

Fitting distributions to data

and why you are probably doing it wrong

By David Vose

All colored illustrations are taken from ModelRisk outputs
For more information on the ModelRisk software, visit www.vosesoftware.com

A common problem in risk analysis is fitting a probability distribution to a set of observations for a variable. One does this to be able to make forecasts about the future. The most common situation is to fit a distribution to a single variable (like the lifetime of a mechanical or electrical component), but problems also sometimes require the fitting of a multivariate distribution: for example, if one wishes to predict the weight and height of a random person, or the simultaneous change in price of two stocks.

There are a number of software tools on the market that will fit distributions to a data set, and most risk analysis tools incorporate a component that will do this. Unfortunately, the methods they use to measure the goodness of fit are wrong and very limited in the types of data that they can use. This paper explains why, and describes a method that is both correct and sufficiently flexible to handle any type of data set.

Fitting a single distribution

The principle behind fitting distributions to data is to find the type of distribution (normal, lognormal, gamma, beta, etc) and the value of the parameters (mean, variance, etc) that give the highest probability of producing the observed data. For example, Figure 1 shows the normal distribution with parameters that best fit a particular data set. The data were randomly generated from a Normal distribution with mean and standard deviation of 4 and 1 respectively. The data set consists of 1026 values, which is many more than one usually has to work with, so the parameter estimates (4.026 and 1.038) are close to the true values.

Usually, of course, 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. The normal distribution, for example, is a good candidate if the random variation of the variable under consideration is driven by a large number of random factors (none of which dominate) 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.

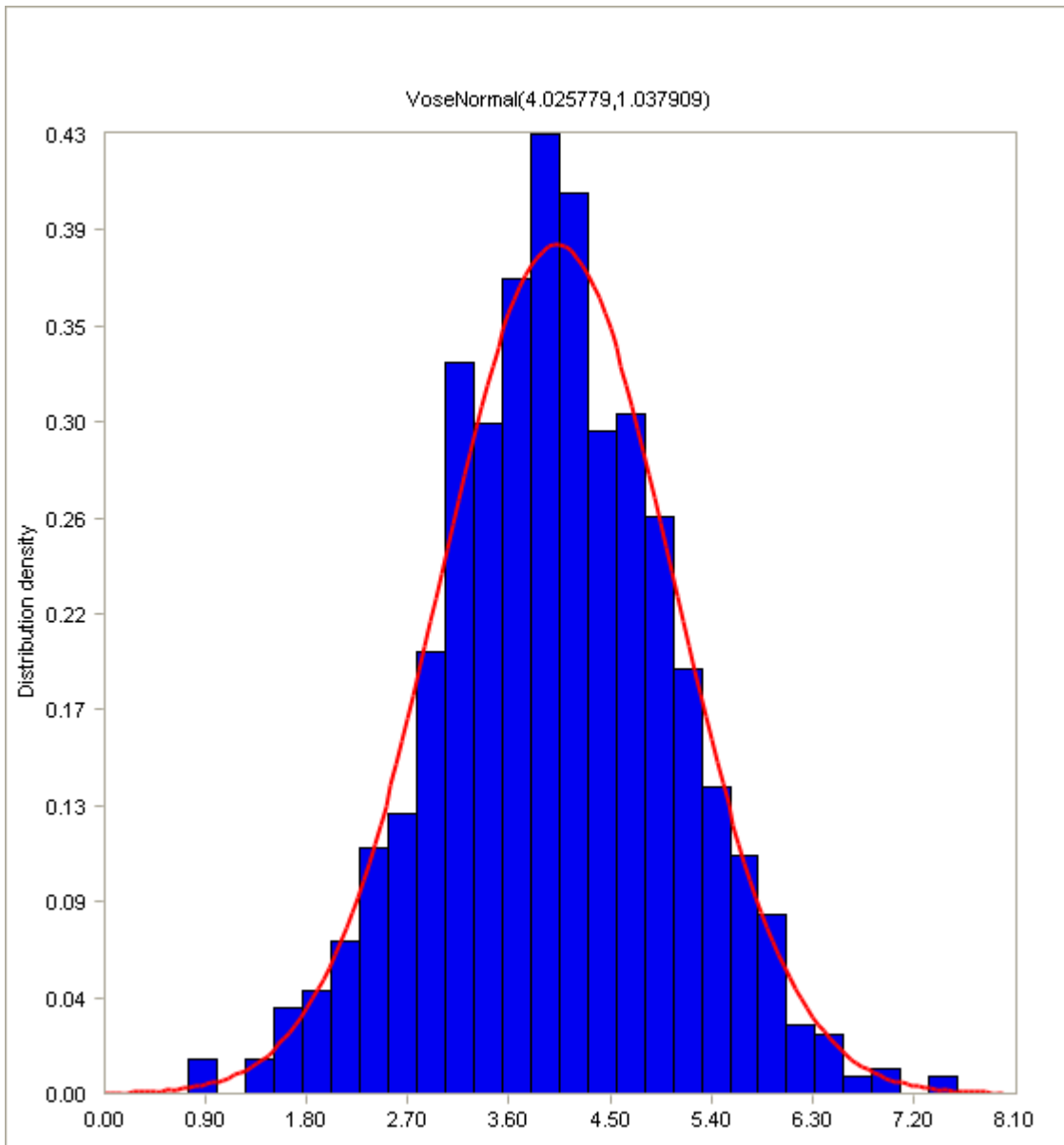


Figure 1: Normal distribution fit to a set of data. The graph compares the probability density of the fitted distribution with a histogram of the data, normalised so that they have the same area.

