A.4.4 for a nonrepairable system, $\lambda \Delta t$ is the approximate probability that the system will fail during a time interval of length Δt , given that it has not yet failed. The precise name for λ is the **hazard rate**, or **hazard function**, although it is often also called the failure rate.

Suppose that the system must run for some mission time, and the data value for that mission is either the mission time, if the system runs to the end without failing, or the failure time, if the system fails during the mission. The outcome, failure or success, is also recorded. The total data set consists of the data from a number of missions.

Now consider the possibility that λ is not constant. Therefore, we write it as $\lambda(t)$. An estimate of $\lambda(t)\Delta t$ at some time t is the number of systems that failed during the interval $(t, t + \Delta t]$ divided by the number of systems that had not yet failed by time t . This leads to the following rather unsatisfactory estimate of $\lambda(t)$. Divide the mission time into little intervals, each of length Δt , with the intervals so short that hardly any of them contain more than one failure time. In an interval with no recorded failures, estimate $\lambda(t)$ by 0. In an interval $(t, t + \Delta t]$ with one failure, estimate $\lambda(t)\Delta t$ by $1/n_t$, where n_t is the number of systems that had not yet failed by time t . Therefore, the estimate of $\lambda(t)$ there is $1/(n_t\Delta t)$. For intervals with more than one failure, set the numerator to the number of failures.

This estimate consists of a number of spikes, at times when failures were observed. Because it is so unsmooth, this estimate is not at all attractive. However, it motivates a very simple estimate of the **cumulative hazard function**, defined as

$$\Lambda(t) = \int_0^t \lambda(u) du .$$

In this definition, the argument t of Λ is the upper limit of integration. Here Λ and λ are related in the same way that a c.d.f. and a density are related. In particular, $\lambda(t)$ is the derivative of $\Lambda(t)$.

A natural and simple estimate of $\Lambda(t)$ is a step function, which is flat except at times when failures occurred. At a time t when a failure occurred, the estimate of Λ jumps by $1/n_t$, where n_t is defined, just as above, as the number of systems that had not yet failed by time t . If exactly simultaneous failures occur, for example because of roundoff in reporting the failure times, the estimate of Λ jumps by the number of failures divided by n_t . This plot is due to Nelson (1982). The full name of the plot is the **cumulative hazard function plot**. This technique is illustrated with the following example.

Example 6.11 EDG failure-to-run times.

Grant et al. (1996) state that 23 failures to run occurred during the EDG tests performed approximately once every 18 months. All these failures were reported by plants subject to Regulatory Guide RG1.108, and there were approximately 665 such tests performed at these plants during the study period. These tests require the EDG to run for 24 hours. Of the 23 failure reports, 19 reported the times to failure. The 19 reported times are given below, in hours.

0.17	0.33	2.67	6.00	11.50
0.23	0.35	3.00	8.00	13.00
0.25	0.93	4.00	10.00	17.78
0.33	1.18	5.50	10.00	

Grant et al. assume that the lack of a reported time is statistically independent of the time at failure, so that the 19 reported times are representative of all 23 times.

There were approximately 665 such tests. Therefore, the cumulative hazard plot jumps by $1/665$ at time 0.17 hours, by $1/664$ at time 0.23 hours, and so forth, until it jumps by $1/647$ at time 17.78. It is important that the duration of all the tests is known to be 24 hours. This fact guarantees that none of the EDGs drop out early, so that after 18 failures 647 EDGs are still running. Actually, this is only approximate, because it ignores the four failures with unreported times.

The jumps are almost the same height, because $1/665$ equals $1/647$ to two significant digits. Therefore Grant et al. plot the cumulative number of failures (a jump of 1 at each failure), instead of the estimated cumulative hazard function. The two graphs make the same visual impression, and the cumulative failure plot was easier to explain in the report. This plot is shown here, as Figure 6.36.

The cumulative hazard plot would differ only in that the vertical scale would be different, and the jumps would not be exactly the same size, though the jumps would be almost the same size in this example.

As explained in introductory calculus courses, when a function is graphed as a curve, the derivative of the function is the slope of the curve. Therefore, the slope of a cumulative hazard plot near time t estimates the derivative of Λ at time t . But the derivative of $\Lambda(t)$ is

6.

$\lambda(t)$. Therefore, a constant slope indicates constant $\lambda(t)$, and a changing slope indicates changing $\lambda(t)$.

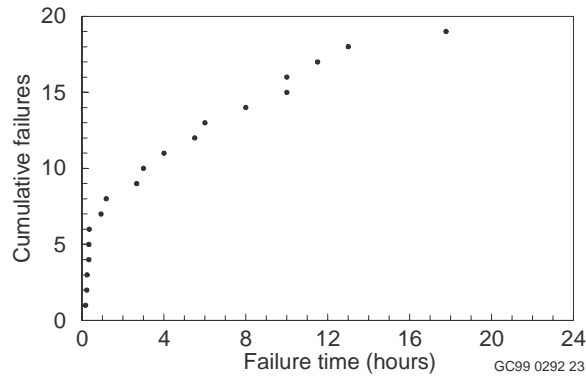


Figure 6.36 Plot of cumulative failure count, a close approximation of plot of cumulative hazard function when only a small fraction of the systems fail.

Grant et al. note that for times less than about one half hour the slope is approximately constant, and steep. It is again constant, but less steep, from about 1/2 hour until about 14 hours, and it is smaller yet after 14 hours. Therefore, Grant et al. estimate three values for λ , corresponding to these three time periods. They comment that the early, middle, and late failures seem to correspond in part to different failure mechanisms.

6.6 Unavailability

The discussion here is presented in terms of trains, although other hardware configurations, from individual components up to an entire reactor, could be considered equally well. A standby train, such as the single train of the HPCI system or a motor-driven train of the AFW system, is normally available if it should be demanded, but sometimes it is out of service for planned or unplanned maintenance. The event of a train being unavailable will be called an **outage** here, and the length of time when it is unavailable is called an **outage time** or **out-of-service time**. The **unavailability** is the probability that the system is unavailable when demanded. More precisely, the **planned-maintenance unavailability** is the probability that the system is out of service for planned maintenance, and the **unplanned-maintenance unavailability** is defined similarly. In summary, outage times are random but the unavailability is a parameter, an unknown constant, denoted here by u . Subscripts "planned" and "unplanned" can be attached to u for clarity if needed.

As mentioned in Section 2.5.2, we assume the following kind of data.

- Only summary data are available, such as the total outage time for each calendar quarter. Such data could be obtained from industry data bases such as EPIX (INPO 1998).

This section is much less detailed and prescriptive than the other sections of this chapter, because data analysis methods are less fully developed for unavailability than for other parameters.

Unavailability data typically are proprietary. Therefore, the methods of this section will be illustrated with the hypothetical data provided in Example 6.12.

Data from many train-months may be reported, (28 in the example, not counting the one month when plant Y was shut down and its trains were not required to be available). The task now is to estimate u_{planned} by a point estimate, and to quantify the uncertainty in the estimate by a confidence interval or a Bayesian distribution.

Denote the exposure time at plant i , train j , and month k by t_{ijk} . Denote the corresponding outage time by O_{ijk} . The upper-case letter emphasizes that the outage time is random. The observed values are denoted by lower case letters o_{ijk} . The corresponding estimator of the unavailability u is the ratio $\hat{U}_{ijk} = O_{ijk}/t_{ijk}$. It is random.

After the data have been observed, the estimate of u is $\hat{u}_{ijk} = o_{ijk} / t_{ijk}$. This gives one such estimate of u for each train-month of data. The estimate from any one train-month is not very good, because it is based on only a small data set.

The data may contain many zeros. Actual data might contain more zeros than shown in this hypothetical example. As a result of the many zeros and few relatively large outage times, the data can be quite skewed. To eliminate some of the zeros and make the data less skewed, the data can be pooled in time periods longer than one month. The sums for the calendar quarter are given in the column on the right of the table in Example 6.12. This assumes that the parameter u does not change from month to month. Denote the summed outage times and exposure times

by o_{ij+} and t_{ij+} , and denote the corresponding estimate by $\hat{u}_{ij+} = o_{ij+} / t_{ij+}$.

Example 6.12 Hypothetical outage times.

Five plants, named U, V, W, X, and Y, each have a chemical and volume control (CVC) system with two trains, named A and B. The first four plants were operating for an entire calendar quarter, but plant Y was operating only for about half the time. Each train was supposed to be available whenever the plant was operating (the exposure time). Exposure times and planned outage times are shown below

	April	May	June	Total
Exposure time	719	744	720	2183
U-A outage time	0	2	0	2
U-B outage time	0.6	0	0.6	1.2
V-A outage time	10.1	1	0	11.1
V-B outage time	9.8	1.3	0	11.1
W-A outage time	13.5	0	5.5	19.0
W-B outage time	8.9	0	0.3	9.2
X-A outage time	0.4	2.1	0	2.5
X-B outage time	0	0	0	0.0
Exposure time	0	446	720	1166
Y-A outage time	0	0.4	5.8	6.2
Y-B outage time	0	0	2.2	2.2

The data can be aggregated further, by summing the quarterly data over the two trains at each plant, and finally by summing the plant data. Assuming that u is the same for both trains and for all plants, this gives estimates \hat{u}_{i++} and \hat{u}_{+++} . The corresponding random quantities are denoted with upper case letters. Notice, this approach pools the numerators and denominators separately and then calculates the ratio. It does not simply average the ratios.

The purpose of this aggregation is to produce multiple observations of an estimator that we denote generically

as X . For example, we might decide to use \hat{U}_{ij+} or \hat{U}_{i++} as the observable X . To carry out the estimation, X must have a distribution that can be analyzed. Bayesian methods will want to use a simple distribution for X , such as lognormal or normal. Frequentist methods will be easiest if X is normal. These distributions do not permit observed values of zero. In addition, the normal distribution does not produce strongly skewed data. We must aggregate enough to obtain data that have these desired characteristics

Table 6.16 gives some sample statistics for these estimates, based on the data of Example 6.12. The sample median is defined in Section 6.7.1.1.2. The sample mean, standard deviation and skewness are defined in Appendix B.2 The skewness is a measure of asymmetry. Positive skewness corresponds to a long tail on the right. Zero skewness corresponds to a symmetrical distribution.

Table 6.16 Sample statistics for estimators of u .

	\hat{U}_{ijk}	\hat{U}_{ij+}	\hat{U}_{i++}	\hat{U}_{+++}
n	28	10	5	1
Mean	3.21E-3	3.29E-3	3.29E-3	3.26E-3
Median	6.95E-4	3.05E-3	3.60E-3	3.26E-3
St. dev., s	5.29E-3	2.81E-3	2.61E-3	—
$s/n^{1/2}$	9.95E-4	8.90E-4	1.17E-3	—
Skewness	1.79	0.61	0.02	—

6.6.1 Frequentist Estimation

The best estimate of u is the sum of the outage times divided by the sum of the exposure times, \hat{u}_{+++} . This ratio of the sums is not quite the same as the average of the ratios, because the various cells are not based on the same exposure times. For example, in Example 6.12 plant Y has fewer exposure hours. Averaging the ratios for the plants treats the data from plant Y with as much weight as the data from the other plants. Summing the outage times and exposure times first, before taking the ratio, gives plant Y only the weight that it should have.

6.

A method to obtain a confidence interval for u uses facts about normally distributed random variables that are presented in Section 6.7.1.2.1 in the context of lognormal random variables, and summarized here.

Suppose that x_1, \dots, x_n are a random sample from a normal distribution with unknown mean μ and standard deviation σ . The maximum likelihood estimate of μ is \bar{x} , the sample mean. A $100(1 - \alpha)\%$ confidence interval for μ is

$$\bar{x} \pm t_{1-\alpha/2}(n-1)s / n^{1/2}$$

where $t_{1-\alpha/2}(n-1)$ is the $(1 - \alpha/2)$ quantile of Student's t distribution with $n - 1$ degrees of freedom, and s is the usual estimate of the standard deviation of X , given in Appendix B. Do not misread the $(n - 1)$ as a multiplier; it is a parameter, the degrees of freedom, of the Student's t distribution. In Table C.3 each row of the table corresponds to one value of the degrees of freedom.

Therefore, a method is to aggregate data until the corresponding estimates appear to be approximately normally distributed, and then construct an overall estimate and confidence interval based on those estimates.

In Example 6.12, the \hat{u}_{ijk} values are not a random sample from a normal distribution. They are too skewed, as is seen by the fact that the mean ($3.21E-3$) is very different from the median ($6.95E-4$), and the skewness (1.79) is far from zero.

Pooling the three months for each train makes the distribution much more nearly symmetrical: the mean and median are within 10% of each other, and the skewness is down to 0.61. When months and trains are pooled, the distribution appears to be fully symmetrical: the mean and median are within 10% of each other, and the sample skewness is virtually zero. This indicates that the values of $x_i \equiv \hat{u}_{i++}$ may be treated as a random sample from a normal distribution. A goodness-of-fit test could be performed, as discussed in Section 6.7, but no departure from normality will be detectable with only five observations.

Therefore, a 90% confidence interval for u_{planned} is

$$3.29E-3 \pm 2.132 \times 1.17E-3 \\ = 3.29E-3 \pm 2.49E-3$$

because 2.132 is the 95th percentile of the Student's t distribution with 4 degrees of freedom. Thus, the lower and upper confidence limits are

$$u_{\text{conf},0.05} = 8.E-4 \\ u_{\text{conf},0.95} = 5.8E-3 .$$

This interval is approximate because the \hat{u}_{i++} values come from an approximately normal distribution.

