College of Engineering

Statistics and Uncertainty

Patrick L. Gurian
Drexel University
pgurian@drexel.edu

Goal

This chapter provides a review of concepts from probability and statistics that are useful for risk assessment. It begins with a review of probability density distribution functions, then covers how these functions are used as models for variability and uncertainty in risk assessment, describes how these functions are fit to it particular cases by estimating parameters, and describes one method, bootstrapping, for quantifying uncertainty in these parameter estimates.

Probability

A probability density distribution function (PDF) describes the probability that some randomly varying quantity, such as the amount of water consumed by an individual, the carcinogenic potency of a chemical toxin, or the concentration of a pollutant in the air, will lie in a particular interval. The PDF is defined as the function that when integrated between limits A and B. gives the probability that the random variable x will fall between those limits A and B. Thus

$$\text{Prob}[A<x<B] = \int_A^B f(x)\,dx$$

where f(x) denotes the PDF.

In typical risk assessments a fairly limited number of functional forms of f(x) are used. For example, the normal distribution is a PDF with the following functional form:

$$f(x) = 1/\{\sigma(2\pi)^{1/2}\}\ \exp\{(x-\mu)^2/2\sigma^2\}$$

where μ and sigma in this equation are parameters, or constants that can be tuned to fit particular applications. For a normal distribution the parameter μ corresponds to the mean and the parameter σ to the standard deviation. By choosing different values of these parameters, the same functional form (normal distribution) can be used to describe many different random variables which have different means and different standard deviations. Figure 1 shows the probability density distribution for a standard normal distribution, which is the normal distribution with mean of 0 and standard deviation of 1. A common notation is to use ~ to denote "is distributed as" and then write an abbreviation for the class of distribution with information on the parameters of the distribution in parentheses. For example, if Z is a random variable that follows a standard normal distribution, this can be written as:

$$Z \sim N(0, 1)$$

where N is a standard abbreviation for a normal distribution and by convention the first number in parentheses is the mean, and the second is the variance (standard deviation squared).