A number of other graphs can help you visualise how well the distribution matches the data. Figure 2 shows the most popular alternative, a comparison of the cumulative curves. Although in this plot it is difficult to see whether the shape of the distribution matches the pattern of the data, it has the advantage that every data value is plotted.

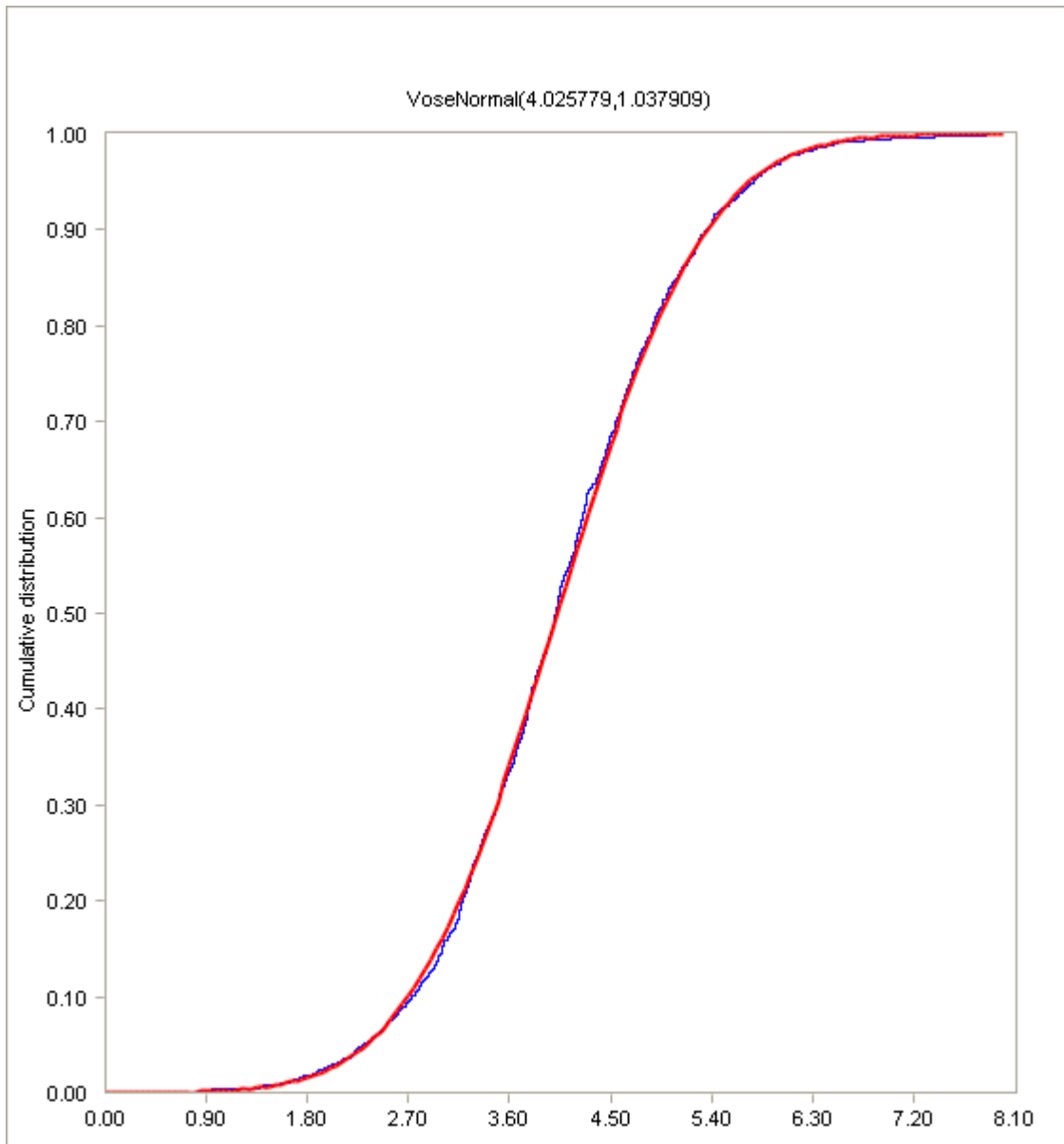


Figure 2: Cumulative plot of data and fitted Normal distribution.

Thus, graphically at least, the normal distribution appears to fit the data very well. We also have to consider whether the properties of the normal distribution are reasonable: a normal distribution stretches from $-\infty$ to $+\infty$, which will almost never reflect reality of course. However, the probability of a value drawn randomly from this fitted normal being less than zero, or greater than 8 are 1 in 19,000 and 1 in 15,500 respectively, which may well be small enough for the purposes of the analysis.

If this is the only distribution type that we intend to use to fit to the data, then the analyst is probably quite comfortable with using a Normal(4.026, 1.038) to represent this variable. However, there remains one problem, namely that the mean and standard deviation estimates are just the best fitting values, they won't be absolutely correct (as we can see, since the true values are 4 and 1 respectively). It may not be a problem when one has large data sets, since the uncertainty in the 'true' values is small (although it might be a problem if the decision is very sensitive to slight changes in these parameters), but it can most certainly be a problem if one has a small data set. In Figure 3 a normal distribution is fit to only

the first 10 data values, plotted in cumulative form. In this figure the red line shows the distribution with the best fitting parameters, and the grey lines show normal distributions with other possible parameter values that the data could have come from. If we had looked only at the best fitting distribution we would have determined that there was about a 55% probability of the random variable falling below a value of 4, but by including the statistical uncertainty about the fitted parameters we now see that this probability could realistically lie anywhere between say 25% and 80%. None of the distribution fitting software that I know of, except [ModelRisk](#), properly considers the statistical uncertainty of the fitted parameters.