Table 6.16 shows that the mean hardly changes as various levels of aggregation are used. Similarly, $s/n^{1/2}$ hardly changes. They change somewhat because of randomness of the data and the unbalancedness of the data — the inequality of the exposure times. This shows that data aggregation is not a magic trick to improve the estimator or reduce the variance of the estimator. The only purpose of the pooling is to eliminate the skewness and permit the use of normal methods.

If the data were more badly unbalanced, the data aggregation could be modified to improve the balance. For example, if plant Y had contributed only one month of data, the two trains could have been aggregated immediately, to have produced an estimate based on two train-months of data. This would be comparable to the three train-months of data used for estimating the unavailability of a single train in any other plant.

Real data may not allow aggregation across plants. That is, the observed unavailability may differ greatly from plant to plant, because of differences in plant-maintenance procedures or in data-reporting policies. In such a case, plant-specific data must be used (assuming that the data reporting for the plant appears reasonable). With several years of such data, one might aggregate within calendar quarters or calendar years or some other time period.

6.6.2 Bayesian Estimation

When X is normally distributed, the conjugate prior distribution is given in Section 6.7.1.2.1 in the context of the lognormal distribution. In particular, the non-informative prior results in \bar{x} as the posterior mean for u , and a $100(1 - \alpha)\%$ credible interval for u is given by

$$\bar{x} \pm t_{1-\alpha/2}(n-1)s / n^{1/2} .$$

The posterior distribution of u follows from the fact that the expression

$$(U - \bar{x}) / (s / \sqrt{n})$$

has a Student's t distribution with $n-1$ degrees of freedom. The parameter u is capitalized here to emphasize that it is the quantity with the probability distribution.

In Example 6.12, based on the noninformative prior, the posterior 90% credible interval goes from 8.E-4 to 5.8E-3, as calculated above.

Although Section 6.7.1.2.1 gives the update formulas when using an informative conjugate prior, no example of this updating is worked out here. The reason is that justification of an informative prior is too complicated to carry out here. The update formulas are straightforward, after the hard work of deciding on a prior.

If X is modeled as having a lognormal distribution, then the same methods can be followed. Work with $\ln(X)$, which is normally distributed, instead of with X . The only complication is that the unavailability u is not one of the prior parameters. Instead, u is the mean of X , $\exp(\mu + \sigma^2/2)$. Therefore, let $\ln(X)$ be normal(μ, σ^2). When the joint posterior distribution of μ and σ^2 has been found, randomly generate values from this distribution. For each randomly generated pair (μ, σ^2), calculate the value of $u = \exp(\mu + \sigma^2/2)$. This sample mimics the desired posterior distribution.

Nonconjugate priors can also be considered. Then the posteriors cannot be found by simple algebraic updating. Instead, random sampling is the most promising tool. For examples of such sampling in more complicated settings, see the discussions of Bayesian analysis of trends in Section 7.2, and of hierarchical Bayes models in Section 8.3.3.

6.6.3 Model Validation

A crucial feature of the simple method proposed above is aggregation of data, to reduce skewness. An implicit assumption when pooling data subsets is that the two subsets correspond to the same distribution. Therefore, one may try to check this assumption.

The methods discussed in detail in Sec. 6.7.2.1 may be used, although the data may not be of good enough quality to show much. In particular, box plots may be used to suggest whether subsets can be pooled. The Kruskal-Wallis test, the nonparametric analogue of the analysis-of-variance test for equal means of normally distributed random variables, can be used to test equality of the unavailability in the data subsets.

To construct box plots or to perform the Kruskal-Wallis test, the data subsets must contain multiple observations. In Example 6.12, the data for Plant i could consist of up to six values of \hat{u}_{ijk} , or two values of \hat{u}_{ij+} , or two or three values of \hat{u}_{i+k} . When so many of the observations are tied at zero, the test probably will not detect a statistically significant difference between trains.

The methods are identical to those of Sec. 6.7.2.1, where they are covered in full detail. Therefore, no further discussion is given here.

6.7 Recovery Times and Other Random Duration Times

The previous analyses have all involved a single parameter, λ or p or u . The analysis of duration times is different because now a distribution must be estimated, not just a single parameter.

A distribution can be estimated in many ways. If the form of the distribution is assumed, such as exponential or lognormal, it is enough to estimate one or two parameters; the parameter or parameters determine the distribution. If the form of the distribution is not assumed, the distribution can be estimated nonparametrically, or characteristics of the distribution, such as moments or percentiles, can be estimated.

To test whether data sets can be combined (pooled), both parametric tests and nonparametric tests exist. The parametric tests typically test whether the means or variances of two distributions are equal, when the distributions have an assumed form. The most common nonparametric tests test equality of the distributions against the alternative that one distribution is shifted sideways from the other.

6.

This section is long, because so many distribution models can be assumed and because the model assumptions can be violated in so many ways. A brief outline of the section is as follows.

- 6.7.1 Characterization of a single distribution
 - Estimation of moments, percentiles, c.d.f.s
 - Fitting of four parametric models (frequentist and Bayesian parameter estimates)
- 6.7.2 Model validation (graphs and hypothesis tests)
 - Poolability, trend
 - Goodness of fit to assumed parametric models
 - Consistency of data with prior for Bayesian parameter estimates
- 6.7.3 Nonparametric density estimation

Many of the methods will be illustrated using the data of Example 6.13, taken from Atwood et al. (1998).

This example shows the times when power could have been recovered, for plant-centered LOSP events, that is, for events not caused by grid problems or by widespread severe weather. (Real life is complicated: sometimes a plant does not restore power as quickly as it could, and the event report states when power was actually restored, and when it could have been restored. The times given by Atwood et al. 1998 as “recovery times” show when power could have been restored, if that time was reported and different from the actual recovery time.) Discussions of this example will use the terms **recovery time** and **duration** interchangeably. Momentary events (duration less than two minutes) and events with no reported duration have been excluded. For common-cause events that affected multiple units at a site, the average recovery time is used.

The group P exists because some plants are permitted to remain at power during certain LOSP events.

Throughout this section, the random variable is denoted by T , because typically the random quantity is a duration time, such as time to recovery of the system. Examples were given in Section 2.6.1: time until restoration of offsite power, duration of a repair time, and others. Let F denote the c.d.f. of T , $F(t) = \Pr(T \leq t)$. It is assumed that n times will be observed, T_1, T_2, \dots, T_n . The assumptions of Section 2.6.2 are repeated here.

- The T_i s are independent,

- Each T_i has the c.d.f. $F(t)$.

Example 6.13 LOSP recovery times.

Atwood et al. (1998) report 115 times of recovery of lost offsite power. The data are categorized into three possible values for plant status: T, S, and P, with meanings explained in the table below. The durations in minutes and the dates (MM/DD/YY) are shown.

P: Plant remained at power throughout LOSP event (8 times)		
6 03/01/80	113 01/18/96	385 04/11/94
45 07/25/85	147 06/03/80	1138 01/03/89
65 07/16/88	355 11/12/90	
S: Plant was shut down before and during LOSP event (62 times)		
2 06/04/84	14 11/16/84	60 06/22/91
2 08/17/87	14 02/01/81	60 06/16/89
2 06/29/89	15 04/27/81	62 07/15/80
2 05/21/94	15 12/19/84	67 03/13/91
3 06/26/93	15 10/12/93	73 08/28/85
3 10/22/84	17 04/26/83	77 03/29/92
3.5 11/21/85	17 10/14/87	97 01/08/84
4 04/22/80	20 03/23/92	120 06/05/84
4 04/04/87	22 08/24/84	120 01/16/81
4 10/20/91	24 07/29/88	127 01/20/96
5 05/03/84	24 07/29/88	132 02/27/95
8 06/24/88	29 03/20/91	136 04/08/93
9 12/26/88	29 09/16/87	140 03/20/90
10 08/01/84	29 05/14/89	155 03/05/87
10 04/28/92	35 04/02/92	163 10/08/83
10 12/23/81	37 03/21/87	240 11/14/83
11 10/04/83	37 05/19/93	240 03/07/91
11 07/24/91	37 07/09/90	335 04/29/85
12 06/22/93	43 05/07/85	917 10/21/95
12 07/19/86	53 09/11/87	1675 11/18/94
14 02/26/90	59 10/16/87	
T: Plant tripped because of LOSP event (45 times)		
2 02/28/84	20 08/21/84	90 02/12/84
4 11/21/85	20 07/16/84	90 03/29/89
4 11/17/87	20 06/27/91	90 06/17/89
5 08/16/85	24 06/15/91	95 12/31/92
6 05/03/92	25 10/03/85	95 12/31/92
10 09/10/93	29 06/22/82	95 10/16/88
10 10/12/93	38 07/17/88	96 12/27/93
11 07/26/84	40 02/11/91	100 01/28/86
13 10/07/85	45 01/16/90	106 06/03/80
14 08/13/88	45 03/25/89	118 07/23/87
15 02/16/84	46 01/01/86	118 07/23/87
15 09/14/93	57 10/19/92	277 04/23/91
19 10/25/88	60 03/21/91	330 02/06/96
20 12/12/85	60 10/22/85	388 07/14/87
20 03/27/92	62 07/15/80	454 08/22/92

A data set satisfying these assumptions is called a **random sample** from the distribution. Sometimes the

T_i s are called **independent identically distributed** (i.i.d.). The term random sample also refers to the observed values, t_1, t_2, \dots, t_n . The data are used to estimate properties of the distribution. This can also be described as estimating properties of the population, where the **population** is the infinite set of values that could be randomly generated from the distribution.

6.7.1 Characterization of Distribution

6.7.1.1 Nonparametric Description

The tools in this subsection are called nonparametric because they do not require any assumption about the form of the distribution. For example, the distribution is not assumed to be lognormal, exponential, or of any other particular form.

6.7.1.1.1 Moments

To estimate the population mean μ or a population variance σ^2 , two simple estimators are the **sample mean**, defined as

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n T_i$$

and the **sample variance**, defined as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (T_i - \bar{T})^2 .$$

The sample mean and sample variance are known to be **unbiased** for the population mean and variance, respectively. In other words, $E(\bar{T}) = \mu$ and $E(S^2) = \sigma^2$. These statements are true regardless of the distribution F , requiring only the assumptions of a random sample. The **sample standard deviation**, S , is the square root of the sample variance.

These are all-purpose estimators, but they are not the only possible estimators. For example, the variance of an exponential distribution is the square of the mean. Therefore, a good estimator of the variance would be the square of the estimator of the mean. This estimator relies heavily on the assumption of exponentiality, whereas the above estimators make no such assumptions. General principles of estimation are discussed in Appendix B.4.1.

6.7.1.1.2 Percentiles

To estimate percentiles of a distribution, it is useful to put the data in ascending order from the smallest to the largest observation. The recovery times in Example 6.13 have been arranged this way. The variables obtained by ordering the random sample are called the **order statistics**, and are denoted by $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(n)}$. The observed values are written $t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(n)}$. Some important estimates based on the order statistics are the sample median, other sample percentiles, and the sample range. The general definition of the 100 q th sample percentile, where $0 < q < 1$, is a number t such that the fraction of observations that are $\leq t$ is at least q and the fraction of observations that are $\geq t$ is at least $1 - q$.

For example, the **sample median** is defined to be t such that at least half (because $q = 0.5$) of the observations are $\leq t$ and at least half (because $1 - q = 0.5$) are $\geq t$. This boils down to the following. If n is odd, the sample median is the "middle" order statistic, $t_{(m)}$ with $m = (n + 1)/2$. If n is even, with $m = n/2$, there is no unique "middle" order statistic. Any number between the two middle order statistics, $t_{(m)} \leq t \leq t_{(m+1)}$, could be used, although nearly everyone uses the average of the two middle order statistics $(t_{(m)} + t_{(m+1)})/2$ as "the" sample median.

The other sample percentiles are defined similarly, with some averaging of two order statistics if necessary. Note that the sample 90th percentile is $t_{(n)}$ if $n < 10$, the sample 95th percentile is $t_{(n)}$ if $n < 20$, and so forth.

Order statistics that are sometimes used are: the lower and upper **quartile**, defined as the 25th and 75th percentiles; percentiles that include most of the distribution, such as the 5th and 95th percentiles; and the extremes, $t_{(1)}$ and $t_{(n)}$. The **interquartile range** is the upper quartile minus the lower quartile. The **sample range** is the difference between the largest and smallest ordered observations, $t_{(n)} - t_{(1)}$. Be careful with interpretation. As data continue to be collected, the sample interquartile range stabilizes at the interquartile range of the distribution, but the sample range does not stabilize at all — it just grows every time a new t is observed that is outside the former observations.

6.

The sample median has the advantage of not being strongly influenced by extreme observations. The sample mean, on the other hand, can be strongly influenced by even one extreme data value. The sample variance is even more sensitive to extreme values, because it is based on squared terms. Therefore, the sample standard deviation, defined as the square root of the sample variance, is also sensitive to extreme terms. Other measures of dispersion, such as the interquartile range, are much less sensitive to extreme values. In general, sample percentiles are much less sensitive to extreme observations than are sample moments.

The recovery times of Example 6.13 have sample moments and percentiles given in Table 6.17.

Table 6.17 Statistics based on the recovery times (minutes) of Example 6.13.

	P	S	T
<i>n</i>	8	62	45
Stand. deviation	373.2	241.4	99.9
95th %ile	1138	240	330
75th %ile (upper quartile)	370	73	95
Mean	281.75	92.3	73.4
50th %ile (median)	130	24	40
25th %ile (lower quartile)	55	10	15
5th %ile	6	2	4

For the P group, the sample median is taken as the average of the two middle numbers. Even though the S group has an even number of observations, its sample median is unique, because t_{31} and t_{32} happen to be equal! The T group has an odd number of observations, so its sample median is unique, t_{23} .

