Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Trapezoidal estimates for the area under a curve (AUC) or average function value for a set of data points over a range can be used when the data are noisy, or the spacing is too sparse or irregular to fit a deterministic equation or to use data smoothing techniques.

From the SIMFIT main menu select [A/Z], open program **average** and inspect the default test file `average.tf1` which contains this data set.

x	y
1	1
2	2
3	3
4	4
5	5
6	5
7	4
8	3
9	4
10	5
11	8
12	10
13	5
14	9
15	4

Note that program **average** creates a continuous model for the data by joining adjacent points by straight line segments then using the intersection of this model with the Y_{crit} threshold to determine the range of X values where the function lies above or below this threshold.

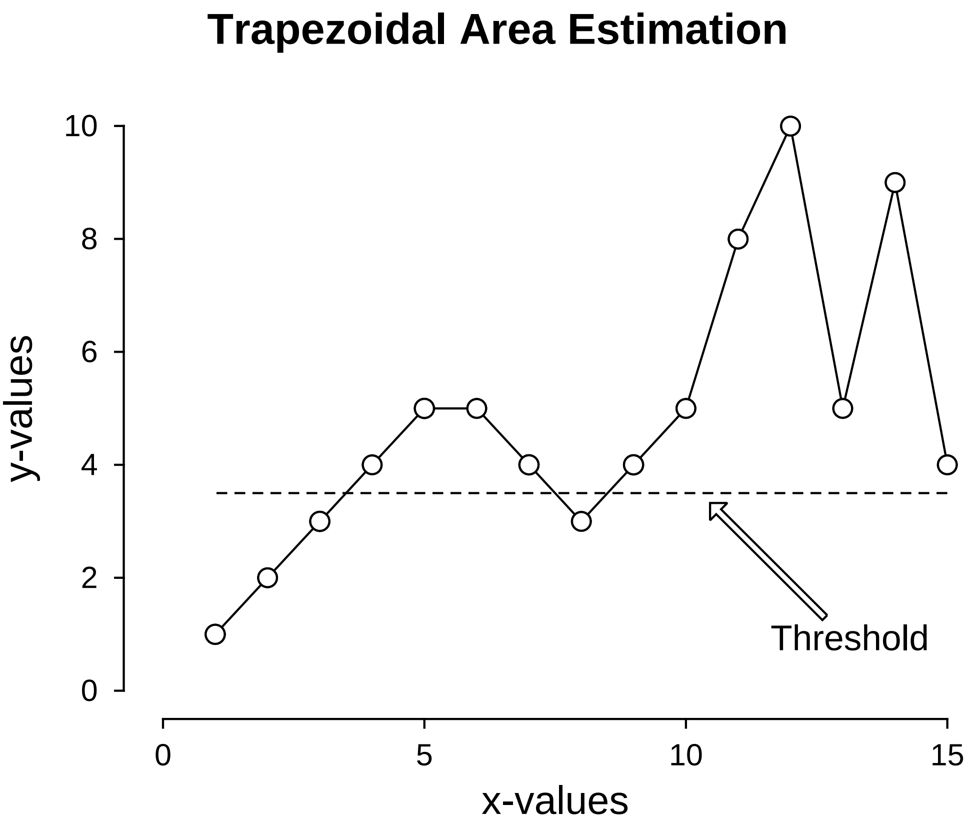
So, proceeding to analyze these data leads to the following results for the default state and the changes resulting from altering the Y threshold from $Y_{test} = 1$ to $Y_{test} = 3.5$.

X_{min}	X_{max}	Y_{min}	Y_{max}					
1	15	1	10					
$Begin = x_1$	$End = x_2$	Above	(as %)	Below	(as %)	Y_{test}	Y_{area}	Y_{mean}
1	15	14	100	0	0	1.0	69.5	4.96
1	15	10.5	75	3.5	25	3.5	69.5	4.96

The results displayed in this table have the following interpretations.