Choosing between two or more fitted distributions

It is relatively rare that we are convinced a variable should be represented by one specific type of distribution. There are many types of distribution (ModelRisk has more than 100!) and there can be quite subtle but useful differences in the models underlying them. 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 a good start, though one should appreciate that the data pattern, particularly 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 still have a number of candidate distributions to choose from, which is where a statistical comparison of their fits comes into play.

Goodness of fit statistics and why they are wrong

There are three goodness of fit statistics that are almost ubiquitously used in distribution fitting software: Chi-Squared; Kolmogorov-Smirnoff; and Anderson-Darling

I describe each statistic and their problems in Appendix I for the mathematically curious but, in summary, their fundamental flaws are:

- 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 big 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 distribution, so fitting software products usually use a generic correction for the other distribution types which can be very rough;
- None of these goodness of fit statistics penalise distributions for the number of parameters they use. Thus, a distribution with four 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 – a problem in statistics known as over-fitting;
- None of these goodness of fit statistics can correctly handle truncated, censored or binned data;
- The method of fitting (usually maximum likelihood, or method of moments) is inconsistent with the measurement of degree of fit. It makes sense that if one were to consider one of these statistics are describing the level of fit, the fitting algorithm would try to find parameters that optimise the goodness of fit statistic being used; and
- None of these goodness of fit statistics give a proper statistical weighting to the plausibility of each candidate distribution

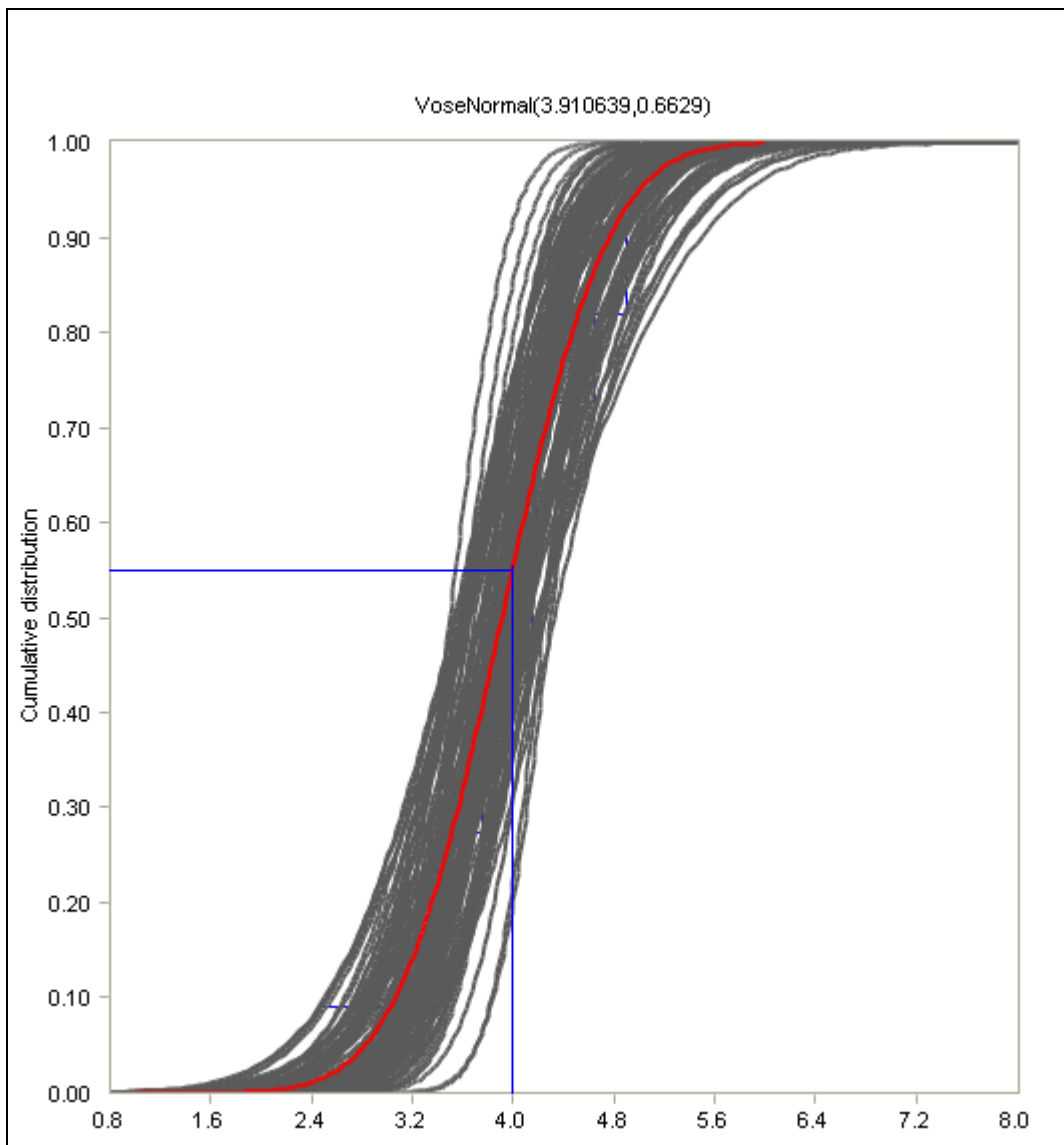


Figure 3: Normal distribution fit to small data set shown with statistical uncertainty about the fit

