Introduction

This IWO Data Analyzer (IDA) aims at analyzing failure data. It is assumed that a 2 parameter Weibull distribution applies. Generally, the Weibull distribution applies to systems where a lowest positive value (like time or strength) represents the performance of the whole system. E.g., the strength of a chain in a test with increasing pulling force, is determined by the link that breaks at the lowest pulling force. Or the time to failure of an electric component in a voltage test, is determined by the position that has the shortest lifetime.

Motivation for IDA and approach

The IWO Data Analyzer aims to support decision-making after early failures in networks or product batches. Large data sets enable accurate analysis. But, these may not be available with early failures. Still, stakes may be high and timely decisions may have to made. Data may be scarce and incomplete. The inherent large uncertainties may be acceptable as decisions do not depend on subtleties. A 17% or 88% of a new failure within a year may lead to the very same decision to replace all products preventively.

Interestingly, emergency situations due to early (unexpected) defects are often caused by a single, dominant mechanism (cf. [A]). This is the reason why early failures of even complicated products can often be analyzed in terms of a single Weibull distribution.

IDA aims to be widely accessible in support of resilience. Although the background is reliability of the electric power supply, the techniques in IDA are useful in the many situations where Weibull distributions apply. Also Small and Medium Enterprises or educational institutes may find the IDA a convenient tool for data analysis.

IDA can provide a foundation for the evaluation and decision-making after a few products failed and the quality of the remaining products is suspect because of that. Typical questions that come to mind are:

- Are the early failures exceptional or normal considering specifications?
- Is this the beginning of many faults or an incident?
- When can a next failure be expected?

In case the purpose is to analyze a set of failure data and/or to compared to another group, this tool may prove to be useful just for that. The input tab and ‘wplot’ tab may be all that is needed then.

Finally, since the motivation is based on studying failures in grids, the data in IDA are referred to as times mostly. However, this is not a fundamental requirement; any variables can be analyzed assuming these have positive values: e.g., the pulling force or voltage mentioned above. However, for convenience, the variable will often be referred to as time to breakdown.

Incomplete data - censoring

As mentioned, IDA can analyze situations where early failures cast doubt on the quality of the as yet surviving rest of the batch. Failure data that remain unknown at the evaluation are called censored (or suspended). IDA is designed to handle not only the actually observed failure times, but also such censored data. Observed and censored data must be processed differently though.

The number of data in a set is called the ‘sample size’. Three sample sizes are relevant: ‘**n**’ as the **total** of failures; ‘**r**’ as the number of **observed** failures; ‘**k**’ as the number of **censored** failures. They relate as: **n = r + k**.

In more detail, assume at the time of evaluation: n test objects are involved; r failure times t_1, .., t_r were observed and k objects survived with times s_1,..,s_k. The knowledge of both t-values and s-values are relevant when we estimate the failure distribution. Therefore both t_i (i=1,..,r) and s_j (j=1,..,k) are used as input.

IDA can process 2 sets (d1 and d2), each being a combined set of observed and censored data. If all objects were simultaneously put to the test, then the s-data may have the same value, but if the test period of the survivors varied, their censored times will differ.

In addition to the data sets d1 and d2, IDA also features a reference distribution (ref) of which the parameters and samples size n can be specified. In that case IDA produces 2 data sets: the expected log-times and the median times. The first are used in Weibull plots, the latter for estimating the 50% probability failure times.

Not the one and only possible approach

Although analyzing data may seem straightforward, alternative approaches exist. The IDA is based on several choices that we had to make to meet our requirements. Others may favor alternative approaches e.g. for the parameter estimation or plotting positions in the graph. IWO aims to be transparent in our choices and are open to feedback and debate.

One goal is that we hope to serve a wide audience with the IDA. For that reason we targeted methods can be applied with software that is widely available in offices, such as spreadsheets.

Another goal is user-friendliness. IWO aims at a relatively light-weight computing tool: modest in memory claim, fast (instant results on input of data) and ease of use for a wide audience, implementable in widely mastered office software.

As for the distribution, we used a very common failure distribution, namely the 2-parameter Weibull distribution. A 3-parameter Weibull is sometimes used, namely when there is a delay before actual degradation starts. This is not done here, but if there is a known or assumed delay, then this delay may be subtracted from the input data before the Weibull-2 analysis is performed. Basically the delay is regarded as a time shift of the starting point. Another often used distribution is the Exponential distribution. This is applicable if the items are not aging or when the products are frequently revived by maintenance. This corresponds to a Weibull-1 distribution with β=1.

