Basic Analysis of Data

Department of Chemical Engineering
Prof. Geoff Silcox
Fall 2008

1.0 Reporting the Uncertainty in a Measured Quantity

At the request of your supervisor, you have ventured out into the plant and have measured the yield for a compound in a plug flow reactor in a series of ten experiments. The conditions in the plant are always varying, but you have tried to make your measurements during periods of relative stability. Your results (Box et al., 1978) are tabulated in Table 1 in the order in which they were collected. What is the best way to present the results to your supervisor? How can you estimate the uncertainty in the yield? Your first thought is to give her Table 1.

Table 1 Reactor Yield

Run No.	Yield
1	89.7
2	81.4
3	84.5
4	84.8
5	87.3
6	79.7
7	85.1
8	81.7
9	83.7
10	84.5

You also have a dim recollection of some statistical quantities. You dig out your statistics notes and find the following definitions. For a set of n observations, y_i,

$$\text{sample average} = \bar{y} = \sum y_i / n \tag{1}$$

$$\text{sample variance} = s^2 = \sum \left(y_i - \bar{y}\right)^2 / \left(n-1\right) \tag{2}$$

$$\text{sample standard deviation} = s = \sqrt{\sum \left(y_i - \bar{y}\right)^2 / \left(n-1\right)} \tag{3}$$

$$\text{variance of sample average} = V(\bar{y}) = s^2 / n \tag{4}$$

$$\text{standard deviation of sample average} = u = s/\sqrt{n} \qquad (5)$$

The sample average, standard deviation, and standard deviation of the average are given in Table 2 for the data in Table 1. These quantities, with the exception of u, were calculated using the built in functions in Excel.

Table 2 Statistical Quantities for Data in Table 1

Statistic	Value
$\bar{y}$	84.24
s	2.902
u	0.9176

You want to report the average value of the yield but you would also like to provide an estimate of the uncertainty in the average. Two terms are central to any discussion of uncertainty analysis (Kline and McClintock, 1953). An *error* is defined as the difference between the true and the observed value of some quantity. Only estimates of the error can be made because the true value is unknown. The *uncertainty* is a possible value that an error might have. The uncertainty is an estimate of the experimental error. The definitions of uncertainty and error are not the same as the definitions of accuracy and precision.

Uncertainty analysis is useful for (i) estimating the possible values that an error may have, (ii) determining if unaccounted errors must be included in an estimation of results, and (iii) designing experiments to minimize errors. Uncertainty analysis generally considers four different types of errors: (i) errors from scale interpolation, (ii) errors from time-wise jitter, (iii) bias errors, and (iv) calibration errors.

Uncertainties are always determined for a particular confidence level. A confidence level of 95 percent is commonly used in engineering. A 95 percent confidence level means that 95 percent (or 19 out of 20) of the measurements will fall within the uncertainty interval, i.e., the odds are 20:1. The higher the confidence level, the larger the uncertainty.

An uncertainty is meaningless unless it includes a confidence level. For example, the measured value of y should be reported as

$$y = \bar{y} \pm \delta y \quad (95\ \%\ \textit{confidence level}) \qquad (6)$$

where $\bar{y}$ is the best estimate of the yield (the sample average) and δy is the uncertainty interval. For errors that follow a normal probability distribution, 95 percent of the measurements will fall within plus or minus 1.96 standard

deviations of the best estimate. Hence, the uncertainty for a 95 percent confidence level for the data in Table 1 is given by

$$\delta y = 1.96u \tag{7}$$

where u is calculated from Eqn. 5. In your report, you need to give Tables 1 and 2 and you should report the yield as

y = 84 ± 2 (confidence level 95 %)

In general, only one significant figure should be used for the estimated uncertainty. Uncertainties for confidence levels ranging from 0.80 to 0.99, based on the normal distribution, are given in Table 3.