The solution

The goodness of fit statistics described above were developed decades ago, and their problems are well documented. More modern approaches to evaluating a goodness of fit use information criteria. Let:

n = number of observations (e.g. data values, frequencies)

k = number of parameters to be estimated (e.g. the Normal distribution has 2: mean and standard deviation)

L_{max} = the maximised value of the log-Likelihood for the estimated model (i.e. fit the parameters by MLE (Appendix II) and record the natural log of the Likelihood)

There are three information criteria that use these values as follows:

1. SIC (Schwarz information criterion, aka Bayesian information criterion)

$$SIC = k \ln[n] - 2 \ln[L_{\max}]$$

The SIC (Schwarz, 1978) is the most strict in penalising loss of degree of freedom by having more parameters.

2. AIC (Akaike information criterion)

$$AIC = \left(\frac{n - 2k + 2}{n - k + 1} \right) - 2 \ln[L_{\max}]$$

The AIC (Akaike 1974, 1976) is the least strict of the three in penalising loss of degree of freedom.

3. HQIC (Hannan-Quinn information criterion)

$$HQIC = \left(\frac{n - 2k + 2}{n - k + 1} \right) \ln[\ln[n]] - 2k \ln[L_{\max}]$$

The HQIC (Hannan and Quinn, 1979) holds the middle ground in its penalising for the number of parameters.

The ModelRisk software applies these three criteria as a means of ranking each fitted model, whether it be fitting a distribution, a time series model or a copula (a type of correlation structure). The user can then rank each fitted distribution by the information criterion of choice: in reality, the different information criteria will often provide the same ranking, unless there are few data or there is a wide variation in the number of parameters of the candidate distributions. The information criteria have none of the problems described above for the other 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 (see Appendix II) and be consistent with the goodness of fit statistic;
- The information criteria penalise distributions with greater number of parameters, and thus help avoid the over-fitting problem;
- Since the basis of these statistics is the log-likelihood (Appendix II), information criteria can be used with truncated, censored and binned data; and
- They provide a basis for giving a weighting to how confident one can be in each distribution, which allows one to use a technique known as *Bayesian model averaging*. This means that one can blend several candidate distributions and get a better overall fit, a technique implemented in ModelRisk

Appendix I: Goodness-of-Fit Statistics

The three most commonly used goodness of fit statistics are the Chi Squared (χ^2), Kolmogorov–Smirnov (K–S) and the Anderson–Darling (A–D). The lower the value of these statistics, the closer the fitted distribution appears to match the data.

Goodness-of-fit statistics are not intuitively easy to understand or interpret. They do not provide a true measure of the probability that the data actually comes from the fitted distribution. Instead, they provide a probability that random data generated from the fitted distribution would have produced a goodness-of-fit statistic value as low as that calculated for the observed data.

Critical Values and Confidence Intervals for Goodness-of-Fit Statistics

Analysis of the χ^2 , K–S and A–D statistics can provide confidence intervals proportional to the probability that the fitted distribution could have produced the observed data.

Critical values are determined by the required confidence level α . They are the values of the goodness-of-fit statistic that have a probability of being exceeded that is equal to the specified confidence level. Critical values for the χ^2 test are found directly from the χ^2 distribution. The shape and range of the χ^2 distribution are defined by the degrees of freedom ν , where:

$$\nu = N - k - 1$$

N = number of histogram bars or classes

k = number of parameters that are estimated to determine the best-fitting distribution.

Figure 4 shows a descending cumulative plot for the $\chi^2(11)$ distribution, i.e. a χ^2 distribution with 11 degrees of freedom. This plots an 80% chance (α , the confidence interval) that a value would have occurred that was higher than 6.988 (the critical value at an 80% confidence level) for data that was actually drawn from the fitted distribution, i.e. there is only a 20% chance that the χ^2 value could be this small. If the analyst is conservative and accepts this 80% chance of falsely rejecting the fit, his confidence interval α equals 80% and the corresponding critical value is 6.988 and he will not accept any distribution as a good fit if its χ^2 is greater than 6.988.

Critical values for K–S and A–D statistics have been found by Monte Carlo simulation (Stephens 1974, 1977; Chandra et al. 1981). Tables of critical values for the K–S statistic are very commonly found in statistical textbooks. Unfortunately, the standard K–S and A–D values are of limited use for comparing critical values, particularly if there are fewer than about 30-50 data points. The problem arises because these statistics are designed to test whether a distribution with *known* parameters could have produced the observed data. If the parameters of the fitted distribution have been estimated from the data, the K–S and A–D statistics will produce conservative test results, i.e. there is a smaller chance of a well-fitting distribution being accepted. The size of this effect varies between the type of distribution being fitted.

Modifications to the K–S and A–D statistics have been determined to correct for this problem as shown in Tables 1 and 2, where n is the number of data points and D_n and A_n^2 are the unmodified K–S and A–D statistics respectively. However, Figure 5 shows that the ‘all others’ class (i.e. where the distribution being fit is not one of those listed) can give quite inaccurate results.