Finally, a similarity index is introduced to evaluate the similarity of two distributions by their distribution densities.

Disclaimer

IDA is meant to be freeware for non-commercial use. It is anticipated that this can be used for studies related to asset management, failure investigations, education and product improvement. A disclaimer is in place: IDA comes with no warranty and IWO is not accountable for consequences, errors and/or conclusions based on the use of this freeware spreadsheet and educational tool.

Definitions, equations and references

This sheet gives definitions, equations and references for the IWO Data Analyzer (IDA).

Notation

In literature, theoretical parameters are often written with Greek and estimated parameters with Roman characters. In the present spreadsheet, theoretical and estimated parameters α, β and θ are all written with Greek characters for quick recognition. Theoretical parameters are defined with the reference (set 'r'). Estimated parameters appear with the data sets 'd1' and 'd2'. Estimated parameters are often specified ± error. The estimated parameters α, β and θ are not unbiased.

Weibull distribution

IDA is based on the Weibull-2 distribution:

F(t) = $1 – \exp(-(t/\alpha)^{\beta})$ [**cumulative distribution function**]

t = variable [e.g., time to breakdown]

α = the scale parameter [same dimension as t]

β = the shape parameter [dimensionless]

Other forms of the Weibull distribution are:

R(t) = 1-F(t) [**Reliability function**]

f(t) = $dF/dt = -dR/dt = \beta\cdot(t^{\beta-1})\cdot R(t)/(\alpha^{\beta})$ [**distribution density function**]

h(t) = $f(t)/R(t) = \beta\cdot(t^{\beta-1})/(\alpha^{\beta})$ [**hazard rate**]

The **mean** or expected <t> is: **<t> = θ** = $\alpha\cdot\Gamma(1+1/\beta)$

where Γ(u) is the Gamma function of u.

Adjusted rank

For plotting and parameter estimation by regression, the data are ranked in increasing order of magnitude and indexed with the **rank index 'i'** ($1\leq i\leq n$).

If a set of r<n failure times t_i is observed and k=n-r failure times s_j are censored, random censoring may occur: i.e., some survivors may have a shorter operational lifetime s_j than one or more observed failure times t_i. It is therefore possible that one or more survivors will fail earlier than one or more presently observed t_i. Consequently, the rank index i of some already failed items becomes uncertain, because as yet survivors may fail some time in between. Methods exist to deal with this uncertainty.

An estimated '**adjusted rank**' **I(i)** is advised in [B]. It averages the possible ultimate rankings. The adjusted ranking can be calculated as follows:

- **I(0)** = 0 by definition
- **I(i)** = $I(i-1)+[n+1-I(i-1)]/[n+2-C_i]$

C_i = the original rank index i increased with the number of censored times up to t_i. Note: adjusted ranks I(i) are not necessarily integers, but real numbers $1\leq I(i)\leq n$.

Regression based parameter estimation

(Weighted) Linear Regression was chosen for estimation of the Weibull parameters for 3 reasons. Firstly, plotting and parameters estimation are fully in line then. Secondly, **(W)LR** yields analytical results, which are fast and enable error calculation. Thirdly, the variance of the unbiased WLR β estimator is equal to that of the Maximum Likelihood estimator.

Both the (Ordinary) Linear Regression (LR) and the Weighted LR (WLR) are used in IDA. Some remarks:

- The variables for (W)LR are the expected plotting positions $<Z_i>$; the covariables are the log(t)
- The $<Z_i>$ are defined as: $<\{\log(-\ln(1-F_i))\}>$
- Adjusted rank indices I(i) is applied
- [B] advises $<Z_i>\approx\log(-\ln(1-F_i))$ with $F_i\approx(i-0.44)/(n+0.25)$. For alternatives, see [C].
- For WLR, the weights $\mathbf{w_i = 1/v_i}$, with $v_i=\mathrm{var}(Z_i)$
- [A] defines an accurate approximation for v_i:

$$\mathbf{v_i = (i-0.5)^{-1}-0.1\cdot(i-0.3445)^{-3}+ ..}$$
$$\mathbf{.. +0.125\cdot(n+0.343)^{-1.656}\cdot(n+0.8-i)^{-0.75}\cdot(i-1)^{1.4}}$$

Note: the adopted (W)LR methods and weights $w_i=1/v_i$ are suitable for non-integer I(i).

