Fitting Distributions to Dose Data

Daniel Stancescu, Ph.D.

May 1, 2014

Fit statistics

A common problem in statistical analysis is fitting a probability distribution to a set of observations for a variable. The principle behind fitting distributions to data is to find the type of distribution (Normal, Lognormal, Weibull, etc.) and the value of the parameters (mean, standard deviation, etc.) that give the highest probability of producing the observed data. Usually, we do not know that the data came from any specific type of distribution, though we can often guess at some good possible candidates by matching the nature of the variable to the theory on which the probability distributions are based. For example, the Normal distribution is a good candidate if the random variation of the variable under consideration is driven by a large number of random factors in an additive fashion, whereas the Lognormal is a good candidate if a large number of factors influence the value of the variable in a multiplicative way. It is relatively rare that we are convinced a variable should be represented by one specific type of distribution. Thus, one usually tries to fit several types of distributions to the data set and then compare how well they fit the data. A visual comparison is usually a good start, though one should keep in mind that the data pattern for small data sets, will not usually look like the same pattern one would see if the dataset was large. It is also important to consider whether the properties of the fitted distribution, particularly the range and any skewness, are appropriate. However, we usually have a number of candidate distributions to choose from, which is where a statistical comparison of their fits comes into play. For radiation dose data, we know that this type of data usually follows a Lognormal, Normal, or a Weibull distribution.

Fit statistics are used for two related, but distinct purposes: model selection, and fit validation. Model selection is the process of picking one particular fitted distribution type over another, while fit validation is the process of determining if a particular fitted distribution is a good fit for the data.

There are three classical goodness-of-fit statistics that are very commonly used in most of the distribution fitting software: Chi-Squared, Kolmogorov-Smirnoff, and Anderson-Darling. These goodness-of-fit statistics were originally developed as tests for fit validation, and were not directly meant to be used as tools for deciding between alternate distributions. They are used to measure how well a distribution fits the input data and how confident we are that the data was produced by the distribution function. Here are some of the reasons these goodness of fit statistics are all technically inappropriate as a method of comparing fits of distributions to data:

- The Chi-Squared statistic depends on specifying the number of histogram classes into which the data will be grouped, and there is no 'golden rule' that gives the correct number to use. It also makes some pretty strong assumptions that only come close to being valid when one has a very large data set.
- The Kolmogorov-Smirnoff and Anderson-Darling statistics were designed to test the goodness of fit of distributions with defined parameter values, not those where the parameters are estimated from the data. Corrections are possible for only a very few types of distributions, so the fitting software usually use a generic correction for the other distribution types which can be very rough.
- None of these goodness-of-fit statistics penalize distributions for the number of parameters they use. Thus, a distribution with three parameters may well fit the data better because it has a lot more flexibility in shape than a two-parameter distribution, but the apparent improvement is spurious, which can lead to the problem of over-fitting.
- None of these goodness-of-fit statistics can correctly handle truncated, censored or binned data.
- The method of fitting, usually the Maximum Likelihood Method, or the Method of Moments is inconsistent with the measurement of degree of fit.
- None of these goodness-of-fit statistics give a proper statistical weighting to the plausibility of each candidate distribution.

In addition to the goodness-of-fit statistics described above, there are some more modern approaches that were specifically developed for model selection. These tests, called 'Information Criteria' are better suited for this task since they take into account, among other things, the number of parameters of the fitted distribution, and so they fit distributions according to the principle of parsimony, which states that the best model is the simplest model that is consistent with all the available data.

There are several 'Information Criteria' tests, the most common used being the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). The AIC statistic is the least strict in penalizing loss of degree of freedom, while the BIC statistic is somewhat stricter in penalizing loss of degree of freedom by having more parameters.

The Information Criteria have none of the problems described above for the previous goodness-of-fit statistics:

- They are based on calculating the log likelihood of the fitted distribution producing the set of observations. This means that one can use Maximum Likelihood as the fitting method and be consistent with the goodness of fit statistic.
- The Information Criteria penalize distributions with greater number of parameters, and thus help avoid the over-fitting problem.
- Since the basis of these statistics is the log-likelihood, Information Criteria can be used with truncated, censored and binned data.

