Formulae for Statistical Treatment of Experimental Data

1. *n* independent experiments

Average of *n* independent experiments

$$\bar{x} = \frac{\sum x_i}{N}$$

Variance

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{N}}{N-1} = \frac{1}{N-1}\sum (x_i - \bar{x})^2$$

Standard deviation of the mean

$$\sigma_m = s / \sqrt{N}$$

Uncertainty interval

$$\bar{x} \pm t\frac{s}{\sqrt{n}} = \bar{x} \pm t\sigma_m$$

where *t* is student's factor for a given probability and $N-1$ degrees of freedom.

<u>Numerical example</u>

Seven independent experiments led to the values given below for the O–H bond dissociation enthalpy in phenol, $\mathrm{X} = DH^{o}(\mathrm{PhO-H})$. Calculate the average value and the associated uncertainty interval.

N	*X* / kJ mol^{-1}	*X*–*X*$_{mean}$	(*X*–*X*$_{mean}$)^2
1	366.3	-5.01	25.14
2	372.1	0.79	0.62
3	375.0	3.69	13.58
4	374.3	2.99	8.91
5	369.3	-2.01	4.06
6	372.4	1.09	1.18
7	369.8	-1.51	2.29

Sum of *X*	2599.2	
X_{mean}	371.31	
Sum of $(X–X_{mean})^2$	55.79	
Sum^2/ 7	965120.09	
Variance		9.30
Sigma m squared		1.33
Sigma m		1.15
Uncertainty (2×Sigma m)		2.3

The result is 371.2±2.3 kJ mol^{-1}. The factor 2, which multiplies the standard deviation of the mean, is the so-called “thermochemical convention”. An alternative procedure, as pointed out above, is using student’s *t*-factor for 6 degrees of freedom and for a given probability.

The variance can be easily calculated with the Descriptive Statistics package (Tools, Data Analysis) of Microsoft Excel. Note that this package does not provide the standard deviation of the *mean* directly. The value for the student’s factor may be calculated by using the statistical function “TINV” with arguments 0.05 (1–0.95 probability) and the corresponding number of degrees of freedom.

2. Error propagation for a given function

Error propagation for $Y = f(x_1, x_2, x_3)$

Error in Y (σ_i are the σ_m for each x_i)

$$\sigma_Y^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2$$

Examples: $y = mx + b$

$$\sigma_Y = |m|\sigma_x \; ; \; \sigma_Y^2 = m^2 \sigma_x^2$$

$y = a / x$

$$\sigma_Y = \frac{|a|\sigma_x}{\bar{x}^2} \; ; \; \sigma_Y^2 = \frac{a^2}{\bar{x}^4} \sigma_x^2$$

$y = \ln x$

$$\sigma_Y = \frac{\sigma_x}{\bar{x}} \; ; \; \sigma_Y^2 = \frac{\sigma_x^2}{\bar{x}^2}$$

3. Error propagation for a given function

Average of several values with different uncertainties (σ_i):

$$\bar{x} = \frac{\sum_i x_i / \sigma_i^2}{\sum_i 1/\sigma_i^2}$$

$$\frac{1}{\sigma^2} = \sum_i \frac{1}{\sigma_i^2}$$

Numerical example

The example is identical to the one given above. The only (important!) difference is that now each value was itself obtained from several independent experiments (even through different techniques), so that the uncertainty intervals affecting each value are provided. Each value and associated uncertainty (standard deviation of the mean) were calculated as described in the example above. What is the mean value and the uncertainty in this case?

N	X / kJ mol^{-1}	σ	X / σ^2	1 / σ^2
1	366.3	8	5.72	0.01563
2	372.1	8	5.81	0.01563
3	375.0	2.9	44.59	0.11891
4	374.3	4	23.39	0.06250
5	369.3	4	23.08	0.06250
6	372.4	4	23.28	0.06250
7	369.8	8	5.78	0.01563

Sum of X/σ^2	131.7
Sum of $1/\sigma^2$	0.353
X_{mean}	372.7
Variance	2.83
Sigma	1.68

The result is 372.7±1.7 kJ mol^{-1}.