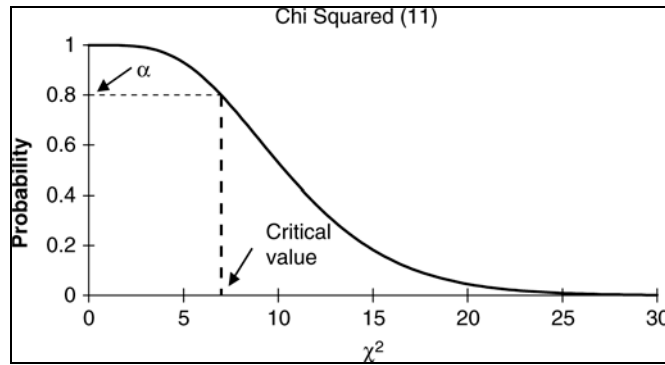


Figure 4: The critical value at an 80% confidence interval (α) for a ChiSquared(11) distribution.

Distribution	Modified test statistic
Normal	$\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}} \right) D_n$
Exponential	$\left(D_n - \frac{0.2}{n} \right) \left(\sqrt{n} + 0.26 + \frac{1}{2\sqrt{n}} \right)$
Weibull and Extreme Value	$\sqrt{n} D_n$
All others	$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}} \right) D_n$

Table 1: Kolmogorov–Smirnov statistics.

Distribution	Modified test statistic
Normal	$\left(1 + \frac{4}{n} - \frac{25}{n^2} \right) A_n^2$
Exponential	$\left(1 + \frac{0.6}{n} \right) A_n^2$
Weibull and Extreme Value	$\left(1 + \frac{0.2}{\sqrt{n}} \right) A_n^2$
All others	A_n^2

Table 2: Anderson–Darling statistics.

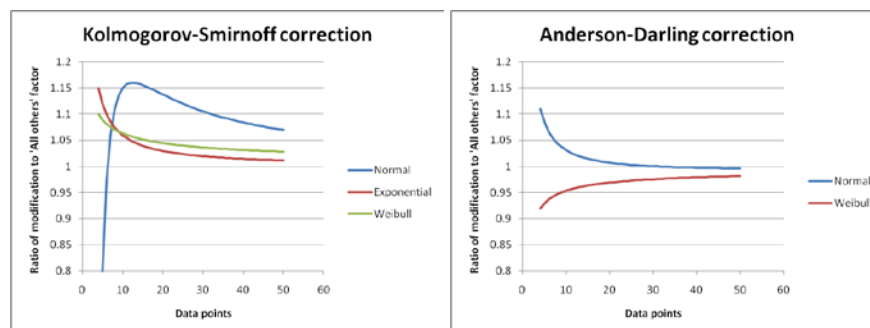


Figure 5: The K-S and A-D corrections as a proportion of the correction used for ‘all other’ distribution. There are inconsistencies here: for example, in the left pane, the Weibull distribution can look like the Normal or the Exponential depending on the value of the shape parameter, yet has a quite different correction for a specified number of data points. Many other distributions approach a Normal shape, yet the K-S and A-D statistics can be 10-15% different if one specifies a Normal distribution rather than the generic ‘all others’ group.

The Chi Squared Goodness-of-Fit Statistic

The Chi Squared (χ^2) statistic measures how well the expected frequency of the fitted distribution compares with the observed frequency of a histogram of the observed data. The Chi Squared test makes the following assumptions:

1. The observed data consists of a random sample of n independent data points.
2. The measurement scale can be nominal (i.e. non-numeric) or numerical.
3. The n data points can be arranged into histogram form with N contiguous classes that cover the entire possible range of the variable.

The Chi Squared statistic is calculated as follows:

$$\chi^2 = \sum_{i=1}^N \frac{\{O(i) - E(i)\}^2}{E(i)}$$

where $O(i)$ is the observed frequency of the i th histogram class or bar and $E(i)$ is the expected frequency from the fitted distribution of x -values falling within the x -range of the i th histogram bar. $E(i)$ is calculated as

$$E(i) = \{F(i_{\max}) - F(i_{\min})\} * n$$

where:

$F(x)$ = distribution function of the fitted distribution

(i_{\max}) = the x -value upper bound of the i th histogram bar

(i_{\min}) = the x -value lower bound of the i th histogram bar

Since the χ^2 statistic sums the *squares* of all of the errors $\{O(i) - E(i)\}$, it can be disproportionately sensitive to any large errors, e.g. if the error of one bar is three times that of another bar, it will contribute nine times more to the statistic (assuming the same $E(i)$ for both).

χ^2 is very dependent on the number of bars N that are used. By changing the value of N , one can quite easily switch ranking between two distribution types. Unfortunately, there are no hard and fast rules for selecting the value of N .

By equating the calculation to a Chi-Squared distribution, one is assuming that $(O(i) - E(i))$ follows a Normal distribution with zero mean. Since $O(i)$ is in fact Binomial this approximation only works when there are a large number of data points within each class.

Kolmogorov–Smirnov (K–S) Statistic

The K–S statistic D_n is defined as:

$$D_n = \max[|F_n(x) - F(x)|]$$

Where

D_n is known as the K–S distance

n = total number of data points

$F(x)$ = distribution function of the fitted distribution

$F_n(x) = i/n$

i = the cumulative rank of the data point.

The K–S statistic is thus only concerned with the maximum vertical distance between the cumulative distribution function of the fitted distribution and the cumulative distribution of the data. Figure 6 illustrates the concept for data fitted to a Normal distribution.