The S group has one very extreme value, which influences the moments. The sample mean for this group is larger than the upper quartile — someone who considers the mean to be “the” average could say that more than 75% of the observed times are below average. This happens with skewed distribu-

tions. This is one reason why many people prefer percentiles to moments for describing a skewed distribution.

There are situations in which some of the times are not observed. Sec. 6.5 dealt with such a situation, when the times of interest were times of EDG failure to run, and not all these times were reported. In the present section, nearly all the times are assumed to be observed, with no systematic bias in which times fail to be observable.

6.7.1.1.3 The Empirical Distribution Function

An estimate of $F(t)$ called the **empirical distribution function** (EDF) is defined as follows: For an arbitrary value of $t > 0$, define

$$\hat{F}(t) = (\text{Number of observations} \leq t) / n.$$

The EDF is a step function. It increases by $1/n$ at each observed time if all observations are distinct. More generally, if there are m times equal to t , $\hat{F}(t)$ has a positive jump of m/n at t .

In some settings the function

$$1 - F(t) = \Pr(T > t)$$

is of interest. If T is the time until failure, $1 - F(t)$ is called the **reliability function**, $R(t)$, in engineering contexts, and the **survival function**, $S(t)$, in medical contexts. A suitable word remains to be coined when T is the time until recovery or repair. The empirical reliability function, or the empirical survival function, is defined as 1 minus the EDF. Anything that can be done with F can be translated in terms of $1 - F$, so this discussion will only consider F .

With a little mental exercise, the EDF can be expressed in familiar terms. For any particular t , let p denote $F(t) = \Pr(T \leq t)$. In the data, define a “demand” to be the generation of an observed time, and define the i th observation t_i to be a “failure” if $t_i \leq t$. By the assumptions for a random sample, any observation has probability p of being a failure, and the outcomes (failures or successes) are statistically independent of each other. By its definition, $\hat{F}(t)$ is the number of failures

divided by the number of demands, which is \hat{p} , familiar from Section 6.3.1. Therefore, $\hat{F}(t)$ is an unbiased estimator of $F(t)$ at any t . It is close to $F(t)$ when the number of observations is large, and a confidence interval for $F(t)$ can be constructed, the familiar confidence interval for p .

Figure 6.37 shows the EDF based on the data in Example 6.13 for group T.

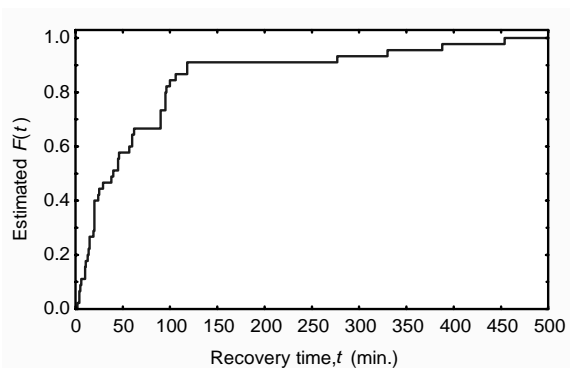


Figure 6.37 Empirical distribution function (EDF) for the data from group T in Example 6.13.

6.7.1.1.4 Histogram Estimate of the Density

The eye smooths the EDF, compensating for its jagged form. To accomplish the same sort of smoothing for a density estimate, group the observed times into bins of equal width, count the number of observations in each bin, and plot the histogram, a form of bar chart with the height of each bin equal to the number of observations in the bin. Some software packages can rescale the height so that the total area equals 1, making a true density estimate. If the vertical axis shows counts, the histogram is *proportional* to an estimate of the density. Books and Ph. D. theses have been written on density estimation, and some modern density estimators are quite sophisticated. A few such are given in Sec. 6.7.3. Nevertheless, the lowly histogram is often adequate for PRA purposes.

Figures 6.38 and 6.39 show two histograms for the data from the above EDF, using two different bin widths. The analyst must decide what bin width gives the most reasonable results, based on belief about how smooth or ragged the true density might

be. Most people would judge Figure 6.39 to be too rough, and would therefore choose wider bins.

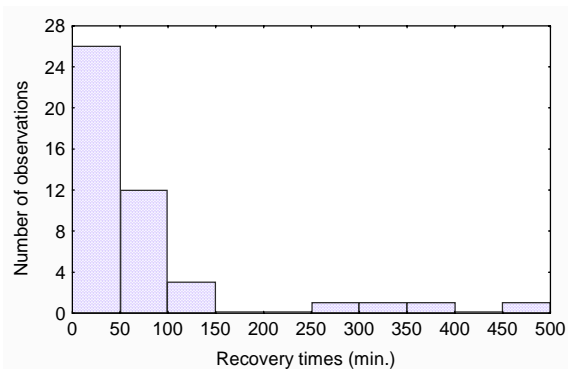


Figure 6.38 Histogram of the data from group T in Table 6.16, with bin width 50.

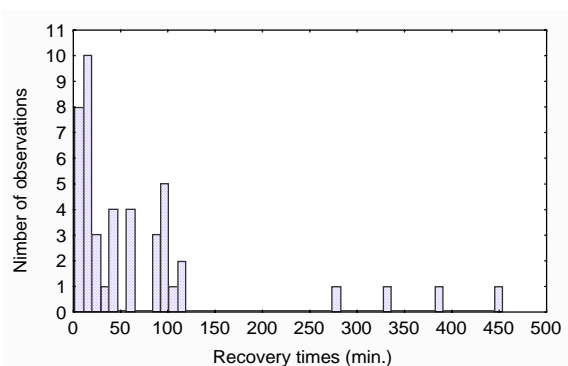


Figure 6.39 Histogram of same data, with bin width 10.

6.7.1.2 Fitting a Parametric Distribution

Sometimes it is desirable to fit some assumed distributional form to data. This subsection gives estimators if the assumed distribution is lognormal, exponential, gamma, or Weibull. Bayesian and non-Bayesian estimates are given, with much of the latter taken from an INEEL report by Engelhardt (1996).

6.7.1.2.1 Lognormal Distribution

This model assumes that T has a lognormal distribution, or equivalently, that $\ln T$ has a normal(μ, σ^2) distribution. Define $Y = \ln T$.

Frequentist Estimates. The usual estimates of μ and σ^2 are the maximum likelihood estimates:

6.

$$\bar{y} = \frac{1}{n} \sum_i y_i$$

and

$$s_Y^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2$$

These estimates have the same form as those given in Section 6.7.1.1.1 for the mean and variance of T , but these are for $\ln T$. Calculate the estimates of these parameters to determine the estimated normal distribution of $\ln T$, which determines the estimated lognormal distribution of T .

The material below is presented in many statistics books, based on the fact that $\ln T$ has a normal distribution. The distribution of $(n-1)S_Y^2 / \sigma^2$ is chi-squared with $n-1$ degrees of freedom. It follows that a two-sided $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left((n-1)s_Y^2 / \chi_{1-\alpha/2}^2(n-1), (n-1)s_Y^2 / \chi_{\alpha/2}^2(n-1) \right).$$

Here $\chi_q^2(n-1)$ is the q quantile, that is, the $100q$ percentile, of the chi-squared distribution with $n-1$ degrees of freedom.

The distribution of \bar{Y} is normal($\mu, \sigma^2/n$). If σ^2 is known, it follows that a $100(1-\alpha)\%$ confidence interval for μ is $\bar{y} \pm z_{1-\alpha/2} \sigma / \sqrt{n}$, where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution. For example, $z_{0.95}$ gives a two-sided 90% confidence interval.

In the more common case that both μ and σ^2 are unknown, use the fact that

$$(\bar{Y} - \mu) / (S_Y / \sqrt{n})$$

has a Student's t distribution with $n-1$ degrees of freedom. It follows that a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{y} \pm t_{1-\alpha/2}(n-1) s_Y / \sqrt{n},$$

where $t_{1-\alpha/2}(n-1)$ is the $1-\alpha/2$ quantile of the Student's t distribution with $n-1$ degrees of freedom. For example, $t_{0.95}(n-1)$ gives a two-sided 90% confidence interval.

Bayesian Estimation. Bayesian estimates are given here.

Conjugate Priors. The conjugate priors and update formulas are presented by Lee (1996, Sec. 2.13). They depend on four prior parameters, denoted here as $d_0, \sigma_0^2, n_0,$ and μ_0 . The notation here tries to follow the notation used elsewhere in this handbook. It is not the same as Lee's. Quantities with subscripts, such as σ_0^2 or d_1 , are numbers. Quantities without subscripts, σ^2 and μ , have uncertainty distributions.

It is useful to think of having d_0 degrees of freedom, corresponding to $d_0 + 1$ prior observations for estimating the variance, and a prior estimate σ_0^2 . More precisely, let the prior distribution for $\sigma^2 / (d_0 \sigma_0^2)$ be inverted chi-squared with d_0 degrees of freedom. That is, $d_0 \sigma_0^2 / \sigma^2$ has a chi-squared distribution with d_0 degrees of freedom. Therefore it has mean d_0 , and therefore the prior mean of $1/\sigma^2$ is $1/\sigma_0^2$. (See Appendix A.7.7 for more information on the inverted chi-squared distribution.)

An alternative notation for the above paragraph would define the **precision** $\tau = 1/\sigma^2$, and the prior precision $\tau_0 = 1/\sigma_0^2$. Then the prior distribution of $d_0 \tau / \tau_0$ is chi-squared with d_0 degrees of freedom. Although we shall not use this parameterization, it has adherents. In particular, BUGS (1995) uses τ instead of σ^2 as the second parameter of the normal distribution; see Spiegelhalter et al. (1995).

Conditional on σ^2 , let the prior distribution for μ be normal with mean μ_0 and variance σ^2/n_0 . This says that the prior knowledge of μ is equivalent to n_0 observations with variance σ^2 . It is not necessary for n_0 to have any relation to d_0 .

The Bayes update formulas are

$$\begin{aligned} d_1 &= d_0 + n \\ n_1 &= n_0 + n \\ \mu_1 &= (n_0 \mu_0 + n \bar{y}) / n_1 \end{aligned}$$

$$\sigma_1^2 = \left[d_0 \sigma_0^2 + (n-1) s_Y^2 + \frac{n_0 n_1}{n_0 + n_1} (\mu_0 - \bar{y})^2 \right] / d_1$$

Here the subscript 1 identifies the posterior parameters. The posterior distributions are given as follows. First, $\sigma^2 / (d_1 \sigma_1^2)$ has an inverted chi-squared distribution with d_1 degrees of freedom. That is, the posterior mean of $1/\sigma^2$ is $1/\sigma_1^2$, and a two-sided $100(1-\alpha)$ credible interval for σ^2 is

$$(d_1 \sigma_1^2 / \chi_{1-\alpha/2}^2(d_1), d_1 \sigma_1^2 / \chi_{\alpha/2}^2(d_1)).$$

Conditional on σ^2 , the posterior distribution of μ is normal($\mu_1, \sigma^2/n_1$). Therefore, conditional on σ^2 , a two-sided $100(1-\alpha)\%$ credible interval for μ is

$$\mu_1 \pm z_{1-\alpha/2} \sigma / \sqrt{n_1}.$$

The marginal posterior distribution of μ , that is, the distribution that is not conditional on σ^2 , is as follows. The expression

$$(\mu - \mu_1) / (\sigma_1 / \sqrt{n_1})$$

has a Student's t distribution with d_1 degrees of freedom. It follows that a $100(1-\alpha)\%$ credible interval for μ is

$$\mu_1 \pm t_{1-\alpha/2}(d_1) \sigma_1 / \sqrt{n_1}.$$

Noninformative Prior. The joint noninformative prior for (μ, σ^2) is proportional to $1/\sigma^2$. Lee (1997, Sec. 2.13) presents this prior, as do Box and Tiao (1973, Sec. 2.4). Lee points out that when $d_0 = -1$, $n_0 = 0$, and $\sigma_0^2 = 0$, the conjugate prior distribution reduces to the noninformative prior, and the credible intervals then agree numerically with the confidence intervals given above.

Possible Further Analyses. Some data analyses require only the posterior distribution of one or both parameters. In that case, use the above posterior distributions, with either an informative or noninformative prior. Other analyses require more, such as simulation of a set of lognormal times X or a credible interval for the mean of X . If so, simulation of

the quantity of interest is a useful technique. Begin each case of the simulation by generating a value of σ^2 from its posterior distribution. Then generate a value of μ from its distribution conditional on σ^2 . Then do whatever is required next to obtain the quantity of interest: generate a random value of X from the lognormal(μ, σ) distribution, or calculate $E(X) = \exp(\mu + \sigma^2/2)$, or calculate whatever else is needed. Save the quantity of interest produced in this way. Repeat this process as many times as needed to obtain a sample that accurately represents the distribution of interest.

Model Validation. Model validation is discussed in general in Section 6.7.2. Many of the methods given there are applicable to any assumed distribution. Some methods, however, have been developed just for the normal and lognormal distributions. They are contained in Sections 6.7.2.1.2, 6.7.2.2.2, and 6.7.2.3.2.