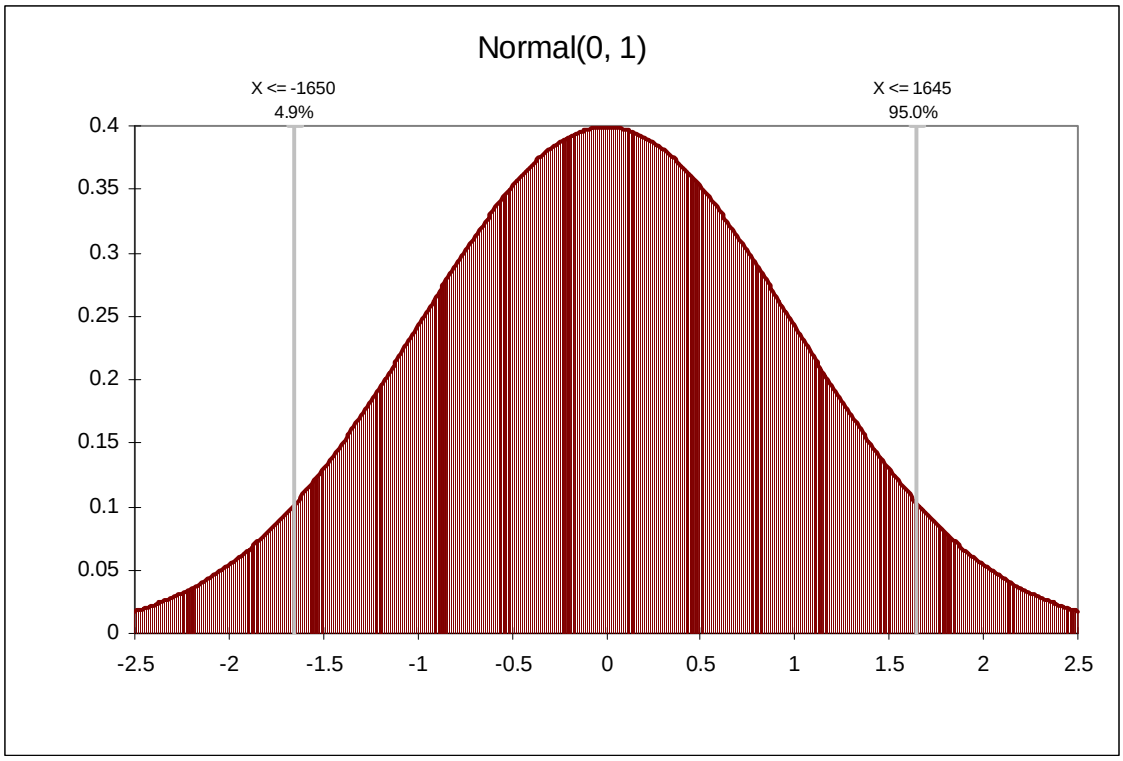

Figure 1. The probability density distribution function for a standard normal variable.

It is often useful to evaluate the probability that a random variable is less than a particular value. For example one might be interested in the probability that a particular risk is below a given regulatory benchmark. This is called a cumulative distribution function (CDF) and is found by integrating the PDF from negative infinity to the particular value, X:

$$F(X) = \text{Prob}[x<X] = \int_{-\infty}^{X} f(x)\,dx$$

where F(X) denotes the CDF. CDF values are probabilities and range from 0 to 1. They are often multiplied by 100 to give percentiles, the percentage change that a random variable is below a specified value. Figure 2 shows the CDF of a standard normal distribution.