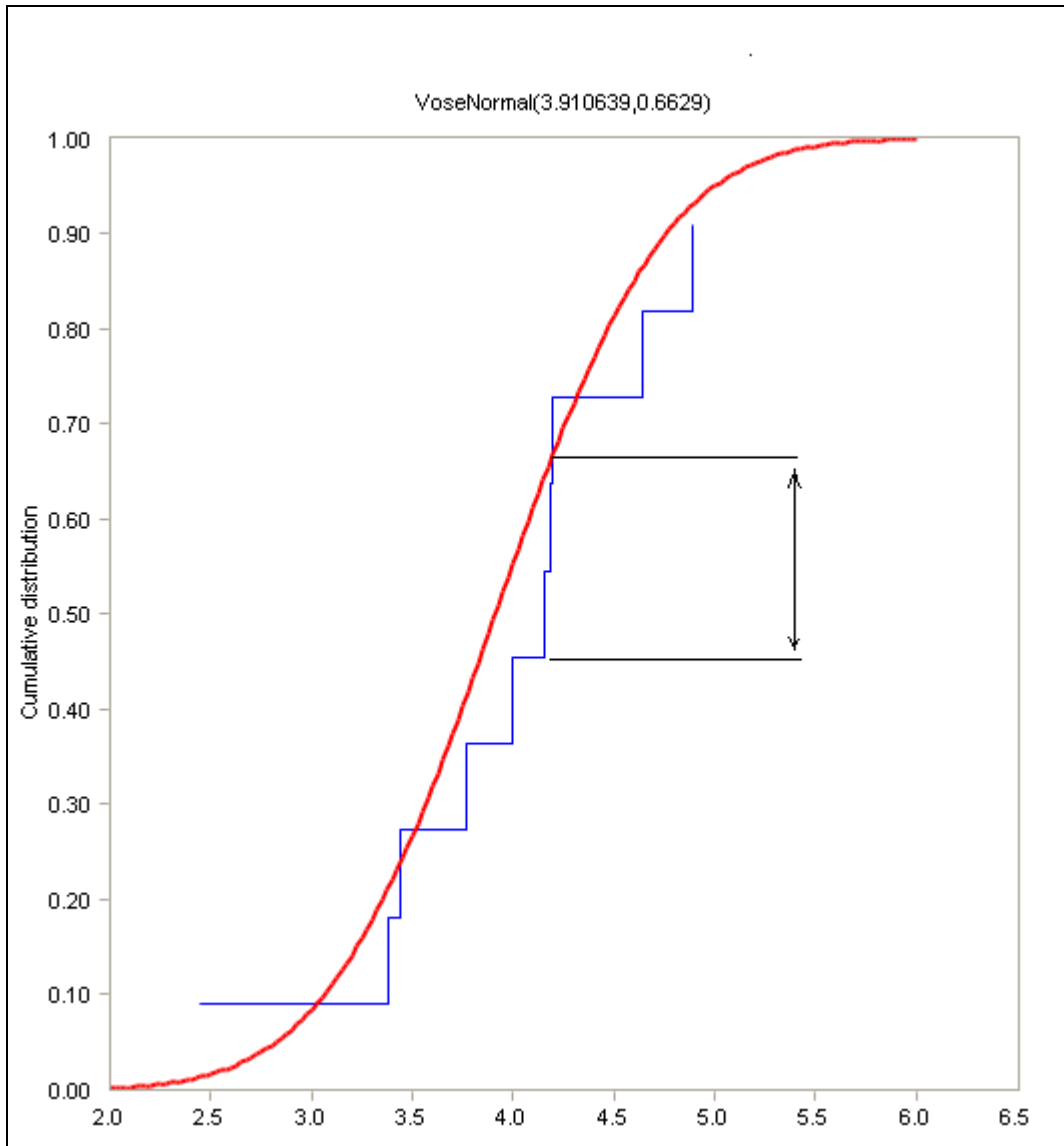


Figure 6: Calculation of the Kolmogorov–Smirnov distance D_n shown as the vertical distance marked with arrows.

The K–S statistic’s value is only determined by the one largest discrepancy and takes no account of the lack of fit across the rest of the distribution.

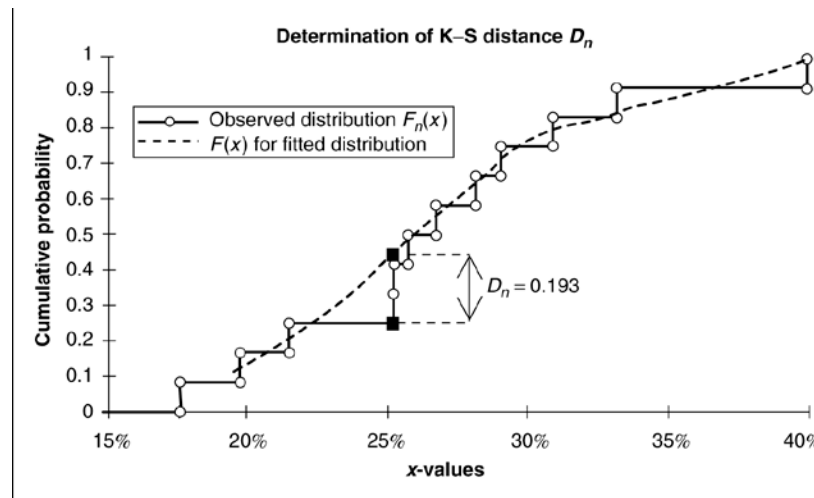
The vertical distance between the observed distribution $F_n(x)$ and the theoretical fitted distribution $F(x)$ at any point, say x_0 , itself has a distribution with a mean of zero and a standard deviation σ_{K-S} given by binomial distribution theory:

