# Automated fitting using AIC

This document describes an algorithm for automated selection of a distribution from a set of feasible distributions. This document details the theory and justification behind algorithm.

The candidate distributions are those that are defined in HAMISH Reliability Basics:

* Weibull
* Normal
* Exponential
* Rayleigh
* Extreme Value
* Log Normal

and we employ the Akaike’s Information Criterion (AIC) to assess the adequacy of the distribution. For a given distribution, , with parameters , AIC is computed as

where is the (maximised) likelihood from the fitting procedure. Note that AIC has a penalty for increasing the number of parameters. The algorithm is as follows:

1. **Input** the candidate distributions, and the failure time data to be used for fitting,
2. Compute the maximum likelihood estimates and AIC values for all .
3. Sort by AIC in ascending order
4. Compute
5. **Return**

This algorithm was implemented in the MATLAB code selectDistributionAIC.m and the code demo.m (both of which can be found in the same folder as this document).

A few notes with AIC:

* AIC can be used to assess the *relative* goodness of fit only. It is silent on how well the best distribution represents the data. It might be good to add a hypothesis test on the adequate distributions to test if we can reject that they are representative of the data set.
* When AIC is computed for 2 times the number of parameters, a low sample size correction is used.
* AIC differences of < 2 are considered to be (statistically) equivalent (Burnham & Anderson, 2002, pp. 70)
* In the code, a small-sample bias correction term is implemented which goes to zero as the number of sample sizes goes to infinity. Despite the fact that this was derived under the assumption of the Gaussian distribution, Burnham and Anderson advocate its use generally for small sample sizes (Burnham & Anderson, 2002, pp. 66). Thus, it was implemented in the code.

In demo.m, a series of random “failure times” are generated and then the above algorithm is applied to select the distribution. Note that distributions that have are considered adequate so that there may be more than one distribution that adequately fits a model. A few results are demonstrated:

* Fig. 1 shows the result of applying the algorithm to sets of failure time data of size when the true underlying distribution is a Weibull with scale = 200 and shape = 5.
* Fig. 2 shows the result of applying the algorithm to sets of failure time data of size when the true underlying distribution is a Weibull with scale = 200 and shape = 5. We can see that fewer mistakes are made in the selection of the distribution.
* Fig. 3 shows the result of applying the algorithm to sets of failure time data of size when the true underlying distribution is a Weibull with scale = 200 and shape = 2. We can see that the Rayleigh distribution is correctly deemed adequate 90% of the time.
* Fig. 4 shows the result of applying the algorithm to sets of failure time data of size when the true underlying distribution is uniform. we can see that the algorithm shows that Weibull and Normal distribtuions are often equally representative of the data (but clearly not *good* representations)! This is the major limitation of this algorithm right now.



Fig. 1



Fig. 2



Fig. 3



Fig. 4