4. Linear regression

$$y = mx + b$$

$$m = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$b = \overline{y} - m\overline{x}$$

Regression coefficient:

$$r = \frac{\sum_{i=1}^{n}\left[(x_i - \overline{x})(y_i - \overline{y})\right]}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Variance of the slope:

$$s_m^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2-\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{(n-2)\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2} \quad \text{or} \quad \frac{s_m}{|m|}=\sqrt{\frac{r^{-2}-1}{n-2}}$$

Variance of the intercept:

$$s_b^2 = s_m^2\left(\sum_{i=1}^{n}x_i^2\right)/n$$

Uncertainty of $y_0 = mx_0 + b$, for a given x_0:

$$s_{y_0}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]^2}{(n-2)\sum_{i=1}^{n}(x_i-\bar{x})^2}\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

Uncertainty of $x_0 = (y_0 - b)/m$, for a given y_0:

$$s_{x_0}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]^2}{(n-2)m^2\sum_{i=1}^{n}(x_i-\bar{x})^2}\left[\frac{1}{n}+\frac{1}{l}+\frac{(y_0-\bar{y})^2}{m^2\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

Standard deviation of the fit:

$$\sigma = \sum_i (mx_i + b - y_i)/(n-2)$$

$$F = \frac{\sigma^2}{\sigma_y^2} \quad \text{where} \quad \sigma_y \approx s_y = \frac{\sum_i\sum_j\left(y_{i,j}-\bar{y}_i\right)^2}{\sum_i n_i - 1}; \qquad \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}y_{i,j}$$

Better fits yield small *F*-values. The *F* value calculated for a given fit must be smaller than *F* tabulated for $N-2$ degrees of freedom (where N is the number of data points) and a given probability.

Uncertainy intervals:

$$y = (m \pm \sigma_m \times t)x + b \pm \sigma_b \times t$$

where $\sigma_m = s_m$ and $\sigma_b = s_b$ are the standard deviations of the slope and the intercept, respectively, and *t* is student's factor for $N - 2$ degrees of freedom (where N is the number of data points) and a given probability.

Numerical example

According to the Benson scheme, the standard enthalpies of formation of any *n*-alkane in the gas phase, can be calculated using the group terms [C–(H)$_3$C] and [C–(H)$_2$(C)$_2$]:

$$\Delta_\mathrm{f} H^\mathrm{o}\left[\mathrm{CH_3(CH_2)_n CH_3, g}\right] = 2\left[\mathrm{C - (H)_3(C)}\right] + \mathrm{n}\left[\mathrm{C - (H)_2(C)_2}\right]$$

In other words the enthalpies of formation of $CH_3(CH_2)_nCH_3$ vary linearly with n. This correlation can be used to derive the above Benson terms. This example illustrates how.

n	$\Delta_fH^o(CH_3(CH_2)_nCH_3,g)$ / kJ mol^{-1}	error
0	-83.8	0.3
1	-104.7	0.5
2	-125.7	0.6
3	-146.9	0.8
4	-166.9	0.8
5	-187.6	1.3
6	-208.5	1.3
7	-228.2	0.6
8	-249.5	1.3
9	-270.8	2.5
10	-289.4	2.1

The regression analysis using the above formulae yields:

$$\Delta_\mathrm{f} H^\mathrm{o}\left[\mathrm{CH_3(CH_2)_n CH_3, g}\right] = 2 \times (-84.3636 \pm 0.3483 \times t) + (-20.61818 \pm 0.058869 \times t) \times \mathrm{n}$$

where *t* is student's factor for 9 degrees of freedom and for a given probability (e.g. *t* = 2.262 for 95% probability).

The calculation of $\sigma_m = s_m$ and $\sigma_b = s_b$ can be easily done with any spreadsheet. With Microsoft Excel (Tools, Data Analysis, Regression). The value of σ_m appears under "Standard Error" and "X Variable"; the value of σ_b appears under "Standard Error" and

“Intercept” (the value for the student’s factor may be calculated by using the statistical function “TINV”, as described above). It is, however, fairly easy to make mistakes in the selection of data. The above example can be used to test the package.