The AIC and BIC statistics are calculated from the log-likelihood function by the following expressions:

$$AIC = -2\ln(L_{max}) + k(2n/(n-k-1))$$

$$BIC = -2\ln(L_{max}) + k\ln(n)$$

where:

- $n =$ number of data values;
- $k =$ number of parameters to be estimated (e.g. the normal distribution has $k = 2$, corresponding to the mean and standard deviation);
- $L_{Max} =$ the maximized value of the log-likelihood function for the estimated distribution (i.e. fit the parameters by the Maximum Likelihood Method, and compute the log of the likelihood).

The AIC and BIC are clearly very similar. Both the AIC and the BIC statistics have two terms in their equations, the first one which measures the deviance of the model fit (or the model lack of fit), while the second term is a penalty term for the additional parameters in the model. Therefore, as the number of parameters k increases, the lack of fit term decreases while the penalty term increases. Conversely, as the number of parameters k decreases, the lack of fit term increases while the penalty term decreases. The model with the smallest AIC value is deemed the 'best' model since it minimizes the difference from the given model to the underlying model for the data. It is important to note, however, that the AIC and BIC statistics do not provide a measure of the absolute goodness of a particular fit. That is, the actual values of the AIC and BIC statistics do not have meaning, except in relative terms, when you compare one proposed distribution type to another.

The theoretical underpinnings of both AIC and BIC statistics rely on Bayesian analysis and the two different forms come from different assumptions for the Bayesian 'priors'. The AIC tends to penalize the number of parameters less strongly than the BIC. There is a lot of discussion in the literature about which one is more appropriate, and the jury appears to still be out. Our recommendation was to use the AIC statistic, since is the most common and most well-known of all the Information Criteria statistics.

There are situations when just based on the visual comparison of the raw data and the distributions fitted, it is very hard to say which distribution fits the data better. In Figure 1, the raw data indicates that the underlying CLL dose conversion factor distribution lies between a Lognormal distribution with 2 parameters, and a Weibull distribution with 3 parameters. In this particular case, the AIC value corresponding to fitting the Weibull-3 distribution is -163,749, and the AIC value corresponding to fitting the Lognormal-2 distribution is -163,261. So, the distribution with the lowest AIC value, i.e. the Weibull distribution with 3 parameters is considered the better fit between these two distributions. While the visual comparison in this case might show that either one of these distributions will provide an appropriate fit to the data, the AIC statistic allows us to choose from these two competing distributions, based on clear defined criteria.