6.7.1.2.2 Exponential Distribution

The exponential distribution is presented in Appendix A.7.4, with two possible parameterizations. The first uses $\lambda = 1/E(T)$, and the second uses $\mu = 1/\lambda = E(T)$. In data analysis, sometimes one parameter seems more natural and convenient and sometimes the other does. In the two parameterizations, the likelihood function is

$$\lambda^n \exp(-\lambda \sum t_i)$$

or

$$\mu^{-n} \exp(-\sum t_i / \mu).$$

Frequentist Estimation. It can be shown that the MLE of μ is the sample mean, \bar{t} . Therefore, to estimate the distribution, estimate μ by \bar{t} . This determines the estimated exponential distribution. The corresponding estimate of $\lambda \equiv 1/\mu$ is $1/\bar{t}$.

For a $(1-\alpha)$ confidence interval, or equivalently a $100(1-\alpha)\%$ confidence interval, the lower limit for λ is

$$\lambda_{\text{conf. } \alpha/2} = \frac{\chi_{\alpha/2}^2(2n)}{2\sum t_i}$$

and the upper limit is

6.

$$\lambda_{\text{conf}, 1 - \alpha/2} = \frac{\chi_{1-\alpha/2}^2(2n)}{2\sum t_i}.$$

(See Martz and Waller 1991.) Confidence limits for $\mu = 1/\lambda$ are obtained by inverting the confidence limits for λ . For example, the lower confidence limit for μ equals 1 divided by the upper confidence limit for λ .

Bayesian Estimation. Now consider Bayesian estimation.

Conjugate Prior. The gamma distribution is a conjugate prior for λ . That is, let t_1, \dots, t_n be independent observations from an exponential(λ) distribution. Let the prior distribution of λ be gamma(α_0, β_0). This uses the same parameterization as when λ is a Poisson parameter, so that β_0 has units of time and the prior mean of λ is α_0/β_0 . A direct calculation shows that the posterior distribution of λ is also gamma, with posterior parameters

$$\begin{aligned}\alpha_1 &= \alpha_0 + n \\ \beta_1 &= \beta_0 + \sum t_i.\end{aligned}$$

The subscript 1 identifies the posterior parameters. The prior parameters have a simple intuitive interpretation – the prior information is "as if" α_0 duration times had been observed with total value β_0 .

The percentiles of the posterior distribution are given by

$$\lambda_p = \frac{\chi_p^2(2\alpha_1)}{2\beta_1}$$

Therefore, for example, a two-sided 90% credible interval has end points

$$\lambda_{0.05} = \frac{\chi_{0.05}^2(2\alpha_1)}{2\beta_1}$$

and

$$\lambda_{0.95} = \frac{\chi_{0.95}^2(2\alpha_1)}{2\beta_1}.$$

There are two possible ways to perform the corresponding analysis in terms of μ . (a) One way is to perform the above analysis in terms of λ , and then translate the answer into answers for $\mu = 1/\lambda$. Be

careful when doing this. The percentiles translate directly, with the p th percentile $\mu_p = 1/\lambda_p$. The moments do not translate directly, however. For example, the mean of μ is not 1 divided by the mean of λ . (b) The other way is to let μ have an inverted gamma distribution. This distribution is defined in Appendix A.7.7.

Either analysis gives exactly the same results. The second approach is just a disguised version of the first approach, using a different distribution to avoid introduction of the symbol λ .

Noninformative Prior. The Jeffreys noninformative prior for λ can be expressed as a gamma(0, 0) distribution. This is an improper distribution, that is, it does not integrate to 1, but it results in proper posterior distributions as long as some data have been observed. Note, this prior is slightly different from the Jeffreys prior when the data have a Poisson distribution. When the gamma(0, 0) prior is used with exponential data, the posterior parameters reduce to

$$\begin{aligned}\alpha_{\text{post}} &= n \\ \beta_{\text{post}} &= \sum t_i.\end{aligned}$$

Then the Bayes posterior credible intervals are numerically equal to the confidence intervals. If the purpose of a "noninformative" prior is to produce intervals that match confidence intervals, this purpose has been perfectly accomplished.

Discussion. The above work has illustrated some facts that are true in general. When the observations have a discrete distribution, such as Poisson or binomial, the so-called noninformative priors do not produce credible intervals that exactly match confidence intervals. This is related to the fact that confidence intervals from discrete data do not have exactly the desired confidence coefficient. Instead, they are constructed to have *at least* the desired long-run coverage probability. The situation is different when the observations are continuously distributed, as in the present case with exponentially distributed times. In this case, the confidence intervals have exactly the desired long-run coverage probability, and posterior credible intervals based on the noninformative prior are numerically equal to the confidence intervals.

Nonconjugate priors can also be used, of course. The procedure is very similar to that in Section 6.2.2.6, but now using the exponential likelihood given above. Therefore it is not discussed here.

Model Validation. Model validation is discussed in general in Section 6.7.2. Many of the methods given there are applicable to any assumed distribution. A few methods, however, have been developed just for the exponential distribution. They are mentioned in Sections 6.7.2.3.1 and 6.7.2.4.2.

6.7.1.2.3 Gamma Distribution

The distribution of T is gamma(α , τ) if the density is

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

Note, this is a different parameterization from the previous section and from Equation 6.3. This parameterization is related to the earlier parameterization by the relation $\tau = 1/\beta$. In the present context, t and τ both have units of time.

The MLEs of the parameters are given by Bain and Engelhardt (1991, p. 298) or by Johnson et al. (1994, Sec. 17.7). They are the solutions of the equations

$$\tau = \bar{t} / \alpha$$

$$\ln(\alpha) - \psi(\alpha) = \ln(\bar{t} / \tilde{t}) ,$$

where $\psi(u) = \Gamma'(u)/\Gamma(u)$ is the digamma function, calculated by some software packages, and

$$\tilde{t} = \exp\left[(1/n)\sum \ln t_i\right] ,$$

the geometric mean of the observed times. The second equation must be solved by numerical iteration. Bain and Engelhardt (1991, p. 298) give a table of approximate solutions, which may be interpolated.

The MLEs of the two parameters determine the estimated gamma distribution.

Bayes estimation is more complicated than elsewhere in Chapter 6 because the gamma distribution has two parameters, and these two parameters must have a joint

distribution. Martz and Waller (1991, Sec. 9.5.2) cite Lwin and Singh (1974) for an analysis that was feasible in the 1970s. A simpler approach today would use the freely available package BUGS (1995), described in Section 8.2.3.3.3. BUGS is designed for models with many unknown parameters, and should make short work of a model with only two. The joint prior distribution would not need to be conjugate.

6.7.1.2.4 Weibull Distribution

A three-parameter Weibull distribution is given in Appendix A.7.5. A two-parameter form of the Weibull distribution is given here, by setting the location parameter θ to zero. The density is

$$f(t) = (\beta / \alpha)(t / \alpha)^{\beta-1} \exp\left[-(t / \alpha)^\beta\right] .$$

As with the gamma distribution, the maximum likelihood equations do not have closed-form solutions. The estimates must be found by iteratively solving

$$\frac{\sum t_i^\beta \ln(t_i)}{\sum t_i^\beta} - \frac{1}{\beta} - \frac{1}{n} \sum \ln t_i$$

and

$$\alpha = \left(\frac{1}{n} \sum t_i^\beta\right)^{1/\beta} .$$

Zacks (1992, Section 7.5) gives the following simple method for solving the first equation. Begin with $\hat{\beta}_0 = 1$. Then repeatedly solve the equation

$$\hat{\beta}_{n+1} = 1 / \left[\frac{\sum t_i^{\hat{\beta}_n} \ln(t_i)}{\sum t_i^{\hat{\beta}_n}} - \frac{1}{n} \sum \ln t_i \right]$$

with $n = 0, 1, 2, \dots$. The value of $\hat{\beta}_n$ converges quickly to the MLE $\hat{\beta}$. Then set

$$\hat{\alpha} = \left(\frac{1}{n} \sum t_i^{\hat{\beta}}\right)^{1/\hat{\beta}} .$$

6.

For more information, see Zacks (1992) or Bain and Engelhardt (1991).

Alternatively, a simple approximate graphical estimate is based on the hazard function. Plots of the cumulative hazard were discussed in Sec. 6.5.2. It can be shown that the cumulative hazard function of the Weibull distribution is

$$H(t) = (t/\alpha)^\beta.$$

Therefore, estimate the cumulative hazard function as explained in Section 6.5.2, by jumping at each observed time, with the size of the jump equal to 1 divided by the number of times that have not yet been equalled or exceeded. The jump at $t_{(1)}$ is $1/n$, the jump at $t_{(2)}$ is $1/(n-1)$, and so forth until the final jump at $t_{(n)}$ is 1. Call this estimate $\hat{H}(t)$. The equation for the Weibull cumulative hazard function can be rewritten as

$$\log H(t) = \beta \log t - \beta \log \alpha, \quad (6.20)$$

which is linear in $\log t$. Therefore, plot $\log[\hat{H}(t)]$ against $\log t$, that is, plot $\hat{H}(t)$ against t on log-log paper, and fit a straight line to the plot by eye. Pick a point on the line and substitute those values of t and $\hat{H}(t)$ into Equation 6.20. This is one equation that β and $\log \alpha$ must satisfy. Pick a second point on the line and obtain a second equation in the same way. Solve those two equations for β and $\log \alpha$, thus obtaining estimates of β and α . In the calculations, it does not matter whether natural logarithms or logarithms to base 10 are used, as long as the same type is used everywhere.

This plot also gives a diagnostic test of whether the Weibull distribution is appropriate. The degree to which the plotted data follow a straight line indicates the degree to which the data follow a Weibull distribution.

Just as in Sec. 6.7.1.2.3, where T has a two-parameter gamma distribution, Bayes estimation is complicated here by the multiple parameters. Martz and Waller (1991, Sec. 9.1) cite a number of early papers using various prior distributions. However the easiest Bayesian approach nowadays would be to assign

convenient diffuse priors to the parameters and use BUGS (1995), described in Sec. 8.2.3.3.3.

6.7.2 Model Validation

This section considers several topics. First, the usual investigations of the model assumptions are considered: whether subsets of the data all correspond to the same distribution, whether the distribution changes with time, and whether the times are serially correlated instead of statistically independent. In addition, the distribution may have been modeled by some parametric form, so the goodness of fit is investigated. Finally, if parameters have been estimated in a Bayesian way, the consistency of the data with the prior must be investigated.

The order described above follows the actual order of analysis. First the analyst would check to see what data subsets can be pooled and whether the usual assumptions seem to be satisfied. Only then would it be appropriate to try to fit some standard distribution to the data.

6.7.2.1 Poolability of Data Sources

To illustrate the methods here, this subsection will consider the three groups of data in Example 6.13, corresponding to three conditions of the plant during the LOSP event. As elsewhere in this chapter, graphical methods are considered first, and statistical tests second.

6.7.2.1.1 Graphical Methods

A simple graphical method of comparison is to overlay the EDFs for the different data subsets on a single graph. Then look to see if the EDF are intertwined, indicating that the subsets may be pooled, or if they are separated, shifted sideways from each other, indicating that the data subsets may not be pooled. This method is simple, but the graph can become very cluttered, especially if a moderate or large number of subsets must be compared. The same comment can be made for comparing separate histograms of the data subsets.

A graph that has come into fashion is the **box-and-whisker plot**, often simply called a **box plot**. The lower and upper edges of the box are the lower and

upper quartiles of the data. Thus, the box can be thought of as containing half the data, with 1/4 of the remaining data on each side. The median is marked somehow. The "whiskers" are two lines coming out of the box and going out to cover the range of most of the data. A few outlying points are plotted individually.

Figure 6.40 shows a box plot of the group T data from Example 6.14 generated using the STATISTICA (1995) software. The median is marked by a small square in the box. The software documentation does not seem to give precise definitions of "most of the data," or of the difference between an outlier and an extreme point. Also, this release of the software seems to have a small bug, in that the maximum (excluding outliers) is labeled as 11, when it should be 118.

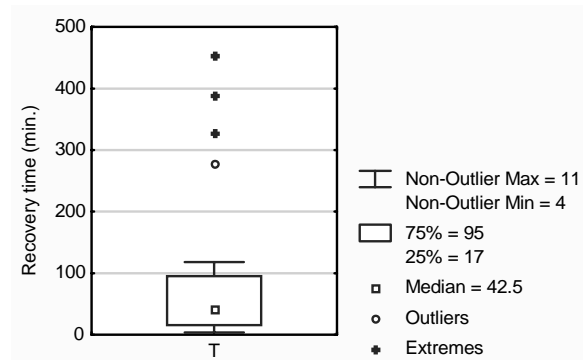


Figure 6.40 One form of a box plot. The box shows the lower and upper quartiles, with the median marked. The whiskers show most of the range, from 4 to 118, and individual outlying points are plotted.

Figure 6.41 shows the same box plot as drawn by a different software package, SAS/INSIGHT (1995). As before, the box shows the lower and upper quartiles, and the median is marked, this time with a stripe. The whiskers are restricted in length, going out to the extreme value if possible, but never more than 1.5 times the interquartile range. Any points beyond that distance are shown as individual dots.

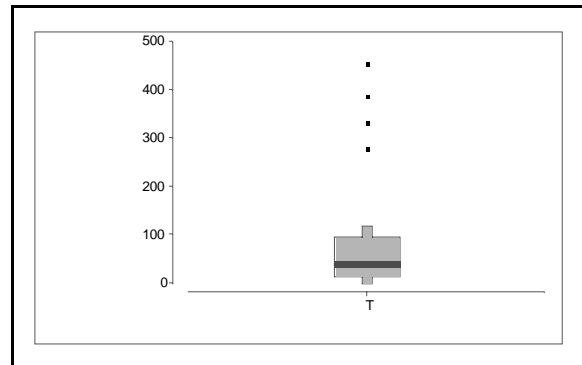


Figure 6.41 A different style box plot of the same data. The box shows the upper and lower quartiles, with the median indicated by a stripe. The whiskers show much of the range, with dots marking outliers.