$$\sigma_{K-S} = \sqrt{\frac{F(x_0)[1 - F(x_0)]}{n}}$$

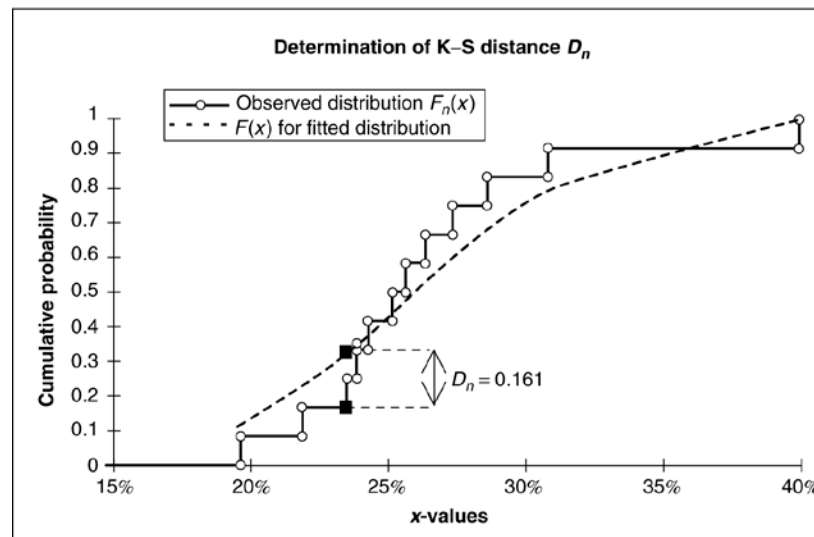
The size of the standard deviation σ_{K-S} over the x -range is shown in Figure 7 for a number of distribution types with $n = 100$. The position of D_n along the x -axis is more likely to occur where σ_{K-S} is

greatest which, as Figure 7 shows, will generally be away from the low probability tails. This insensitivity of the K-S statistic to lack of fit at the extremes of the distributions is addressed in the Anderson–Darling statistic.

The enlightened statistical literature is quite scathing about distribution fitting software that use the KS statistic as a goodness of fit if one has estimated the parameters of a fitted distribution from data. This was not the intention of the K-S statistic, which assumes that the fitted distribution is fully specified. In order to use it as a goodness-of-fit measure that ranks levels of distribution fit one must perform simulation experiments to determine the critical region of the K-S statistic in each case.



(a) Distribution is generally a good fit except in one particular area



(b) Distribution is generally a poor fit but with no single large discrepancies

Figure 6: How the K-S distance D_n can give a false measure of fit because of its reliance on the single largest distance between the two cumulative distributions rather than looking at the distances over the whole possible range.

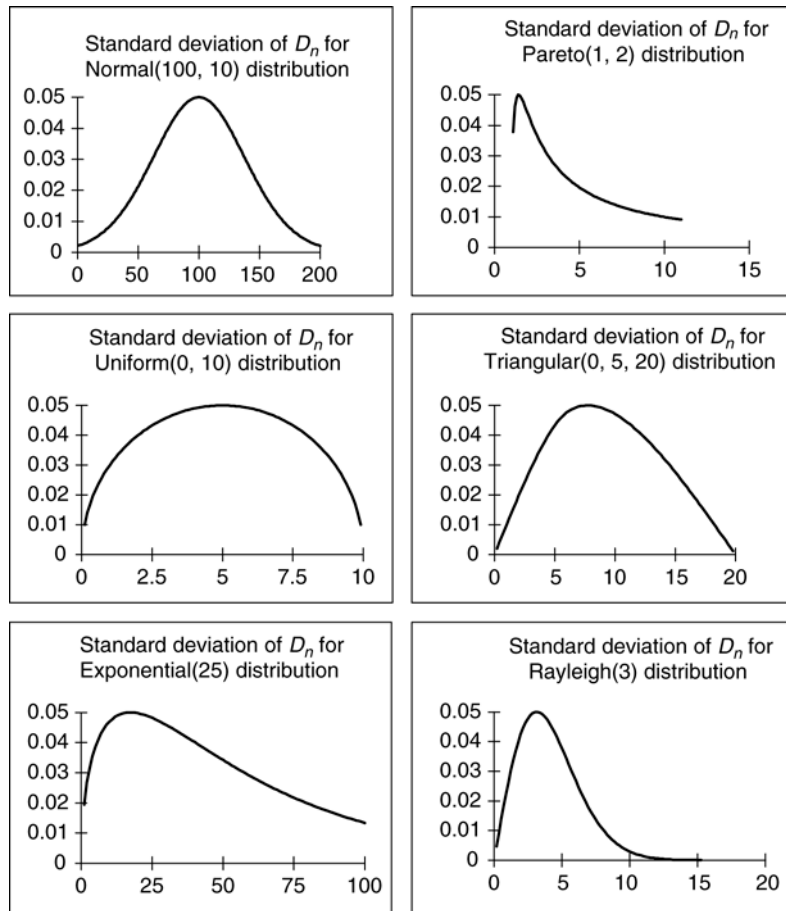


Figure 7: Variation of the standard deviation of the K-S statistic D_n over the range of a variety of distributions. The greater the standard deviation, the more chance that D_n will fall in that part of the range, which shows that the K-S statistic will tend to focus on the degree of fit at x-values away from a distribution's tails.