Confidence limits and intervals

An **A%** confidence limit of a variable t gives value $\mathbf{t_{A\%}}$ below which A% of the t-values lie. A typical statement is that A% of the variable t values falls below $t_{A\%}$ and 1-A% above. The A% and 1-A% confidence limits $t_{A\%}$ resp. $t_{1-A\%}$ form the boundaries of a **1-2A% confidence interval**.

Two types of confidence limits are in use in IDA: Beta and Regression. **Beta limits** apply to random selection of n objects from an (infinitely) large group. The A% Beta limit marks the value $t_{i,n,A\%}$ below which A% of the ith ranked samples with size n fall. It is related to random selection. (Note: the word 'Beta' refers to the Beta function [C] and *not* to the Weibull shape parameter; avoid confusion!). These limits do not depend on the scatter of the actual data, but can be calculated for any given distribution with known (estimated or defined) parameters. Therefore, it is also possible to calculate these limits for the reference distribution. Beta intervals are useful to identify ranges in which data may be expected, like the next time to failure if sample are independent but from a same distribution.

Regression limits are due to scatter in the observed data around the best fit. This scatter about the best fit is supposed to be Gaussian and provides a means for the confidence in the best fit [C]. The more scatter, the wider the regression confidence intervals. As the reference distribution data (set ‘ref’) are on the defined line, the scatter (as determined by the residues) is zero. The data sets (‘d1’ and ‘d2’) will normally feature scatter about the best fit and the regression intervals will be non-zero. Regression intervals are useful to identify outliers.

References

[**A**] Ross, R., P.A.C. Ypma, G. Koopmans, 2021, Weighted Linear Regression based Data Analytics for Decision Making after Early Failures, 10th IEEE PES International Innovative Smart Grid Technologies Conference, Brisbane, paper 148, 5 pp.

[**B**] IEEE Std 930-2004 Guide for the Statistical Analysis of Electrical Insulation Breakdown Data, New York: The Institute of Electrical and Electronics Engineers, Inc., 2004. - Also: IEC 62539:2007.

[**C**] Ross, R., 2019, Reliability Analysis for Asset Management of Electric Power Grids, Wiley IEEE, ISBN 9781119125174, DOI:10.1002/9781119125204, 520 pp.

Getting started with the IWO Data Analyzer (IDA)

Data analysis can be carried out in the following steps:

0. If this file is the original file, you may wish to make a copy with a meaningful name to save your work
1. **Data input** with main distribution characteristics: see **tab 'input'**.
2. Producing and fine-tuning a **Weibull plot** with or without confidence limits: see **tab 'wplot'**.
3. Various types of **analysis**: see **tab 'analysis'**.
4. Viewing **data details** and **intermediate results**: see **tabs 'd1', 'd2' and 'ref'**.

Generally it is a good idea to store a blank IWO Data Analyzer (IDA) file as a template and to save a version for each analysis project that you want to keep with a filename of your choice. As a suggestion, you can build a filename by combining a date in the format YYMMDD, a short project name, a version index and some keywords. In a directory, such files can be ordered alphabetically to provide a convenient overview. Steps 1-4 are discussed in the following sections.

Input of data - tab 'input'

The input tab is the place to start your data analytics. The tab 'input' has fields where **names** can be assigned to the **project** itself, to the two **data sets** and to the **reference** set. These names will then be used throughout the IDA with the plot legends, in the data tabs and on the analysis tab. Such names are useful if plots and results are copied into presentations or reports.

As mentioned on the Introduction page at the 'Incomplete data – censoring' section, when analyzing the results from a test on **n objects**, there may be **r observed failure data t_i** ($1 \le i \le r \le n$) and **k censored data s_j** ($1 \le j \le k = n - r$) which are usually the breakdown respectively survival times. Both the t- and s-data must be specified on the tab 'input'. The input fields for sets d1 and d2 consist of a left column for the t-data and a right column for the s-data.

There is no need to rank the t- and s-data in increasing order. The IDA will rank the data and count n, r and k. The main analyses are carried out on the tabs 'd1' and 'd2', but a selection of the results also show on top of the data set input columns: the counted numbers n, r, and k; the Weibull scale parameter α and shape parameter β (either the WLR or LR estimators); the mean time to failure (or other relevant variable) θ. The effect of censored data can be studied by adding of removing censored data.

As for the reference set, the Weibull parameters α and β are defined along with a sample size n. With very low β-values (e.g. β=0.2), the Weibull plot on tab 'wplot' will run very flat and cover many decades of variable t. The grid lines and scale markings in the plot are designed for max 10 decades (say, β>0.3). The plot should remain correct.