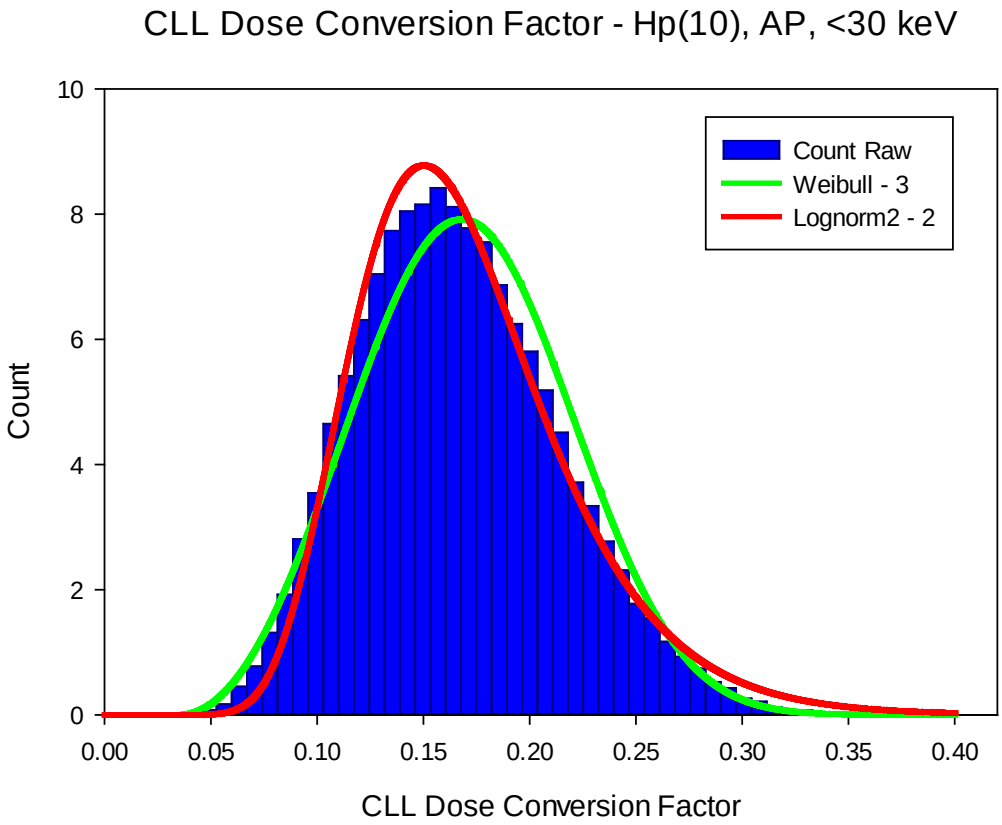

Figure 1: Comparison of two distributions for the CLL Dose Conversion Factor data.

Weibull distribution

The Weibull distribution has either two parameters (shape and scale), or three parameters (shape, scale, and location). The Weibull distribution is very flexible; actually, it consists of a family of distributions that can assume the properties of several distributions. When the shape parameter is 1, the Weibull distribution is identical to the exponential distribution. When the shape parameter is less than 1, the Weibull distribution becomes a steeply declining curve. When the shape parameter is equal to 2, a special form of the Weibull distribution, called the Rayleigh distribution, results. When the shape parameter is set to 3.25, the Weibull distribution approximates the shape of the Normal distribution. Figure 2 shows how the Weibull distribution changes when the shape parameter increases from 1.5 to 5, and the scale is held constant at 1. Figure 3 shows how the Weibull distribution changes when the shape parameter increase from 1 to 4, and the scale is held constant at 2. The Weibull distribution has a range that is bounded at the lower end, which can be in certain situations a great advantage compared to the Normal distribution, and have a density function that gradually approaches to zero at the upper end. While it is obvious that a Weibull distribution can closely approximate the shape of a Normal or a Lognormal distribution, it is also pretty clear that it can also model data that

might not be coming from either one of these distributions, but it's something that it's close to both the Normal and Lognormal distributions.