Box plots were invented by Tukey (1977), and are still being modified according to individual taste. Any form of the plot that is produced by a convenient software package is probably adequate.

The example here is typical, in that the data are skewed, and the most obvious feature of the box plots given here is the long distance from the box to the largest value. Box plots are supposed to focus on the bulk of the data, with only moderate attention given to the extremes. Therefore, there are visual advantages to transforming skewed data by taking logarithms. Therefore, all the remaining box plots shown in this section will use $\log_{10}(\text{recovery time})$ instead of the raw times.

Figure 6.42 shows side-by-side box plots of the three data subsets in Example 6.13. Incidentally, the box plot of $\log(\text{time})$ is different from the box plot of time plotted on a logarithmic axis — the logarithms of large times tend not to be considered as outliers. This can be seen by comparing Figure 6.40 with the group-T portion of Figure 6.42.

6.

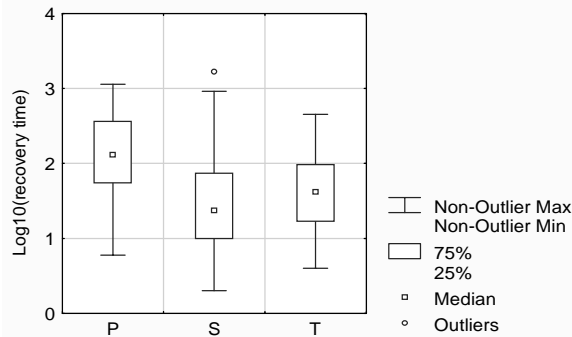


Figure 6.42 Side-by-side box plots of the three groups of data from Table 6.16, based on $\log_{10}(\text{recovery time})$.

Figure 6.42 shows that group P seems to have somewhat longer recovery times than the other groups. There seems to be little difference between groups S and T. Tests will be given below to investigate whether this visual impression is correct.

6.7.2.1.2 Statistical Tests

Tests Based on Normality. Warning: these tests are only valid if normality or lognormality can be assumed. If each data subset corresponds to a lognormal distribution, work with $Y = \log(T)$. Either natural logs or base-10 logs can be used, because $\log_{10}(T) = \ln(T)/\ln(10)$, so both are normally distributed if either is.

When Y has a normal distribution, standard tests based on normal theory can be used, as given in many statistic books. These tests investigate whether μ , the mean of Y , is the same in each data subset, under the assumption that the variances are the same. For added sophistication, tests of equality of the variances can also be performed.

- To compare the means of two data subsets, perform a Student's t test.
- To simultaneously compare the means of two or more data subsets, perform a one-way analysis of variance test.
- To compare the variances of two data subsets, perform an F test.
- To compare variances of two or more data subsets, use some version of a likelihood ratio test, such as Bartlett's test or a Pearson-Hartley test, as discussed by Bain and Engelhardt (1992, p. 426).

These tests are not considered further here, because they rely heavily on the assumption of normality. This is especially true of the tests later in the list. Most statistical software packages will happily perform these tests, no questions asked. The analyst must ask whether the assumption of normality is well enough established to justify the use of the tests.

Nonparametric Tests Based on Ranks. For general use when normality or lognormality is not well established, nonparametric tests are preferable. The books by Conover (1999) and Hollander and Wolfe (1999) are excellent summaries of standard tests. As before, let $Y = \log(T)$, but do not assume that Y has a normal distribution or any other particular distribution. Tests for location assume that various data subsets have distributions that are shifted sideways from each other. The shapes are the same, but the medians may be different. This is the nonparametric analogue of assuming that the distributions are normal with a common variance but possibly different means. Tests for dispersion assume that the shapes are the same, but possibly with different location and scale parameters. This is the nonparametric analogue of assuming normal distributions with possibly different means and variances.

To test equality of two medians against a shift alternative, use the Wilcoxon-Mann-Whitney test. This test was introduced in Sec. 6.3.3.2.2. In the present context, let W denote the sum of the ranks of times for the first data subset, when all the times are considered together. The ranks are the same whether or not the logarithmic transformation is performed.

For example, to compare group P to group S in Example 6.13, arrange all 70 times from the two groups in ascending order, and mark the times corresponding to group P. The smallest time from group P is 6 minutes. This has rank 12, because it is preceded by 11 values in group S from 2 to 5 minutes. The other ranks are found similarly. Ties are handled by assigning the average rank to all tied values. The rest of the test was explained in Section 6.3.3.2. It is not detailed here, because the test is normally performed by a computer.

To test whether two *or more* data subsets can be pooled, the test of choice is the Kruskal-Wallis test. It

tests whether the distribution of T is the same in all the data subsets, against the alternative that the distributions have the same shape but different medians. The test is based on a sum of ranks for each data subset. Those who want details can look in Conover (1999) or Hollander and Wolfe (1999); everyone else can just let the computer do the test.

When the Kruskal-Wallis test is applied to the data of Example 6.13, it rejects equality of the distributions with p -value 0.026. This is consistent with the graphical comparison in Figure 6.42 — clear evidence of a difference, though not extreme overwhelming evidence. Based on these analysis, Atwood et al. (1998) dropped group P from the analysis of durations, and combined groups S and T. Group P consists of LOSP durations when the plant remained at power throughout the event. The authors comment on reasons why plant personnel might be very deliberate in restoring offsite power while the plant is still up and running.

To test for equality of dispersion of two data subsets, the rank-like test of Moses is recommended. This requires splitting each data subset into two or more parts, and so is not suitable for very small data sets. See Hollander and Wolfe or documentation of a statistical software package for details of applying this test.

A well-known nonparametric test has not been developed for testing equality of dispersion of more than two data subsets. Therefore, graphical comparisons, such as side-by-side box plots, should be an important part of the analysis.

Nonparametric Test Based on EDFs. A well-known test for comparing two data subsets is the two-sample Kolmogorov-Smirnov test. It is based on comparing the empirical distribution functions for the two data sets. The test statistic is

$$D = \max_t [|\hat{F}(t) - \hat{G}(t)|]$$

where $\hat{F}(t)$ and $\hat{G}(t)$ are the empirical distribution functions from the two data sets. Many software packages can perform this test.

6.7.2.2 No Time Trend

This section will be illustrated by an extension of Example 6.13, taken directly from Atwood et al. (1998).

Based on the above type of analysis of Example 6.13, the LOSP study (Atwood et al. 1998) pooled the data from groups S and T, but excluded group P. That report also combined common-cause pairs of events at multiple units into single site-events (one pair of shutdown events, two pairs of trip events, and two pairs that involved a shutdown reactor and a reactor that tripped). This gave a total of 102 site events instead of the 107 in Example 6.13. They are sorted by event date and listed as Example 6.14. Times are in minutes, and dates are MM/DD/YY.

6.7.2.2.1 Graphical Methods

One natural way to examine the data for a trend is through a scatter plot of the observed values against calendar time. Often, as in Example 6.14, a few large values are outliers. They will determine the scale of the vertical axis. Compared to those large values most of the other values are very small, hugging the horizontal axis. In such a case, the observed values should be transformed, typically by taking logs.

Figure 6.43, from the LOSP study (Atwood et al. 1998), shows a plot of $\log_{10}(\text{recovery time})$, for the data of Example 6.14. Does this plot show a trend in time? Visually, any trend appears to be very slight. The section below, which considers statistical tests, will re-examine this example.

6.

Example 6.14 LOSP recovery times and event dates.

4	04/22/80	3.5	11/21/85	40	02/11/91
106	06/03/80	4	11/21/85	240	03/07/91
62	07/15/80	20	12/12/85	67	03/13/91
120	01/16/81	46	01/01/86	29	03/20/91
14	02/01/81	100	01/28/86	60	03/21/91
15	04/27/81	12	07/19/86	277	04/23/91
10	12/23/81	155	03/05/87	24	06/15/91
29	06/22/82	37	03/21/87	60	06/22/91
17	04/26/83	4	04/04/87	20	06/27/91
11	10/04/83	388	07/14/87	11	07/24/91
163	10/08/83	118	07/23/87	4	10/20/91
240	11/14/83	2	08/17/87	77	01/29/92
97	01/08/84	53	09/11/87	20	03/23/92
90	02/12/84	29	09/16/87	20	03/27/92
15	02/16/84	17	10/14/87	35	04/02/92
2	02/28/84	59	10/16/87	10	04/28/92
5	05/03/84	4	11/17/87	6	05/03/92
2	06/04/84	8	06/24/88	454	08/22/92
120	06/05/84	38	07/17/88	57	10/19/92
20	07/16/84	24	07/29/88	95	12/31/92
11	07/26/84	14	08/13/88	136	04/08/93
10	08/01/84	95	10/16/88	37	05/19/93
20	08/21/84	19	10/25/88	12	06/22/93
22	08/24/84	9	12/26/88	3	06/26/93
3	10/22/84	45	03/25/89	10	09/10/93
14	11/16/84	90	03/29/89	15	09/14/93
15	12/19/84	29	05/14/89	12.5	10/12/93
335	04/29/85	60	06/16/89	96	12/27/93
43	05/07/85	90	06/17/89	2	05/21/94
5	08/16/85	2	06/29/89	1675	11/18/94
73	08/28/85	45	01/16/90	132	02/27/95
25	10/03/85	14	02/26/90	917	10/21/95
13	10/07/85	140	03/20/90	127	01/20/96
60	10/22/85	37	07/09/90	330	02/06/96

A possibly more helpful plot is a cumulative plot of recovery time against chronological sequence. The vertical axis shows cumulative recovery time, that is, cumulative duration of LOSP events. No logarithmic transformation is made, because a sum of durations is easy to interpret, but a sum of $\log(\text{duration})$ is harder to interpret. Also, logarithms can be negative, so a cumulative plot of logarithms would not necessarily be monotone.

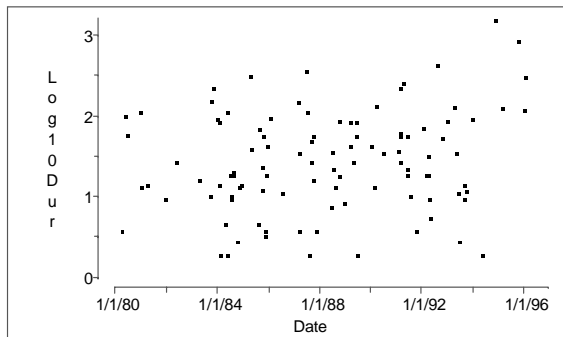


Figure 6.43 $\text{Log}_{10}(\text{recovery time})$ plotted against event date, for data from groups S and T in Example 6.14.

What should the horizontal axis show? If it shows event date, the slope of the curve represents average LOSP duration per calendar time. If, instead, the horizontal axis shows event sequence number, that is, the cumulative number of events, then the slope represents average LOSP duration per event. The latter is more meaningful in a study of durations.

Finally, a diagonal line, connecting the origin to the final point, provides a reference guide, so that the eye can better judge the straightness of the plot.

Figure 6.44 shows the cumulative duration plot for the data of Example 6.14.

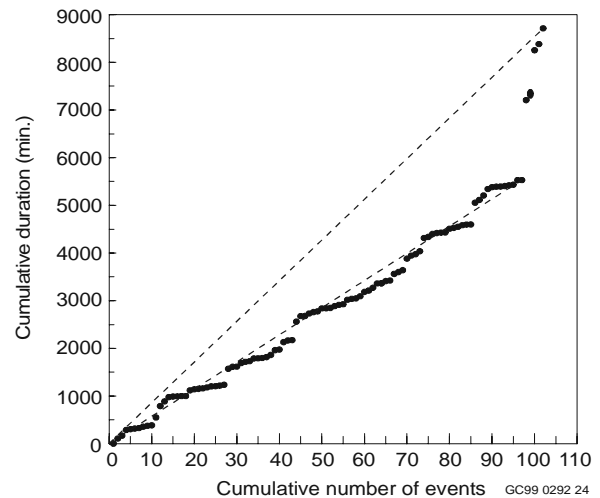


Figure 6.44 Cumulative duration of LOSP events versus cumulative number of events.

The cumulative plot clearly departs from the diagonal straight line, because of two large duration times near the right of the plot. The LOSP report mentions that one of those two times is conservatively large. The LER narrative states that recovery could have been performed earlier, but it does not give an estimated possible recovery time. The LOSP report used times when recovery would have been possible, when such times were available, but for this event the report was forced to use the actual recovery time.

In Figure 6.44, a second dashed line connects the origin (0, 0) to the 97th point, just before the first of the two large jumps. The cumulative plot stays close to this line until the large recovery times occur. Thus, any "trend" is the result, not of a gradual increase in recovery time, but of a couple of outlying

values, one of which is conservatively large. Figures 6.43 and 6.44 both reveal the two large recovery times. In this example, however, the cumulative plot seems more informative than the scatter plot, because the log-transformation in Figure 6.43 makes the large times appear less dramatic.

6.7.2.2.2 Statistical Tests

Test Based on Normality. Data from a scatter plot may be fitted with a straight line by least squares. Most software packages then test of the hypothesis that the slope is zero, assuming normally distributed scatter around the line.

The cited LOSP report fitted a straight line to the data in Figure 6.43 by least squares. The trend was reported as statistically significant at the 0.03 level.

This conclusion of a statistically significant trend seems surprising, based on the minimal apparent trend in the figure. The report authors did not have the insights given by the cumulative plot, but they critiqued the calculation in several ways.