Weibull plot - tab 'wplot'

On the top righthand side of the tab 'wplot', the appearance of the graph can be fine-tuned by switching the **grid lines** on/off. The grid switched 'on' facilitates convenient reading of the plots with a grey background grid, while the grid switched 'off' yields smooth graphs that are often preferred for reporting.

Next, the **extremes of horizontal axis** are automatically set to the minimum number of decades that show all data, fits and confidence boundaries. The inputs 'start' and 'end' enable to set user-defined axis extremes. A maximum of 10 full decades can be shown, but with a higher span only the grid lines of the first 10 decades are shown, although the plots remain plotted correctly.

Below the general project section are boxes dedicated to **each of the three distributions**. The left-hand side of each box gives a **summary of analysis results**: numbers n, r and k; Weibull parameters α and β as well as the mean θ; the correlation coefficient ρ and similarity indices S (with the other two distributions). The right-hand side of each box is interactive. Firstly the **type of regression analysis** can be chosen to be weighted (**WLR**) or ordinary (**LR**). The plot and all results on 'wplot' automatically follow this choice, but only for the respective set. The F-scale extremes of the fits can be set by 'start F' and 'end F'. This enables to give all fits the same vertical range if desired.

Both **Beta and regression confidence limits** can specified for each data set, such as the 1% and 99% limits to show the 98% **confidence interval** (see also the "Confidence .." section on the "Definitions, equations and references" page. Each set also has a **calculator** for **t↔F conversion**.

Analysis and inferences - tab 'analysis'

Various analyses can be performed with the tab 'analysis'. The tab shows boxes that are arranged in 4 columns and 5 rows. The boxes in the left column provide information about the possible analyses. The 3 columns on the right allow the analyses based on the two data sets (d1 & d2) and the reference distribution (ref).

The top row provides overall information and the key numbers for each of the distributions. On the top the employed regression type is shown. If another regression type is desired, this can be selected on the tab 'W-plot'. The shown Weibull parameters and mean are used for the analyses.

The following analyses can be carried out per set:

1. **Convert time and probability** of failure and estimate the other statistical quantities (R, f, h)
2. Quantification of **similarity between distributions**
3. Finding the **optimum servicing interval** for Period Based Maintenance
4. Estimating the **time of the next failure**

<u>ad 1.</u> Probability and Time: these are two independent calculators to convert the variable t and probability F into each other and into the values of reliability R, probability density f and hazard rate h. The conversion is based on the applicable Weibull parameters.

<u>Usage</u>: fill in t and/or F in the appropriate fields and the other values follow as indicated by the arrow.

<u>ad 2.</u> Similarity S_{fg}: this is a quantification of how much two distributions f and g are alike based on Weibull parameters and time range. Here, the S_{fg} indicates how much the distribution resembles each of the other two distributions. S_{fg} has a value between 0 and 1; $S_{fg}=1$ means f and g are identical; $S_{fg}=0$ means f and g have nothing in common. First, S_{fg} is determined for the full time range $[0,\infty)$. In the second part, the user can define an evaluation range $[0,t_{eval}]$.

In practice, a few early failures may occur (say r out of n failed). From these failures times Weibull parameters can be estimated which define the **observed** distribution f. The distribution g may be **specified** or be another observed distribution. The first S_{fg} in the box indicates how much the observed and the reference distribution resemble **until infinity**. The second S_{fg} covers the similarity over a **given period**, e.g. the time until the last failure. Typically, statements can be made like: "*Until the last failure the similarity between the observed and reference distribution is $S_{fg,t}$. If this trend continues the similarity will be $S_{fg,\infty}$*". Possibly, $S_{fg,t}$ is acceptable, but $S_{fg,\infty}$ not and parties may wish to take precautions. Note: **similarity is not the same as compliance**. (Performance beyond a level means higher quality, but also leads to lower similarity; note: S_{fg}=1 means perfect match).

Usage: The S_{fg} calculation between the distribution over the full range requires no action; just put in data on the 'input' tab. For S_{fg} over a partial range, specify t_{eval} in one of the relevant boxes at row 22 and the S_{fg} with the other two distributions is calculated (if these exist).

ad 3. Optimum PBM cycle: this optimizes the **servicing cycle T based on costs alone** (but other values may be expressed here instead). A balance is calculated between **CAPEX** (installing a new product) and **OPEX** (servicing the product by renewing it periodically). The optimum cycle has the lowest cost of ownership (**TOTEX** rate). The cost unit (**CU**) can be **any chosen currency**. The performance of the optimized cycle is expressed in expected lifetime, average hazard rate, highest hazard rate at end of cycle T, the reliability after the first cycle, and the three cost rates for CAPEX, OPEX and TOTEX. If an **alternative, chosen, cycle T** is filled in, the performance of the chosen cycle is assessed and **compared** to the optimized cycle.