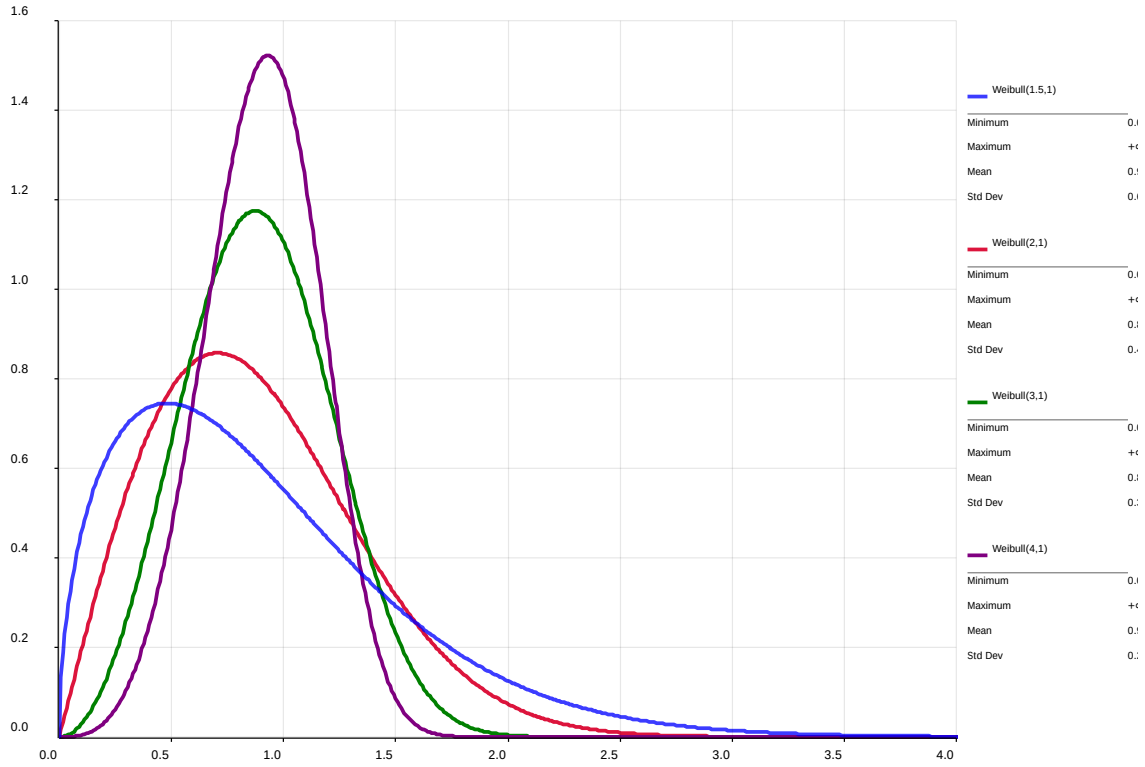

Figure 2: Weibull distributions for different shape parameters, and a fixed scale parameter.

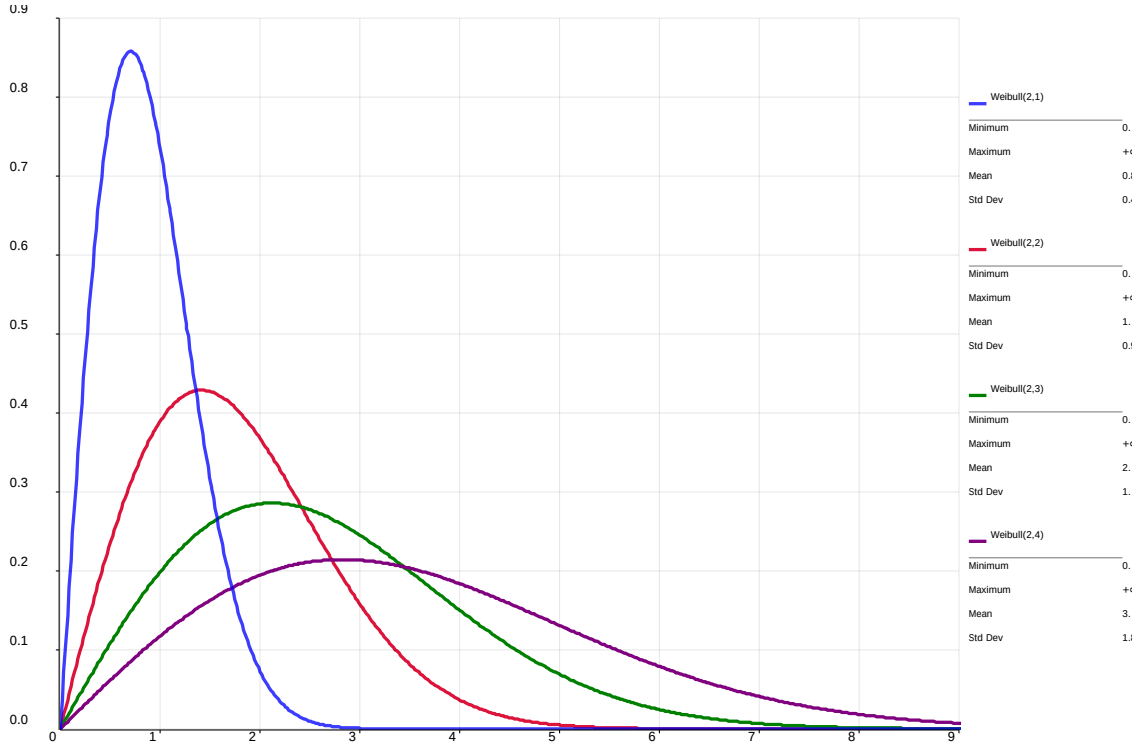

Figure 3: Weibull distributions for different scale parameters, and a fixed shape parameter.

The Weibull distribution was recently added as one of the input distribution for entering the radiation doses in the IREP v5.7 version. While the Weibull distribution was used for a long time to model radiation doses (Darby 1982, Ogundare 2003, Datta 2007), it was the introduction of the CLL cancer risk model in the IREP v.5.7 version that contributed to the addition of the Weibull distribution as one of input distributions for entering the radiation doses in IREP. The reason for this is that some of the distributions for the number of lymphocytes and B-lymphocytes for several organs were modeled using Weibull distributions (Apostoaei 2012), and as a result of using the probabilistic model developed for the CLL cancer risk model, some of the dose distributions that were obtained from this probabilistic model were better fitted by a Weibull distribution, rather than the distributions available in IREP in the previous version.