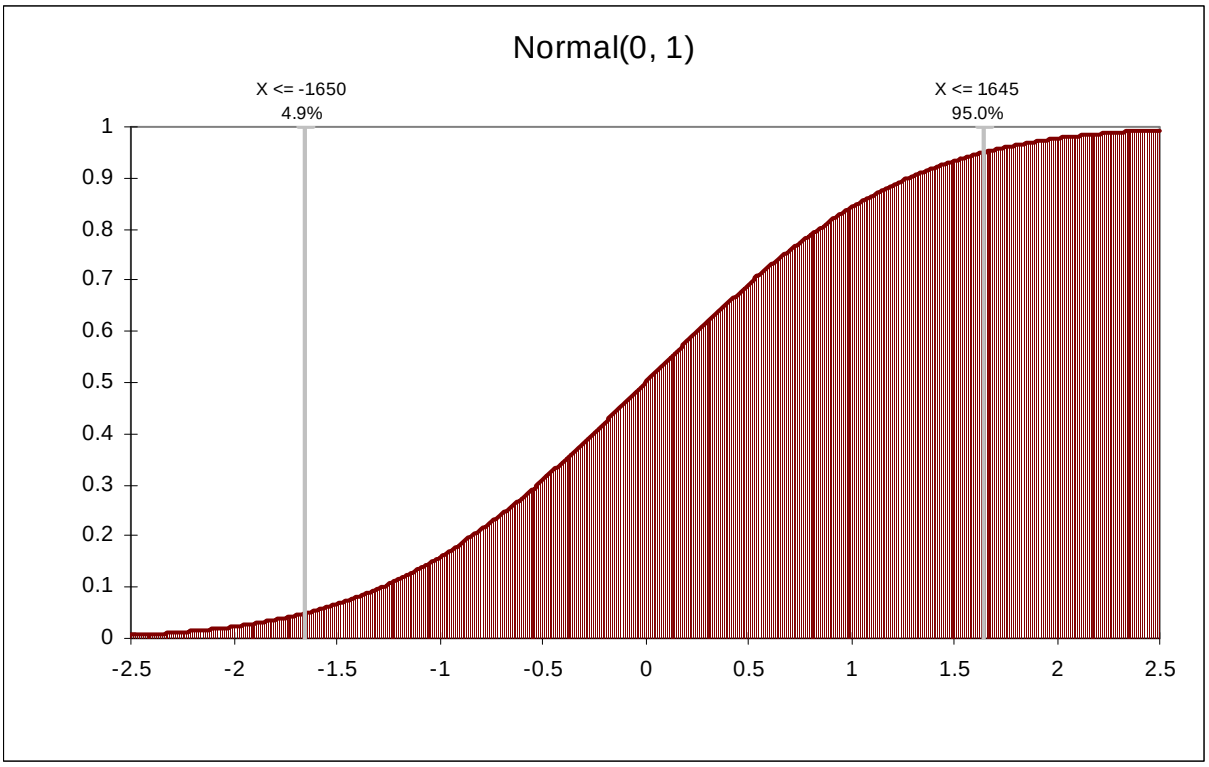

Figure 2. The cumulative distribution function for a standard normal distribution.

Evaluating the CDF of a normal distribution require numerical integration. To avoid having to carry out this integration for each of the infinite number of normal distributions, one makes use of the fact that the following transformation will convert any normally distributed random variable, denoted by x, to a standard normal variable, denoted by Z:

$Z=(x - \mu)/\sigma$

Note that this transformation does not change the order of different values of x. Thus the highest value of x will correspond to the highest value of Z, the median value of x will correspond to the median value of Z, the 10^{th} percentile value of x will correspond to the 10^{th} percentile of Z, etc. Thus if one knows the CDF for a standard normal distribution, one can transform X to the corresponding value of Z, evaluate the standard normal CDF at Z and this equals the CDF value of X. To facilitate this approach, CDF values for a standard normal distribution are widely available in standard reference tables. The transformation described above can then be applied to determine the CDF for any of the infinite number of normal distributions.

Example. Finding the CDF of a normal distribution
Suppose annual repair costs for a particular car follow a normal distribution with a mean of 300 and standard deviation of 100:

Annual repairs ~ N (300, 100^2)

and we wish to find the probability that repairs will exceed $450 in a given year. The first step is to find the Z value corresponding to the value of $450:

$Z = (X-\mu)/\sigma$

$Z = (450-300)/100 = 1.5$

The next step is to find the CDF value of Z in a standard table found in nearly every introductory statistics textbook:

$F(Z) = F(1.5) = 0.933$

Note, however, that the CDF is the probability of a random variable being less than a given value. To find the probability of repair being less than $450, we make use of the fact that the probability of an event and its complement (defined as the event not happening) add to one. Thus

Prob [repair<450] = prob[Z<1.5] = 1- prob[Z>1.5] = 1 - 0.933

Prob [repair >450] = 0.067

Parameter Estimates

In the example above the probability distribution was specified. The mean and standard deviation of the repair costs were assumed to be known perfectly. It is common not to know the parameters of a distribution but instead have a data and wish to "tune" the parameters of a PDF to fit the particular data. This process of fitting parameter values to match observations is referred to as parameter estimation. One natural approach might be to observe repair costs for a number of cars and calculate the arithmetic mean and standard deviation of these costs. Then set the mean of the model distribution equal to the observed mean and the standard deviation equal to the observed standard deviation. This is an example of a technique known formally as the method of moments. While the approach is straightforward in this case, it is difficult to generalize to more complicated statistical models, such as the dose-response models used in risk assessment. The emphasis here will instead by on maximum likelihood estimation, because this method is very generally applicable.