Table 3 Uncertainty as a Function of Confidence Level for Normally Distributed Errors

Confidence Level	Uncertainty, δy
0.80	1.28u
0.90	1.64u
0.95	1.96u
0.99	2.58u

The assumption made above in the calculation of uncertainty is that the errors are normally distributed. For finite samples this is not strictly correct and uncertainties should actually be estimated using the t-distribution (Box et al., 1978; Gonick and Smith, 1993). The t-distribution is more spread out than the normal distribution and hence the uncertainties based on it are larger. We will not pursue this further in these notes.

2.0 Propagating Uncorrelated Uncertainties

In the situation described above, we had 10 values of the reactor yield and we were able to estimate the uncertainty interval on the sample average using (5), (6), and (7). In many cases we do not have this opportunity because the experiments are too costly or time consuming. We may be limited to a single sample. But we often know the uncertainty level in the variables that are used to calculate a desired quantity and we can use these to estimate the uncertainty in the final result. The equations below allow you to estimate the uncertainty in a single sample experiment and they give you a way of propagating the uncertainties through a calculation.

In what follows, σ is the estimated uncertainty on each quantity. From the equations outlined below, you can estimate the uncertainty in a dependent variable given the uncertainties in the independent variables. If

$$a = b + c$$

or

$$a = b - c$$

then

$$\sigma_a^2 = \sigma_b^2 + \sigma_c^2 \tag{8}$$

If

$$f = xy$$

or

$$f = x/y$$

then

$$\left(\frac{\sigma_f}{f}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2 \tag{9}$$

In general,

$$f = f(x_1, x_2, \ldots, x_n)$$

and

$$\sigma_f^2 = \sum \left(\frac{\partial f}{\partial x_i}\sigma_i\right)^2 \tag{10}$$

Equations 8, 9, and 10 only apply for small uncertainties. A more general way to propagate uncertainties is to use a numerical approach as follows (Lyons, 1991). This approach applies no matter how large the uncertainties.

(i) Set all x_i equal to their measured values and calculate f. Call this f_o.

(ii) Find the n values of f defined by

$$f_i = f(x_1, x_2, \ldots, x_i + \sigma_i, \ldots, x_n) \tag{11}$$

(iii) Obtain σ_f from

$$\sigma_f^2 = \sum \left(f_i - f_o \right)^2 \tag{12}$$

If the uncertainties are small this should give the same result as (10). If the uncertainties are large, this numerical approach will provide a more realistic estimate of the uncertainty in f. The numerical approach may also be used to estimate the upper and lower values for the uncertainty in f because the f_i in (11) can be calculated with $x_i+\sigma$ replaced by $x_i-\sigma$.

For example, consider the measurement of the velocity of an air stream inside a duct using a pitot-static tube (Kline and McClintock, 1953). If c is the air velocity, Δp is the pressure difference, and p_a and T_a are the pressure and temperature of the air, then the gas velocity is given by

$$c = \sqrt{\frac{2(\Delta p) R T_a}{p_a}} \tag{13}$$

Suppose that T_a is measured with a calibrated, mercury-in-glass thermometer, p_a with a Bourdon gage, and Δp with a U-tube manometer. The readings are

$\Delta p = 8.0 \pm 0.1$ in. H_2O (confidence level 95 %)
$T_a = 527.1 \pm 0.2°R$ (confidence level 95 %)
$p_a = 14.7 \pm 0.3$ psia (confidence level 95 %)

Evaluation of the derivatives, $\partial c/\partial x_i$, and substitution in (10) gives (after taking the square root)

$$\sigma_C = \left\{ \frac{1}{2} \frac{RT_a}{(\Delta p) p_a} \left(\sigma_{\Delta P} \right)^2 + \frac{1}{2} \frac{(\Delta p) R T_a}{p_a^3} (\sigma_{Pa})^2 + \frac{1}{2} \frac{R(\Delta p)}{T_a p_a} (\sigma_{Ta})^2 \right\}^{1/2} \tag{14}$$