Figure 4 shows that a three-parameter Weibull distribution has the lowest AIC and is selected as the best fit to the CLL dose distribution. The dose to the CLL precursors is entered into IREP as a three-parameter Weibull with its associated shape, scale, and shift parameters.

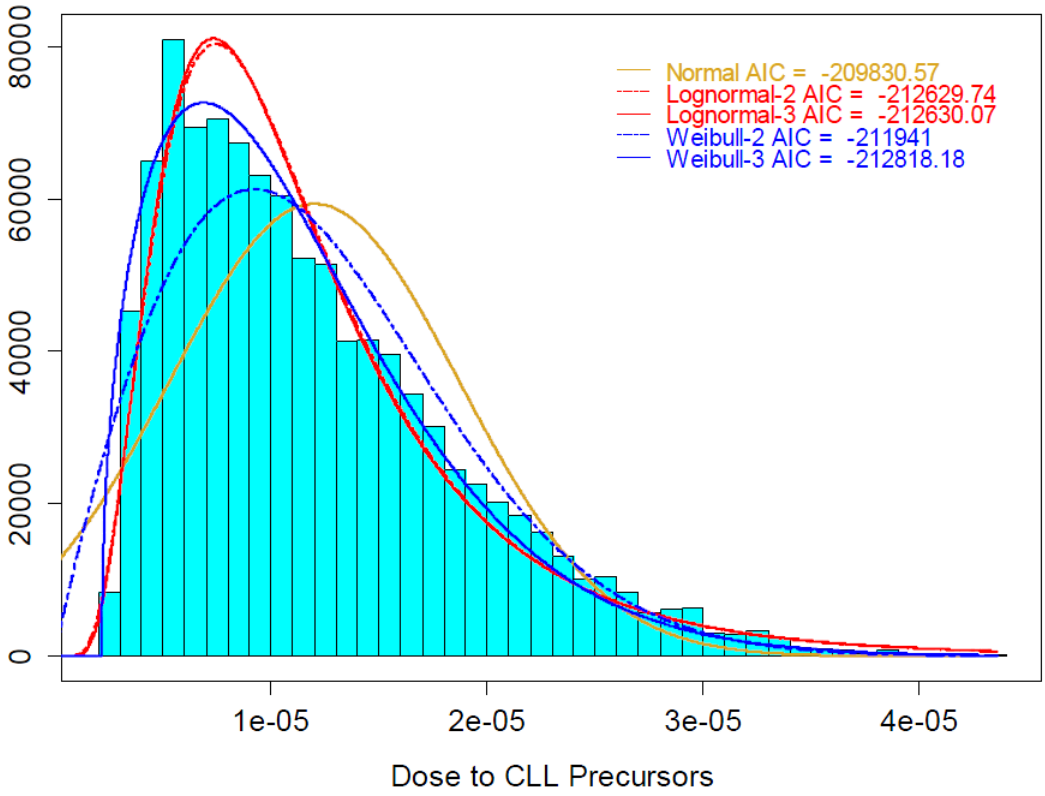

Figure 4: Comparison of the fits of five probability distributions to the distribution of CLL doses.

References

Darby, S.C., Kendall, G.M., Greeslade, E. (1982): "Patterns of dose incurred by workers on the national radiological protection board's dose record keeping service. II Individual Dosemeter Assessments", J. Society Radiological Protection, Vol. 2 (4), pg. 31–38.

Ogundare, F. O., Balogun, F. A. (2003): "Analysis of occupational doses of workers on the dose registry of the Federal radiation protection service in 2000 and 2001", Radiation Protection Dosimetry, Vol. 103(1), pg. 57–62.

Datta D., Singh S., Johnson B. E., Kushwaha H. S. (2008): "Maximum likelihood estimates of mean and variance of occupation radiation doses subjected to minimum detection levels", Radiation Protection Dosimetry, Vol. 129(4), pg. 411–418.

Apostoaei, A. I., Trabalka, J. R. (2012): "Review, synthesis, and application of information on the human lymphatic system to radiation dosimetry for chronic lymphocytic leukemia", http://www.cdc.gov/niosh/OCAS/pdfs/irep/raddoscll.pdf