Maximum likelihood estimation begins with the question "how likely is the data we observed?" For a single observation, x, by definition this is f(x). Typically data are multiple independent observations. The likelihood of those observations occurring together is then the product of the likelihoods of the individual observations:

$L=f(x_1)\ f(x_2)\ f(x_3)\ldots f(x_N)$

where the subscripts indicate the individual observations, N is the number of observations, and L is referred to as the likelihood function. For example, if a baseball team is observed to win one game and lose the next two then:

L = Prob[win] prob[loss] prob[loss]

If we write the probability of winning as p, then this can be written:

$L = p(1-p)^2$

since winning and losing are compliments. Suppose one person posits that the team has a long-run frequency of winning of p=0.5. In this case the likelihood of the observed data is:

$L = (0.5)\ 0.5^2 = 0.125$

If a second person states that the value is only p=0.3, which value do we prefer? One way to assess this is to examine the probability of getting the results we actually observed under these two alternative views of p. If p=0.3 then the likelihood of the sequence of wins and losses that was observed is:

$L = 0.3\ (0.7^2) = 0.147$

The probability of the outcome that actually occurred is higher given the second person's estimate of p than given the first person's estimate of p. Based on this we generally prefer the second person's estimate of p, as it is more consistent with the observed data. The next step is to ask if there is another estimate of p which gives an even higher likelihood of observing the data. Ultimately one seeks the value of p which maximizes the probability of observing the data. Calculus provides a method for doing this. One first differentiates the likelihood function, L, with respect to the parameter, p:

$L = p(1-p)^2$

$dL/dp = (1-p)^2 - 2p(1-p)$

To find a critical point one sets dL/dp=0

$0 = (1-p)^2 - 2p(1-p)$

Now factor out (1-p)

$0 = 1-p -2p$

Rearrange to:

$3p=1$

And solve for p:

$p=1/3$

Thus p=1/3, the observed proportion of wins, is the maximum likelihood estimate of p. (To verify that this is a maximum one can note that the second derivative is negative for this value of

p.) In this case it was possible to find the maximum value of L analytically. In many cases with more complicated likelihood functions, it is not possible to find L analytically. In these cases numerical search algorithms are used to identify a maximum.

For very large data sets, one can imagine that the joint probability of all the observations will be quite low (i.e., L is a product of many numbers each of which is ≤ 1 since all probabilities are ≤ 1). It is often easier for these numerical search algorithms to work with the log of the likelihood, rather than the likelihood:

$$\text{Ln } L = \ln \pi f(xi|\theta)$$

where θ indicates the parameters associated with a particular f(x) and indicates a product. The log of a product can be expressed as the sum of the logs:

$$\text{Ln } L = \Sigma \ln f(xi|\theta)$$

and it is this form that is customarily used in numerical optimization routines.

Variability and Uncertainty

Variability refers to differences in outcomes obtained from a process. Uncertainty is lack of knowledge. Probability theory was originally developed to describe variability in the outcomes of repeated events, that is, the long-run frequency of different events. Those who desire to restrict the use of probably to describing objectively measurable variability are referred to as *frequentists*. Others who view probability more broadly as the subjective assessment of the likelihood of an event are termed *subjectivists*. This subjectivist viewpoint allows probability distributions to be used to describe not only variability but also uncertainty. Uncertainty can result from variability. For example, I may not know the outcome of a coin flip because if varies between heads and tails. However, uncertainty can result from many other sources, such as lack of understanding of the fundamental process at work. The subjectivist view of probability allows for the use of probability to describe one's belief as to the value of a quantity that has not yet been observed (i.e., for which there is no frequency information). For example, one might use probability to describe factors such as the sensitivity of the earth's climate system to a doubling of pre-industrial CO_2, even though this is not strictly speaking a randomly varying quantity. There is one value which is unknown to us.