Usage: fill in cost for replacement (the amount that will be depreciated in the end, i.e. CAPEX) and the servicing costs per cycle T (OPEX). The optimized situation is calculated based on the Weibull parameters and the costs. Fill in a chosen interval to compare the performance.

Note: extreme Weibull parameters or other inputs may lead to unrealistic or extreme results. Various warnings may pop up in red. Changing inputs can help. A particular issue can be: with too long chosen cycle T, the reliability drops to (almost) 0, making PBM useless and letting h run up to infinity. Another issue may be a much too short chosen cycle T, keeping average h at (almost) 0 making the average life infinite but also at enormous maintenance cost. Another issue can be extreme shape parameter values (e.g. β>15). The range between too small and too long chosen cycles T_{choice} around the optimum cycle T_{opt} may then become relatively narrow. Typically, 0.5<β<10 would be suitable.

ad.4 Next failure with right censored data: this calculates some features of the next failure based on a number of r observed failures out of a (known or assumed) total number of n products to fail. For this analysis all products must have been taken into operation at the same time (i.e. **right censoring: all s_j equal and > t_r**). All future failing products have a larger lifetime than the r already failed products. The analysis calculates times with given confidence of failure and the mean time at the next failure. With the reference distribution, the numbers n and r may be chosen in this box and it is allowed to have these deviate from what is specified on the tab 'Input'.

Usage 4.1: if tab 'input' is completed for a data set, then r, n and the last observed t_r are known. An automatically check runs on this being a right censored case and if so, the analysis of the next failure can be done. **The median and mean next time are estimated** with no action required. Specifying **confidence limits** on row 44, leads to an **confidence interval** estimation **for the next failure**.

Usage 4.2: if no failure occurred at a given t_{eval} (larger than t_r), the confidence interval, median and mean next failure time can be determined. Specify the t_{eval} and the confidence limits to **evaluate the situation at t_{eval}.**

Usage 4.3: a similar exercise can be carried out for a **reference set**. On the input tab the sample size is treated as complete, but here the n and r can be specified in P40 and P41. The next failure after the r^th^ failure will be characterized as under Usage 4.1 and Usage 4.2.

Intermediate results - tab 'd1', 'd2' and 'ref'

The foundation of the plot and the analyses can be found on the tabs 'd1', 'd2' and 'ref'. Intermediate and final results the are shown on each tab for that distribution. If data are not available, then blank cells or cells filled with 0 are shown. These tabs are not interactive.

The 9 top rows are used for: in columns A-M - identification of the project and the set; summarizing n, r and k; columns O-X - summary of WLR (intermediate and end results and columns Z-AI intermediate results summary of (ordinary) LR results and intermediate results.

The following **general results** are found in the columns:
A: **index** from 1 running up to 1024. Indices i=1,..,256 are used for data; for combined observed and suspended data: 1,..,512; for limit curves: 1,..1024.
B-C: **ranked observed and suspended data**
D: **adjusted rank I(i)** based on IEEE 930
E: approximated **weights for LR** as defined on the right column here
F-H: **support** results for adjusted ranking as ranking j (indexing all data: ranked mixed observed & suspended), ranked all data, reverse index j' counting down j from highest j with non-zero data to j=1 (and then 0)
L: **index** for limit curves from 1 up to n in 1024 values; with steps (n-1023)/1023
M: **plotting positions** corresponding to L

The following **WLR results** are in the columns:
O: **times** at the plotting position $\langle Z_I \rangle$ with I (cf. col D) based on the **WLR best fit**
P-Q: **Beta A% confidence limits** (A% in cells P11,Q11 copied from 'wplot' tab)
R-S: **regression A% confidence limits** (A% in R11,S11 copied from 'wplot' tab)
T-U: **Beta A% confidence limit curves** with index in L (A% equal to P-Q)
V: **median times** (A%=50%) with index L; support data for regression limits W-X
W-X: **regression A% confidence limit curves** with index in L (A% equal to R-S)

The **LR results** correspond to the WLR results except that they are obtained by giving all data equal weight. The descriptions match those of WLR above:
Z: see O, except **LR fit instead of WLR fit**; AA-AB: see P-Q; AC-AD: see R-S; AE-AF: see T-U; AG: see V; AH-AI: see W-X.