- The calculation assumes that $\log(T)$ is normally distributed around the trend line. The lognormal distribution (without modeling a trend) was found to fit the data well, and the scatter plot appears consistent with normality. Therefore, the calculated p-value of 0.03 is apparently close to correct.
- The evidence for trend was very sensitive to the two values in the upper right of the figure. Dropping either value raised the p-value to 0.08. Further, one of those values was known to be conservatively high, as discussed above. This means that the trend may in part be an artifact of the data coding.
- The magnitude of the trend is small. A linear trend in the mean of $\log(T)$ corresponds to an exponential trend in the median of T . The magnitude of this trend is a factor of 3.6 over the 17 years of the study. This is fairly small from an engineering viewpoint.
- No solid engineering reasons were found to explain the trend.

Section 6.2.3.1.2 of this handbook discusses how test results should be interpreted. It states that calculation of a p-value is only part of the analysis, which should

be followed by some critical thinking. The above bulleted list of considerations illustrates that kind of thinking. Use of a cumulative plot would have helped the report authors even more, revealing that a smooth trend of any kind is inappropriate. The authors of the LOSP study chose not to model a trend, but recognized that additional data might change this decision.

Nonparametric Test. A test for trend that does not assume normality is easy to construct. Such a test is necessary if normality cannot be assumed. If normality can be assumed, the nonparametric test is less powerful for detecting a trend, because it ignores available information, that the data are normally distributed.

The test is the Wilcoxon-Mann-Whitney test, first introduced in Section 6.3.3.2.2. To apply it here, arrange the times sequentially, in order of their event dates. Count an event as *A* if it is *above* the median and as *B* if it is *below* the median. Discard any values that equal the median. Now carry out the Wilcoxon-Mann-Whitney test based on the ranks of the *As* in the sequence of all the events. Because this test is based only on comparisons to the median, it is the same whether or not logarithmic transformations are used.

When this was done with the data from Example 6.14, the median duration was 29. The first duration in Example 6.14 was a *B*, the next three were *A*, and so forth. In all, there were 48 *As* and 50 *Bs*. The *As* had higher average rank than the *Bs*, suggesting an upward trend, but the p-value was 0.09, not quite statistically significant. The nonparametric test is not as sensitive as the parametric test for detecting the small trend, in part because it does not make as much use of the two extreme values seen in Figure 6.44. If the normality assumption were not satisfied, only the nonparametric test would be valid.

6.7.2.3 Goodness of Fit to Parametric Models

One way to model recovery times and other durations is to model the distribution of the durations by some parametric distribution, such as lognormal, Weibull, etc. One must then check to see if the data fit this proposed model well. This section gives graphical methods and statistical tests for such investigations.

6.

6.7.2.3.1 Graphical Methods

The basic idea is to compare nonparametric estimates, which come directly from the data, with estimates based on the fitted model under consideration. For example:

- Compare the histogram to the density from the fitted model.
- Compare the EDF to the c.d.f. of the fitted parametric model. Equivalently, compare the empirical reliability function (1 minus the EDF) to the fitted reliability function.
- Compare the quantiles of the data to the quantiles of the fitted distribution. This plot is called a quantile-quantile plot, or a Q-Q plot. Q-Q plots have become very popular for assessing goodness of fit, although they take getting used to.

These three comparisons are illustrated below, using the data of Example 6.14, and an assumed lognormal distribution. First, the fitted distribution is found by taking natural logarithms of the recovery times, and estimating the mean and variance of their distribution. The estimated mean is 3.389 and the estimated standard deviation is 1.434. The $\ln(\text{time})$ values are modeled as normally distributed with this mean and variance. The raw times have the corresponding lognormal distribution.

Figure 6.45 shows the histogram density with a fitted lognormal density overlaid. Because this distribution is concentrated at small values, the goodness of fit is difficult to judge. Therefore the histogram of the $\ln(\text{time})$ values are also plotted, with a normal density overlaid, in Figure 6.46. Actually, the area under the histogram equals the number of observations, and the density has been rescaled to have the same area.

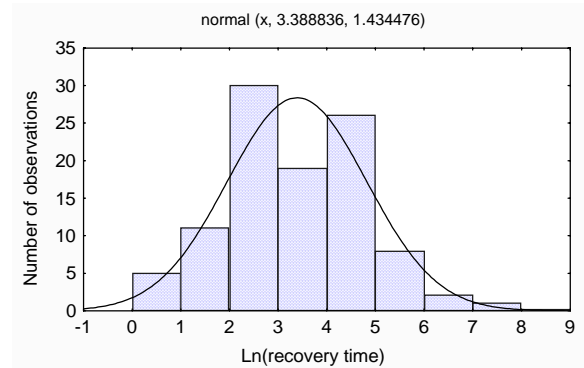


Figure 6.46 Histogram of data from Table 6.19, with a multiple of lognormal density overlaid. The skewness of a normal density overlaid. Fit appears as good as achievable without using a bimodal distribution.

Figure 6.47, from the LOSP report, shows a plot of the reliability function, 1 minus the EDF, with the corresponding fitted function, 1 minus the lognormal c.d.f. The plot in this form is useful for interpreting the degree to which the fitted c.d.f. differs from the empirical c.d.f., because the horizontal axis is in units of time. A plot in terms of $\log(\text{time})$ would not hug the axes so closely. Therefore, discrepancies between the curves would be more visible, but their magnitudes would be harder to interpret in real-world terms.

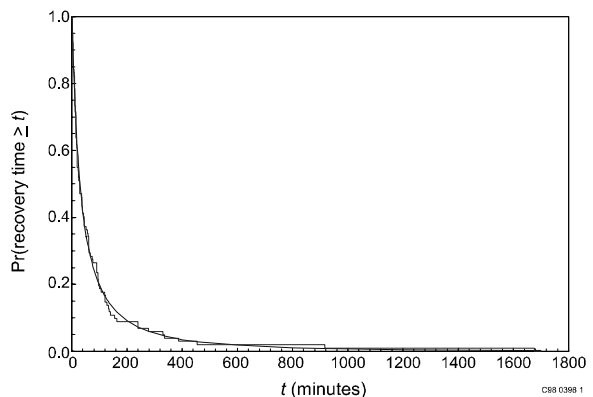


Figure 6.47 Empirical and theoretical reliability functions, where the reliability function is defined as 1 minus the c.d.f.

Finally, Figure 6.48 gives a Q-Q plot. If only one plot could be used, a Q-Q plot would be a strong contender for that one. Users of probability paper will recognize that a plot on probability paper is a form of a Q-Q plot. In Figure 6.48, the software package implemented the Q-Q plot by plotting the ordered values of $\ln(\text{time})$ against the theoretical expected

values of the corresponding order statistics. For example, denote $\ln(t)$ by y . In the implementation of this particular software package, the i th ordered value, $y_{(i)}$, is plotted against the expected value of $Z_{(i)}$, assuming that 102 values of Z are randomly sampled from a standard normal distribution.

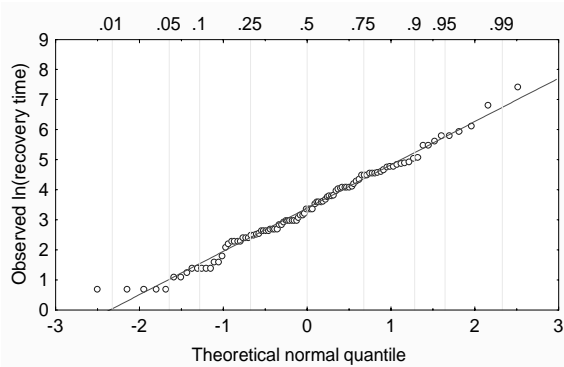


Figure 6.48 Quantile-quantile plot of $\ln(\text{recovery time})$ and fitted normal distribution. The points fall nearly on a straight line, indicating good fit.

The parameters, μ and σ , can be ignored in a Q-Q plot based on the normal distribution, because a normal random variable Y with mean μ and standard deviation σ is related to Z by $Y = \mu + \sigma Z$. This is a linear transformation, and so does not change the linearity or nonlinearity of the plot. In fact, it is not even necessary to obtain estimates of μ and σ . For other distributions, the parameters may need to be estimated before the Q-Q plot can be constructed.

The expected values of the order statistics cannot be constructed without tables or a computer program. Users of probability paper may construct a simpler version, plotting $y_{(i)}$ against the $i/(n+1)$ quantile of a standard normal distribution. Here n is the total number of observations, 102 in the present example. This simpler version gave its name to the plot, a quantile-quantile plot.

For the purpose of illustration, Figure 6.49 gives a Q-Q plot of the same example data, assuming that the raw recovery times have a *normal* distribution. Of course the fit is horrible — no one expects the raw times to have a normal distribution. This lack of fit is shown by strong curvature in the plot. The two largest times show the lack of fit most emphatically, but even without them the plot appears to show a curvature that indicates non-normality.

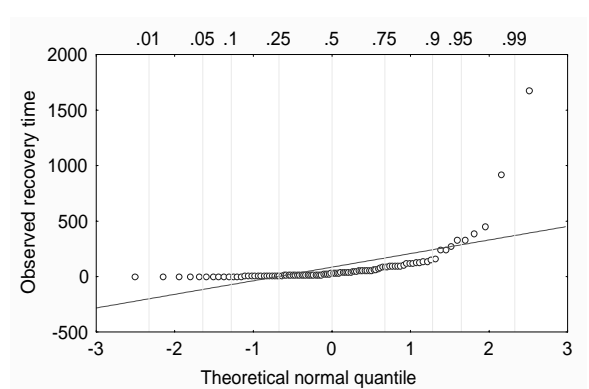


Figure 6.49 Quantile-quantile plot of raw recovery times against fitted normal distribution. The strong curvature indicates bad fit.

The particular form of the distribution can sometimes allow special tricks. Let us leave the present example, and consider investigating whether data t_1, \dots, t_n come from an exponential distribution. Example 6.6, which was deferred from Section 6.2.3.4, will be used to illustrate the method.

The idea of the Q-Q plot is that, when the data come from the assumed distribution, then

$$t_{(i)} \approx F^{-1}[i/(n+1)],$$

where F^{-1} is the inverse of the assumed c.d.f. It follows that

$$F(t_{(i)}) \approx i/(n+1).$$

A plot of these terms is called a **probability-probability plot** or **P-P plot**. We are considering the exponential distribution, so $F(t) = 1 - e^{-\lambda t}$, for some unknown λ . Therefore, when the data come from an exponential distribution, the above approximate equality becomes

$$1 - \exp(-\lambda t_{(i)}) \approx i/(n+1),$$

and therefore

$$-\lambda t_{(i)} \approx \ln[1 - i/(n+1)], \text{ or}$$

$$t_{(i)} \approx -\ln[1 - i/(n+1)]/\lambda.$$

6.

Thus, a plot of the ordered times against $-\ln[1 - i/(n+1)]$ should be approximately linear. The reason for the above mathematical gyrations is to obtain a plot that is linear, regardless of the value of λ . The linearity or nonlinearity of the plot does not depend on whether λ has been estimated well. Nonlinearity is evidence against the assumed exponential distribution.

Example 6.6 contains times between LOSP events, which should be exponentially distributed. A plot of the ordered times against $-\ln[1 - i/(n+1)]$ is shown in Figure 6.50. Because the plot does not show much

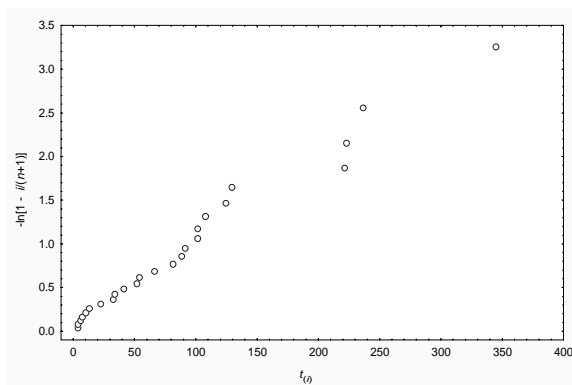


Figure 6.50 Plot for checking exponential distribution in Example 6.6.

curvature, it indicates good fit to the exponential distribution.

6.7.2.3.2 Statistical Tests

The tests in this section are called **goodness-of-fit tests**, because they are intended to test whether the data fit the assumed model well. The null hypothesis is that the data come from a distribution of the assumed form, for example from a lognormal distribution. The null hypothesis does not specify the parameters. Therefore, the null hypothesis includes a family of distributions. The alternative hypothesis is that the data come from some other distribution.

As always, remember that "acceptance" of the null hypothesis does not mean evidence that the null hypothesis is true. It merely means lack of evidence that the null hypothesis is false. For example, the data may be consistent with a lognormal distribution, and also

consistent with a gamma distribution and a Weibull distribution. In such a case, the analyst should not make assertions that are highly dependent on the form of the distribution. For example, a sample of 10 observations may be consistent with many possible distributions. An estimate of the 99.9th percentile of the distribution would be a large extrapolation from the actual data, highly dependent on the assumed form of the distribution. A confidence interval on this percentile would be even worse, because it would give an appearance of quantified precision, when in reality the distribution could have practically any form out in the tail.

Chi-Squared Test. The chi-squared test, seen in Sections 6.2 and 6.3, is also an all-purpose goodness-of-fit test. To apply it in the present context, estimate any unknown parameters of the hypothesized distribution of T . Based on these parameter estimates, divide the time axis into c bins of equal probability. The letter c stands for *cell*, another term for a bin in this context. Based on the recommendations of Moore (1986), choose the number of bins so that n/c is at least 1, and preferably at least 2. Let x_i be the observed number of values of T in the i th bin. Because the bins have equal probability, the expected number of values of T that will fall in any bin is n/c , the number of observations divided by the number of bins. The Pearson chi-squared statistic is

$$X^2 = \sum_j (x_j - e_j)^2 / e_j \quad ,$$

where each e_j equals n/c and each x_j is an observed count.