In many cases both variability and uncertainty are present. For example, a particular drinking water may be contaminated with *Cryptosporidium* oocysts. The amount of oocysts in different water samples will vary. If the oocysts move randomly in the water, that is they move independently, neither clustering together nor dispersion from each other, then the probability that x oocysts will be found in a given water sample can be described by a Poisson distribution:

$$\text{Prob [x oocysts]} = \lambda^x \exp(-\lambda)/x!$$

where λ is a parameter. One property of this distribution is that the mean is equal to λ. One can think of λ as the long-run mean, that is, the mean if an infinite number of samples from the same Poisson distribution are averaged. Because a finite sample is not guaranteed to be perfectly

representative of the population from which it is drawn, even after observing the mean (median, variance, etc.) of a sample, there is still uncertainty as to the population mean (median, variance, etc.). It is uncertainty due to this sampling variability which is quantified through statistical methods.

It is common to first assess an appropriate functional form for f(x) and then, given the chosen form of f(x), assess the uncertainty in the parameters of f(x). In the classical statistical framework, these parameters are fixed values. Based on a sample, estimates are obtained for these parameters. However, given sampling variability, these parameter estimates have variability in them. This variability can be assessed and the standard deviation of the estimate of the parameter quantified. This standard deviation of a parameter estimate is termed a standard error. Thus uncertainty in model parameters (due to sampling variability and only sampling variability) is captured by these standard errors. In the Bayesian framework, model parameters are treated as themselves being random variables. In this framework, the λ value for a Poisson process would follow a probability distribution with a mean (reflecting the central tendency of λ) and a variance (reflecting uncertainty as to the true value of λ). In the classical framework, it is not quite correct to describe the parameters of models as following uncertainty distributions. Instead the standard errors are used to include or exclude different possible values of the model parameters with various levels of confidence. It is not clear that any harm is done in simply treating model parameters as random variables with means equal to their estimates and standard deviations equal to their standard errors. This view essentially adopts the Bayesian framework even for models estimated in a classical framework. It is often a convenient approach to adopt in risk assessments where probability is used quite generally to describe a wide variety of uncertainties.

Bootstrapping

As described above, sampling variability leads to uncertainty in parameter estimates. In some cases an analytical formula is available to estimate the standard error of a parameter. For complicated model forms, such formulae may not be available. Bootstrapping is a generally applicable method to assess the uncertainty in parameter estimates due to sampling variability. The concept underlying bootstrapping is to treat the sample obtained as the PDF of the model. Each point has probability 1/N where N is the number of observations in the sample. This assumption allows us to create alternative samples from the data. One randomly draws N observations from the observed dataset to create an alternate data set. These observations are drawn with replacement, meaning that if a particular observation is not eliminated from subsequent draws after it is sampled. It is this that allows for each alternative dataset to be slightly different from the original dataset, since some observations will appear multiple times in the alternative dataset and some will not be sampled at all.

The next step is to estimate the desired parameter (or summary statistic, such as mean, variance, 90th percentile, etc.) for each of the alternative datasets. Each combination of generating an alternative dataset and estimating the quantity of interest is termed an iteration of the bootstrap procedure. The values for each dataset are then considered discrete samples from the probability distribution of the parameter. Means, variances, and percentiles for the parameter estimate can be found from this discrete sample. Thus the standard error of the parameter estimate can be estimated as the standard deviation of the different parameter estimates obtained from each

iteration. This is essentially a Monte Carlo approach to assessing parameter uncertainty and, as with all Monte Carlo analyses, the number of iterations conducted should be quite large, preferably as large as 10,000 (Morgan and Henrion 1990, Burmaster and Anderson 1994). A smaller number of iterations may be acceptable if convergence of estimates can be observed. In this case one would track the estimate of interest across different iterations. Values will fluctuate greatly at first as the estimate is based on a small number of iterations, but these fluctuations will decrease as a larger sample size is obtained. Convergence is achieved when the estimate no longer shows fluctuations large enough to be of concern to the analysis.

References/Suggested Reading

Burmaster DE, Anderson PD. (1994) "Principles of Good Practice for the Use of Monte-carlo Techniques in Human Health and Ecological Risk Assessments," *Risk Analysis, 14* (4): 477-481.

Morgan, M.G., and Henrion, M. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press.