Anderson–Darling (A–D) Statistic

The A–D statistic A_n^2 is defined as

$$A_n^2 = \int_{-\infty}^{\infty} |F_n(x) - F(x)|^2 \Psi(x) f(x) dx$$

$$\Psi(x) = \frac{n}{F(x)\{1 - F(x)\}}$$

n = total number of data points

where $F(x)$ = distribution function of the fitted distribution

$f(x)$ = density function of the fitted distribution

$F_n(x) = i/n$

i = the cumulative rank of the data point

The Anderson–Darling statistic is a more sophisticated version of the Kolmogorov–Smirnov statistic. It is more powerful for the following reasons:

- $\Psi(x)$ compensates for the increased variance of the vertical distances between distributions (σ_{K-S}^2), which is described in Figure 7.
- $f(x)$ weights the observed distances by the probability that a value will be generated at that x -value.
- The vertical distances are integrated over *all* values of x to make maximum use of the observed data (the K–S statistic only looks at the maximum vertical distance).

The A–D statistic is therefore a generally more useful measure of fit than the K–S statistic, especially where it is important to place equal emphasis on fitting a distribution at the tails as well as the main body. Nonetheless it still suffers the same problem of the K–S statistic in that the fitted distribution should in theory be fully specified, not estimated from the data. It suffers a larger problem in that the confidence region has been determined for only a very few distributions, as shown in Table 2.

Appendix II

Maximum Likelihood Estimators (MLEs)

The maximum likelihood estimators of a distribution type are the values of its parameters that produce the maximum joint probability density for the observed data X . In the case of a discrete distribution, MLEs maximise the actual probability of that distribution type being able to generate the observed data.

Consider a probability distribution type defined by a single parameter α . The likelihood function $L(\alpha)$ that a set of n data points (x_i) could be generated from the distribution with probability density $f(x)$ – or, in the case of a discrete distribution, probability mass— is given by

$$L(X|\alpha) = \prod_i f(x_i, \alpha) \quad \text{i.e.} \quad L(\alpha) = f(x_1, \alpha) f(x_2, \alpha) \cdots f(x_{n-1}, \alpha) f(x_n, \alpha)$$

The MLE $\hat{\alpha}$ is then that value of α that maximises $L(\alpha)$. It is determined by taking the partial derivative of $L(\alpha)$ with respect to α and setting it to zero:

$$\left. \frac{\partial L(\alpha)}{\partial \alpha} \right|_{\hat{\alpha}} = 0$$

For some distribution types this is a relatively simple algebraic problem, for others the differential equation is extremely complicated and is solved numerically instead.

Maximum likelihood methods offer the greatest flexibility for distribution fitting because we need only be able to write a probability model that corresponds with how our data are observed and then maximise that probability by varying the parameters.

Censored data are those observations for which we know only that they fall above or below a certain value. For example, a weight scales will have a maximum value X it can record: we might have some measurement off the scale and all we can say is that they are greater than X .

Truncated data are those values that are not observed above or below some level. For example, at a bank it may not be required to record an error below \$100 and a sieve system may not capture diamonds from a river below a certain diameter.

Binned data are those observations that we only know the value of in terms of bins or categories. For example, one might record in a survey that customers were (0,10], (10,20], 920-40] and (40+) years of age.

It is a simple matter to produce a probability model for each category or combination, as shown in the following examples where we are fitting to a continuous variable with density $f(x)$ and cumulative probability $F(x)$:

Example 1: Censored data

Observations: Measurement censored at *Min* and *Max*. Observations between *Min* and *Max* are *a,b,c,d* and *e*. *p* observations below *Min* and *q* observations above *Max*.

Likelihood function: $f(a)*f(b)*f(c)*f(d)*f(e)*F(\text{Min})^p*(1-F(\text{Max}))^q$

Explanation: For *p* values we only know that they are below some value *Min*, and the probability of being below *Min* is $F(\text{Min})$. We know *q* values are above *Max*, each with probability $(1-F(\text{max}))$. The other values we have the exact measurements for.

Example 2: Truncated data

Observations: Measurement truncated at *Min* and *Max*. Observations between *Min* and *Max* are *a,b,c,d* and *e*.

Likelihood function: $f(a)*f(b)*f(c)*f(d)*f(e)/(F(\text{Max})-F(\text{Min}))^5$

Explanation: We only observe a value if it lies between *Min* and *Max* which has probability $(F(\text{Max})-F(\text{Min}))$, and there are five observations.

Example 3 Binned data

Observations: Measurement binned into continuous categories as follows:

Bin	Frequency
0-10	10
10-20	23
20-50	42
50+	8

Likelihood function: $F(10)^{10}*(F(20)-F(10))^{23}*(F(50)-F(20))^{42}*(1-F(50))^8$

Explanation: We observe values in bins between a *Low* and *High* value with probability $F(\text{High}) - F(\text{Low})$.

References

- Akaike, H. (1974): "A New Look at the Statistical Model Identification," *I.E.E. Transactions on Automatic Control*, **AC 19**, 716-723.
- Akaike, H. (1976): "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion," in R. K. Mehra and D. G. Lainotis (eds.), *System Identification: Advances and Case Studies*, Academic Press, New York, 52-107.
- Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Stat.* **23**, 193–212.
- Chandra, M., Singpurwalla, N. D. and Stephens, M. A. (1981). Kolmogorov statistics for tests of fit for the extreme value and Weibull distribution. *J. Am. Stat. Assoc.* **76**(375), 729–731.
- Hannan, E. J., and B. G. Quinn (1979): "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, B*, **41**, 190-195.
- Schwarz, G (1978). Estimating the Dimension of a Model. *The Annals of Statistics* **6**(2) (Mar., 1978), 461-464.
- Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.* **69**(347), 730–733.
- Stephens, M. A. (1977). Goodness of fit for the extreme value distribution. *Biometrika* **64**(3), 583–588.