If the null hypothesis is true, the distribution of X^2 is approximately chi-squared. The commonly quoted rule is that the degrees of freedom is $c - 1 - p$, where p is the number of estimated parameters. For example, suppose the null hypothesis is that the distribution of T is lognormal, or equivalently, that $\ln(T)$ is normal. Then two parameters must be estimated, μ and σ . Thus, the commonly quoted rule for the degrees of freedom is $c - 3$. In fact, researchers have found that this is not quite correct, for subtle reasons described by Moore (1986, Section 3.2.2.1). The correct degrees of freedom are somewhere between $c - 1 - p$ and $c - 1$. The exact value depends on the form of the distribution in the null hypothesis.

Let us apply this to the data from Example 6.14, and use $Y = \ln(T)$ for convenience. Let H_0 be the hypothesis that Y is normally distributed. As mentioned above, the estimates of μ and σ are 3.389 and 1.434. With 102 observations, it is convenient to take 50 bins, so that each expected count is $102/50 = 2.04$. The bin boundaries are the 0.02, 0.04, ..., 0.98 quantiles of the distribution. These are estimated as

$$y_q = 3.389 + 1.434z_q,$$

where q is 0.02, 0.04, etc., and z_q is a quantile interpolated from a table of the standard normal distribution. For example, $z_{0.02} = -2.054$.

When this is carried out, using a computer to perform the calculations, the value of X^2 is 63.69. The distribution under H_0 is chi-squared with degrees of freedom between 47 and 49. Therefore, the p-value is between 0.053 and 0.077. The test almost rejects normality of $\ln(T)$ at the 0.05 level, in spite of the graphical evidence to the contrary!

Upon examination, the test is seen to be too powerful for its own good. It notices that the values tend to cluster, five occurrences of 2 minutes, six values of 20 minutes but no values of 21 minutes, etc. With 50 cells, each observed time is commonly the sole occupant of a cell. The test notices that the numbers have been rounded to convenient times, such as 20 minutes, and uses this as evidence against normality. In fact, such clustering is a departure from normality, and from any other continuous distribution. But it is not the kind of departure that is of interest to most analysts.

A coarser binning, into fewer cells, would not be distracted by fine clustering, and would search for more global departures from the null hypothesis.

We conclude this discussion of the chi-squared test by considering the exponential example that was deferred from Section 6.2.3.4.

Example 6.6 consists of 25 times. The null hypothesis is that the data come from an exponential distribution. The unknown λ is estimated as the number of events divided by the total observation period, $25/(2192 \text{ days}) = 0.0114$ events per day. This MLE is justified based on the Poisson count of events, as in Section 6.2.1.1. To obtain a moderate expected count in each bin, let us use ten bins. They have

equal estimated probabilities, 0.10 each, if they run from

0 days to $[-\ln(0.9)]/0.0114 = 9.24$ days
 9.24 days to $[-\ln(0.8)]/0.0114 = 19.57$ days
 ...
 201.89 days to infinity.

These calculations are all based on the exponential c.d.f., $F(t) = 1 - \exp(-\lambda t)$. Setting $F(t)$ to 0.1, 0.2, and so forth gives the bin boundaries.

There are four observed times in the first bin, two in the second, and so forth. The expected count in each bin is $25/10 = 2.5$. The calculated value of X^2 is 9.00. This must be compared with the percentiles of the chi-squared distribution. There are $c = 10$ bins, and $p = 1$ estimated parameter. Therefore, the degrees of freedom are between $10 - 1 = 9$ and $10 - 2 = 8$. The value 9.00 is in the middle of both of these distributions, the 56th percentile of one and the 66th percentile of the other. Therefore, the chi-squared test finds no evidence against the exponential distribution. This agrees with the earlier graphical analysis.

Shapiro-Wilk Test for Normality. Many software packages offer the Shapiro-Wilk test for normality. It is based on seeing how closely the order statistics follow theoretical normal values, as displayed for example in Figure 6.48. For testing the normal distribution, the Shapiro-Wilk test is one of the most powerful tests against a wide variety of alternatives. Details are not given here, because all the calculations are carried out by the computer.

With the logarithms of the data of Example 6.14, the Shapiro-Wilk test does not reject normality of $\ln(T)$, giving a p-value of 0.34. This agrees with the visual evidence of Figure 6.45.

Tests Based on the EDF. Several families of tests have been proposed based on the empirical distribution function. The idea is to reject the null hypothesis if the EDF is not "close to" the theoretical c.d.f. Closeness can be measured in various ways, giving rise to a variety of tests. EDF-based tests are appealing because they do not require a choice of bins, but simply use the data as they come.

6.

The most famous such test is the Kolmogorov test, also known as the Kolmogorov-Smirnov test. It is described in Appendix B.3.4. It rejects H_0 if

$$\max | \hat{F}(t) - F(t) |$$

is large, where the maximum is over all values of t . Here, any unknown parameters in F must be estimated; the effect of this estimation is typically ignored.

When SAS (SAS Version 8, 2000) performs the Kolmogorov test of lognormality on the times in Example 6.14, it gives a p-value > 0.15 . That is, it does not calculate the exact p-value, but it does report that the departure from lognormality is not statistically significant.

The Cramér-von Mises test and the Anderson-Darling test are other EDF-based tests, designed to remedy perceived weaknesses in the Kolmogorov test. The Cramér-von Mises test is based on

$$\int [\hat{F}(t) - F(t)]^2 f(t) dt .$$

Here, F is the distribution that is assumed under the null hypothesis, and f is the corresponding density. Thus, the Kolmogorov test looks at the maximum difference between \hat{F} and F , while the Cramér-von Mises test looks at an average squared difference. The Anderson-Darling test is based on

$$\int \{ [\hat{F}(t) - F(t)]^2 / \{ F(t)[1 - F(t)] \} \} dt .$$

This division by $F(t)[1 - F(t)]$ gives greater weight to the tails of the distribution, where departures from F might most likely occur. Thus, this test is intended to be more powerful than the Cramér-von Mises test against common alternative hypotheses. Many statistical packages perform one or more of these tests.

When testing lognormality of the data in Example 6.14, SAS reports a p-value of > 0.25 for the Cramér-von Mises test and also for the Anderson-Darling test. Just as for the Kolmogorov test, SAS does not compute the exact p-value, but it reports that the departure from lognormality is not statistically significant.

6.7.2.4 Consistency of Data with Prior, in Bayesian Parametric Estimation

The issue here is whether the data are consistent with an assumed *informative* prior distribution for the unknown parameters. If a noninformative prior distribution is used, then the question does not arise, because the noninformative distribution is supposed to be consistent with anything.

6.7.2.4.1 Exponential Durations

A quantitative approach is possible when T has an exponential(λ) distribution. In this case all the information of interest about λ is contained in Σt_i , as shown in Section 6.7.1.2.2. Therefore, we can compare Σt_i to what would be expected based on prior belief about λ .

If Σt_i is surprisingly large or surprisingly small, that is, if Σt_i is in either tail of the distribution of ΣT_i , then the prior distribution is questionable. The value Σt_i is in the lower tail if $\Pr(\Sigma T_i < \Sigma t_i)$ is a small probability, and in the upper tail if $\Pr(\Sigma T_i > \Sigma t_i)$ is a small. To be specific, consider the lower tail. The relevant probability is

$$\Pr(\Sigma T_i < \Sigma t_i) = \int \Pr(\Sigma T_i < \Sigma t_i | \lambda) f_{\text{prior}}(\lambda) d\lambda . \quad (6.21)$$

The inner conditional probability can be evaluated by using the fact that the distribution of ΣT_i , given λ , is gamma(n, λ). If the prior distribution of λ is not conjugate, the integral in Equation 6.21 must be evaluated numerically, just as in Sections 6.2.3.5 and 6.3.3.4: either (a) compute the integral using numerical integration, or (b) generate a random sample of λ values from the prior distribution, find $\Pr(\Sigma T_i < \Sigma t_i | \lambda)$ for each such λ , and find the average of these probabilities as the overall probability.

Treatment of the upper tail follows the same pattern.

If the distribution of λ is conjugate, that is, gamma(α, β), for some prior parameters α and β , then Equation 6.21 simplifies. By working out the integrals it can be shown that $\Sigma T_i / (\Sigma T_i + \beta)$ has a beta(n, α) distribution. Equivalently, $\beta / (\Sigma T_i + \beta)$ has a beta(α, n) distribution. These are marginal distributions corresponding to Equation 6.21, from which λ has been integrated out. Therefore, if $\Sigma t_i / (\Sigma t_i + \beta)$ is in either extreme tail of a

beta(n, α) distribution, or equivalently, if $\beta/(\sum t_i + \beta)$ is in either extreme tail of a beta(α, n) distribution, then the gamma(α, β) prior distribution is questioned.

In Example 6.13, suppose that the group of S (shut-down) events are the only ones of interest. Suppose also that the times are assumed to be exponential(λ) – the realism of that assumption is not the subject of the present investigation. Finally, suppose that λ is assigned a gamma(2, 30) prior distribution, roughly equivalent to two prior observed times with total duration of 30 minutes. The shape parameter of only 2 means that the prior is not very informative, so we expect the data to be consistent with it, unless 30 minutes is very unrealistic.

From Table 6.15, we find $n = 62$ and the total of the durations is $62 \times 92.3 = 5722.6$. The beta tables in Appendix C assume that the first beta parameter is smaller than the second, so it is convenient to work with the beta(2, 62) distribution rather than the beta(62, 2) distribution. Therefore, we ask if

$$30/(5722.6 + 30) = 5.2E-3$$

is in either tail of a beta(2, 62) distribution. Table C.5 shows that the 5th percentile of the beta(2, 62) distribution is roughly $6E-3$ (it is an interpolation of $7.01E-3$ and $3.53E-3$ in the table). Table C.6 shows that the 2.5th percentile is roughly $4E-3$. So the observed value is somewhere between the 2.5th and 5th percentiles of the predictive distribution. This means that the prior may need rethinking. It should either be modified or it should be justified more carefully. (In the present example the prior came out of thin air, but the real difficulty is that the durations are not really exponential – the whole exercise is only for illustration.)

6.7.2.4.2 Distributions Having Two or More Parameters

When this topic – comparing the data to the prior – arose in connection with estimating λ or p , there was a single parameter of interest, and a single observed random variable that contained all the information of interest for that parameter. This random variable was the total count of initiating events, the count of failures on demand, or in the previous section the total duration.

However, the present subsection considers a distribution with (at least) two parameters, such as μ and σ or

α and β . No single random variable contains all the information of interest. Therefore, in such cases it is simplest to compare the data with the prior by constructing

1. a prior credible region for the two parameters, or a pair of prior credible intervals, and
2. a posterior credible region *based on noninformative priors*, or alternatively a pair of confidence intervals.

The first case shows what the prior distribution says, and the second case shows what the data say. Compare the answers from 1 and 2 to see if the prior distribution and the data seem consistent.

6.7.3 Nonparametric Density Estimation

The most prevalent methods of estimating a density function are parametric methods. As described in Section 6.7.1.2, the density is specified in terms of a functional form, such as lognormal or Weibull, with unknown parameters. The parameters are then estimated from the data. However, there also exist nonparametric methods for estimation of a density function, some of which are described here.

The simplest and best known method of estimating a density function is to construct a frequency table, and then to plot the histogram. This method was discussed in Sec. 6.7.1.1.4. Two illustrations are given there, Figures 6.38 and 6.39. Both use the 45 recovery times from part T of Example 6.13. The methods discussed below are illustrated with the same set of 45 recovery times.

6.7.3.1 Smoothing Techniques and Kernel Estimators

Smoothing techniques can be motivated by recalling that the density function, $f(t)$, is the derivative of the c.d.f., $F(t)$. The EDF, discussed in Section 6.7.1.1.3 and denoted by $\hat{F}(t)$, is a natural estimator of $F(t)$. Thus, a natural estimator of the density is the differential quotient using the EDF in place of the c.d.f.,

$$\hat{f}_n(t) = \frac{\hat{F}(t+h) - \hat{F}(t-h)}{2h} \quad (6.22)$$

6.

where h is an increment of the variable t . The main problem in applying such an estimator is to choose h small enough so that the differential quotient adequately approximates the derivative, but large enough so that the interval with limits $t \pm h$ contains a sufficient amount of data.

Recall that $\hat{F}(t)$ equals the number of observations having a value less than or equal to t divided by the total number of observations, n . Therefore, Equation 6.22 can also be written as

$$\hat{f}_n(t) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{t-t_i}{h}\right) \quad (6.23)$$

where K is a function defined as $K(u) = 1/2$ if u is between ± 1 , and zero otherwise, and t_i is the i th observation. Notice that an observation t_i only enters into this calculation if $(t_i - t)/h$ is between ± 1 , or in other words if t_i is near t ; specifically if t_i is within h units of t . Thus, the estimate is based on averaging values of $1/2$ when observations are near t . This is a special case of a general type of estimator known as a **kernel density estimator**. The function $K(u)$ is called the kernel and the increment h is called the **bandwidth**. The bandwidth defines a "window" centered at t and having width $2h$, which contains the data involved in the estimate at the point t .

6.7.3.1.1 The Rectangular Kernel

When graphed, the kernel corresponding to Equation 6.23 is a rectangle of height $1/2$ and width $2h$. The resulting estimator is illustrated here with group T of Example 6.13 and two bandwidths.