1. The minimum and maximum data values $X_{min}, X_{max}, Y_{min}, Y_{max}$.
2. The range x_1, x_2 of x selected where $X_{min} \le x_1 \le x \le x_2 \le X_{max}$.
3. The trapezoidal area estimate Y_{area} over the range x_1, x_2.
4. The average function value $Y_{mean} = Y_{area}/(x_2 - x_1)$ over the range x_1, x_2.
5. The critical threshold y value Y_{test} where $Y_{min} \le Y_{test} \le Y_{max}$.
6. The percentage of the range x_1, x_2 where the piecewise linear model lies above the critical threshold.
7. The percentage of the range x_1, x_2 where the piecewise linear model lies below the critical threshold.

To understand the effect of changing Y_{test} as are emphasized in red in the above table, consider the next graph.



Further details are now provided to compare and contrast the use of program **average** with some alternative SimFIT programs.

Estimating AUC using deterministic equations

Observations y_i are often made at settings of a variable x_i as for a regression, but where the main aim is to determine the area under a best fit theoretical curve AUC rather than any best fit parameters. Frequently also $y_i > 0$, which is the case we now consider, so that there can be no ambiguity concerning the definition of the area under the curve. One example would be to determine the average value f_{average} of a function $f(x)$ for $\alpha \le x \le \beta$ defined as

$$f_{\text{average}} = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} f(u)\, du.$$

Another example is motivated by the practise of fitting an exponential curve in order to determine an elimination constant k by extrapolation, since

$$\int_0^{\infty} \exp(-kt)\, dt = \frac{1}{k}.$$

Yet again, given any arbitrary function $g(x)$, where $g(x) \ge 0$ for $\alpha \le x \le \beta$, a probability density function f_T can always be constructed for a random variable T using

$$f_T(t) = \frac{g(t)}{\displaystyle\int_{\alpha}^{\beta} g(u)\, du}$$

which can then be used to model residence times, etc.

If the data do have a known form, then fitting an appropriate equation is probably the best way to estimate slopes and areas.

For instance, in pharmacokinetics you can use program **exfit** to fit sums of exponentials and also estimate areas over the data range and AUC by extrapolation from zero to infinity since

$$\int_0^{\infty} \sum_{i=1}^{n} A_i \exp(-k_i t)\, dt = \sum_{i=1}^{n} \frac{A_i}{k_i}$$

which is calculated as a derived parameter with associated standard error and confidence limits. Other deterministic equations can be fitted using program **qnfit** since, after this program has fitted the requested equation from the library or your own user-supplied model, you have the option to estimate slopes and areas using the current best-fit curve.

Estimating AUC using splines

The SIMFIT spline fitting programs **compare** and **spline** can also be used if the data are extensive enough to fit a meaningful cubic spline reference curve.

Estimating AUC using program average

The main objection to using a deterministic equation to estimate the AUC stems from the fact that, if a badly fitting model is fitted, biased estimates for the areas will result. For this reason, it is frequently better to consider the observations y_i, or the average value of the observations if there are replicates, as knots with coordinates x_i, y_i defining a linear piecewise function. This can then be used to calculate the area for any sub range x_1, x_2 where $X_{min} \le x_1 \le x \le x_2 \le X_{max}$.

Another use for the trapezoidal technique is to calculate areas above or below a baseline, or fractions of the x range above and below a threshold, for example, to record the fraction of a certain time interval that a patients blood pressure was above a baseline value.

Note that, in the previous figure, the base line was set at $y = 3.5$, and program **average** calculates the points of intersection of the horizontal threshold with the linear spline in order to work out fractions of the x range above and below the baseline threshold. For further versatility, you can select the end points of interest, but of course it is not possible to extrapolate beyond the data range to estimate AUC from zero to infinity.

Another feature of program **average** is that the parameters x_1, x_2, Y_{test} that are set for the first data set can be preserved instead of being re-set for each data set. This is intended for a scheme where the first set is intended to be a reference data set. Of course, if these parameters are not consistent with subsequent data sets, warning messages will be issued, and the parameters will be re-set to the normal defaults.

$$\begin{aligned} x_1 &= X_{min} \\ x_2 &= X_{max} \\ Y_{test} &= Y_{min} \end{aligned}$$