Figure 6.51 shows a graph of the estimate of the density when the bandwidth is $h = 25$ minutes.

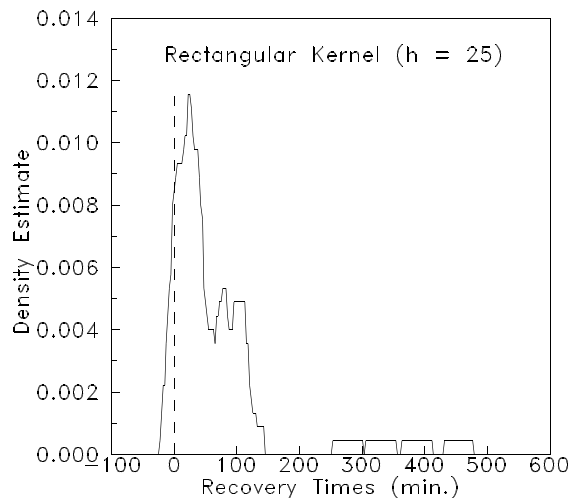


Figure 6.51 Density estimate of the data from group T in Example 6.13, with rectangular kernel and bandwidth 25.

Notice that the estimated density is zero in the interval roughly from 150 to 250 minutes. This corresponds to the fourth and fifth bins of the histogram of Figure 6.38, both of which were empty.

It is also evident that the graph is somewhat jagged, indicating that the bandwidth may be too small so that not enough data are being captured in the window.

The vertical dashed line marks the point $t = 0$, to be discussed later.

Consider now a rectangular kernel estimate with the same data but with a larger bandwidth, $h = 50$ minutes. The results are shown in Figure 6.52.

There is still some jaggedness, but it is somewhat less than in Figure 6.51. There is still a noticeable low point in the vicinity of 200 minutes, but it is narrower than in Figure 6.51.

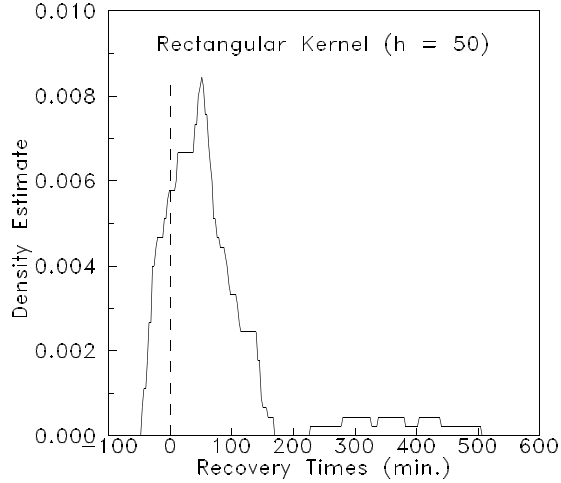


Figure 6.52 Density estimate of the data from group T in Example 6.13 with rectangular kernel and bandwidth 50.

It is clear that by smoothing over a very wide window any features can be smoothed out. For this reason, it is desirable to give some thought to whether there is some explanation of low density. In other words, are these real effects or are they just due to randomness? If the low estimates can be explained by something other than random fluctuation, smoothing would tend to cover it up, but if they are due to randomness, then smoothing should be helpful.

This issue was also seen with histograms. Choosing too narrow bins for the size of the data set caused the shape to be influenced too much by random variation. Choosing too wide bins smoothed out nearly all the variation. The question of how much to smooth and how much roughness to allow is inherent in all forms of density estimation.

6.7.3.1.2 Boundary Problems

Notice that as the bandwidth is increased the interval over which the estimated density is positive becomes wider. This is because the window is picking up more data as it gets wider. This causes the anomaly that the estimated density is positive over negative values of the t axis, even though t represents a positive variable, namely recovery time. The vertical dashed line marks the point $t = 0$ in each figure, and the portion of the

density to the left of this line is substantial. In addition, although many values of t_i are close to zero, the density estimate *decreases* as t moves leftward to zero. Various methods have been suggested for correcting the estimate at the boundary of the possible region.

Silverman (1986) gives a method that is very easy to implement. If the density is allowed to be positive only for $t \geq 0$, augment the data by reflecting it around 0. That is, create a new data set that consists of

$$\{ \dots, -t_2, -t_1, t_1, t_2, \dots \} .$$

Estimate the density based on this data set. Call this estimate $\tilde{f}(t)$. The integral from $-\infty$ to ∞ of $\tilde{f}(t)$ is 1.0, because \tilde{f} is a density. Also, if the kernel is a symmetrical function then \tilde{f} is symmetrical around

6.

zero, that is, $\tilde{f}(-t) = \tilde{f}(t)$. Now define the real density estimate by

$$\hat{f}(t) = 0 \quad \text{for } t < 0$$

$$\hat{f}(t) = 2\tilde{f}(t) \quad \text{for } t \geq 0.$$

Then \hat{f} is a density that is zero for negative t and nonnegative for positive t . It estimates the unknown true density.

Figure 6.53 shows the resulting estimate with the data of this section, when the kernel is rectangular and the bandwidth $h = 50$. This estimate can be compared with Figure 6.52.

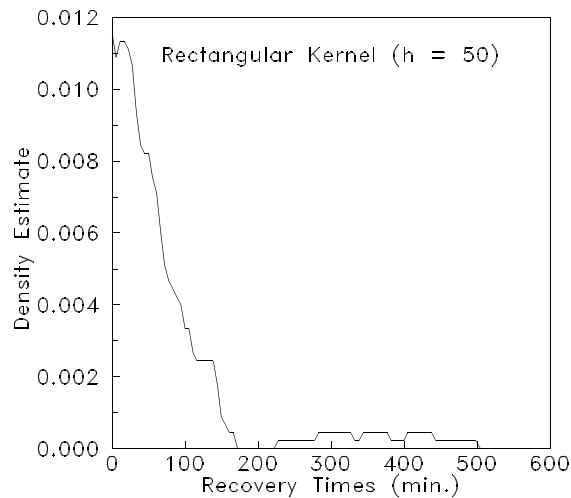


Figure 6.53 Density estimate from group T of Example 6.13, with rectangular kernel and bandwidth 50, forced to be nonzero on positive axis only.

For large t , this estimate is very similar to that of Figure 6.52. However, it is quite different for t near zero. The density is not plotted for $t < 0$, but it equals zero there.

The simple method just given forces the density estimate to have slope zero at the boundary. Those who want to allow a density estimate with nonzero slope at the boundary can see Hart (1997, Sec. 2.5). Technically, Hart's book deals with smoothing a scatter plot, but the method given there can be adapted as follows to smoothing a density estimate: construct a rough histogram density estimate, place a dot at the top

of each histogram bar (including the bars with height zero!), and treat those dots as a scatter plot.

6.7.3.1.3 The Triangular Kernel

It may also be desirable in some cases to give less weight to the data in the extremes of the window and to produce a smoother graph. This can be accomplished by choosing a different function for the kernel. A very simple one which does this is the function $K(u) = 1 - |u|$ if u is between ± 1 , and zero otherwise. The graph of $K(u)$ is an isosceles triangle with base two units in width. This kernel gives more weight to the data in the middle of the window and less to data at the sides of the window. It is also possible, by choosing a kernel function with a smoother graph, to produce a kernel estimate which is also smoother. The normal kernel, given next, is such a smooth kernel.

6.7.3.1.4 The Standard Normal Kernel

A kernel function that is often used is the **standard normal kernel**, equal to the standard normal p.d.f., which is given in Appendix A.7.2. Figure 6.54 shows the density estimate for the same recovery time data, but using the standard normal kernel and bandwidth 25. The density has been made positive on the positive time axis only, by the technique of Section 6.7.3.1.2.

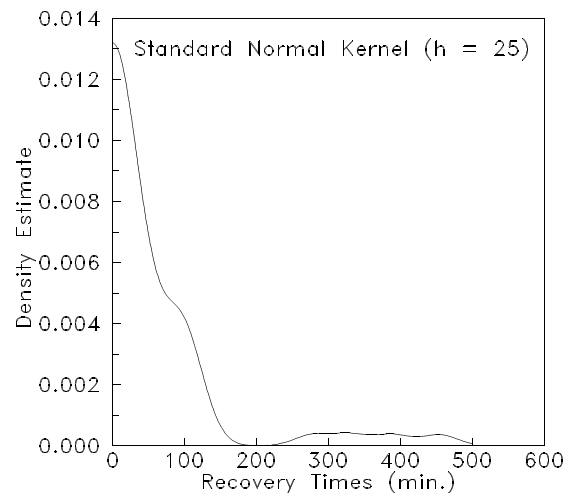


Figure 6.54 Density estimate of the data from group T in Example 6.13, with standard normal kernel and bandwidth 25.

The resulting plot is clearly much smoother than the ones obtained using the rectangular kernel. The increased smoothness is provided by the standard normal kernel, which is differentiable everywhere. The low estimate of density near 200 is still present, but the low spot does not drop to zero as it did in Figure 6.53. This is because the standard normal kernel is always positive valued. Even though this kernel gives less weight to data which are farther from the center of the kernel, it makes use of every observation in the data set. Consequently, with the standard normal kernel, all terms in the density estimate of Equation 6.23 are positive, although the extreme ones will tend to be relatively small.

For the sake of comparison, Figure 6.55 shows the standard normal kernel estimates for bandwidth $h = 50$.

Although the graphs shown in Figures 6.54 and 6.55 retain some general features of the graphs in Figures 6.51 through 6.53, they are somewhat smoother. As mentioned in the case of the rectangular kernel in Section 6.7.3.1.1, this type of smoothing is desirable if the sparsity of data in these intervals is due to randomness, but possibly not if there is an explanation for the sparseness.

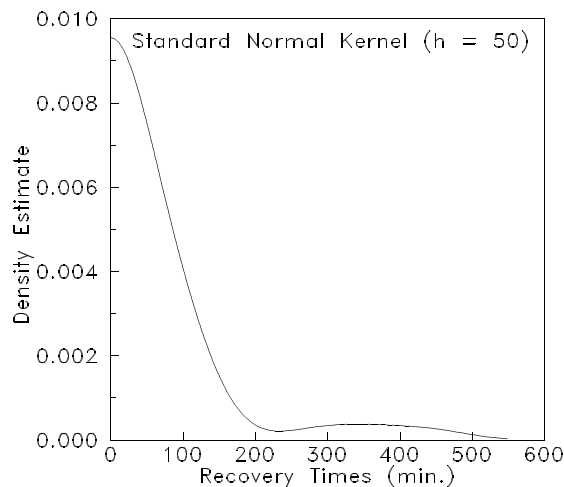


Figure 6.55 Density estimate of the data from group T in Example 6.13, with standard normal kernel and bandwidth 50.

6.7.3.2 Choosing the Bandwidth

General guidelines for choosing a kernel and bandwidth are difficult to formulate. The choice of a bandwidth always involves a trade-off between bias and variability. An attempt to reduce bias generally requires a small bandwidth, but this tends to result in a large variance. On the other hand, choosing a large bandwidth will reduce the variance, but at the expense of increasing the bias. A criterion which accounts for both the bias and variance is based on a quantity called the **mean squared error**, $MSE = \text{mean squared difference between the unknown parameter and its estimator}$. It is easy to show that

$$MSE = (\text{bias})^2 + \text{variance of estimator}$$

so that as the MSE approaches zero, both the bias and the variance of the estimator approach zero.

A reasonable choice of bandwidth should take into account the amount of data, and so the solution must depend on n . Thus, we consider a sequence, $h = h(n)$. The sequence should converge to zero, but not too fast or too slowly. It is known, for example, that under certain fairly modest assumptions, a desirable form for the bandwidth is

$$h(n) = cn^{-1/5}.$$

The main problem is that calculation of the constant c requires more than is typically known about the p.d.f. to be estimated, and it also depends on the choice of a kernel. For example, according to p. 45 of Silverman (1986), for the standard normal kernel and assuming the distribution of the data to be normal with standard deviation σ , the bandwidth which minimizes the integrated MSE asymptotically is

$$h(n) = 1.06\sigma n^{-1/5}.$$

Notice that the constant c in this case requires that the standard deviation be known or at least estimated.

For example, with the recovery time data the sample standard deviation, which is given in Table 6.17, is 99.9 minutes. If this is used to estimate σ , then the optimal bandwidth is $h(n) = 105.9n^{-1/5}$. Using the sample size $n = 45$ yields $h = 49.5$. This choice is very nearly the bandwidth of 50 minutes that was

6.

used in Figure 6.55.

Keep in mind that this choice of bandwidth was derived for the case where both the distribution being estimated and the kernel are normal, so the result would be good with these assumptions. However, this might be a good place to start if trial and error is used to determine what bandwidth to use. In other words, if it is not clear what to assume, then it would be best to try a few different bandwidths and choose one which provides some smoothing, but does not obscure basic features. As Silverman says, "There is a case for undersmoothing somewhat; the reader can do further smoothing 'by eye' but cannot easily unsmooth."

Another problem that often occurs in practice is that the data will be plentiful in some parts of the range, but

sparse in others. This is typical with data from highly skewed distributions. For example, with a positively skewed distribution, such as any of the distributions in Sec. 6.7.1.2, there will tend to be more data in the lower end than in the upper tail. This would suggest the desirability of having a bandwidth that varies with t , so that a shorter increment can be used in the lower end where the data points are more abundant, and a larger increment used in the upper tail where there are not as many points. This idea is not developed here, but such methods exist. For additional reading on this topic see the discussions of the nearest neighbor method and the variable kernel method in Silverman (1986).