Dividing (14) by (13) simplifies (14) by making it dimensionless.

$$\frac{\sigma_C}{c} = \left\{ \left[\frac{1}{2} \frac{\sigma_{\Delta P}}{\Delta p} \right]^2 + \left[\frac{1}{2} \frac{\sigma_{Pa}}{p_a} \right]^2 + \left[\frac{1}{2} \frac{\sigma_{Ta}}{T_a} \right]^2 \right\}^{1/2} \tag{15}$$

Substituting the temperature and pressures in (15) gives

$$\frac{\sigma_C}{c} = \frac{1}{2} \left[1.56x10^{-4} + 4.16x10^{-4} + 1.44x10^{-7} \right]^{1/2} = 0.01197 \; or \;\; 1.2\% \tag{16}$$

From (16) it is clear that improving the measurement of T_a will have little impact on σ_C/c. The largest effect on σ_C/c will come from improving the measurement of

p_a. This could be accomplished by using a pressure transducer or manometer instead of a Bourdon gage.

Now try the same calculation using the spread sheet method. The dimensionless form of (12) is (after taking the square root)

$$\frac{\sigma_f}{f_0} = \left[\sum \left(\frac{f_i}{f_0} - 1 \right)^2 \right]^{1/2} \tag{17}$$

The propagated fractional uncertainties using (15) and (17) are compared in Table 4.

A further advantage of the numerical approach is that it can be used with simulations. In other words, the function f in (12) could be a complex mathematical model of a distillation column and f might be the mole fraction or flow rate of the light component in the distillate.

Table 4 Uncertainties in Gas Velocity Calculated from (15) and (17)

Equation Used	σ_f/f_0
(9)	0.011968
(12) with $+\sigma$	0.011827
(12) with $-\sigma$	0.012113

3.0 Estimating Uncertainties

The uncertainty cannot always be calculated from a set of data, like that in Table 1, using (5), (6), and (7). Instead, you must rely on engineering judgement and experience. Some rough guidelines for estimating the uncertainty with a 95 percent confidence level follow. For analogue scales, use 1/2 the smallest scale division and for digital scales, use 1/2 the value of an increment for the least significant digit. For a reading that is jittery, use the range of the variable that includes roughly 95 percent of the readings.

4.0 Combining Results from Several Sets of Measurements

You may want to combine results from two groups or laboratories. The combination is weighted using the uncertainties as follows (Lyons, 1991; Young, 1963)

$$a = \sum \left(a_i / \sigma_i^2 \right) / \sum \left(1 / \sigma_i^2 \right) \tag{18}$$

and the uncertainty on the combined result is given by

$$1/\sigma^2 = \sum\left(1/\sigma_i^2\right) \tag{19}$$

For example, two labs are reporting the speed of light as (Young, 1963)

c_1 = 299,774 ± 2 km/s (confidence level 95 %)
c_2 = 299,778 ± 4 km/s (confidence level 95 %)

From (18) the most probable value of c is 299,775 km/s and from (19) the uncertainty is 2 km/s.

5.0 Linear Least Squares Fitting

A particular theory will frequently tell us that data should plot as a straight line. This is a powerful way of testing whether data agree with theory. The following example will use this technique to see if a mixing tank can be assumed perfectly mixed. In chemical process control, for example, we frequently assume simplified models for complex processes and design our controllers based on those simplified models. Perfect mixing is a common simplifying assumption.

The mixing tank is sketched in Fig. 1. The perfect mixing hypothesis will be tested using temperature data (Marlin, 1995). The volume of the tank (V) is 2.7 m^3, the volumetric flow rate (F) is 0.71 m^3/min., and the initial steady temperature of the fluid (T_0) is 103.5 C. At the beginning of the experiment, the inlet temperature (T_i) is suddenly changed from 103.5 to 68 C and the data shown in Table 5 are collected. We will test our perfect mixing assumption by seeing if a model of the process, developed with this assumption, agrees with the data.

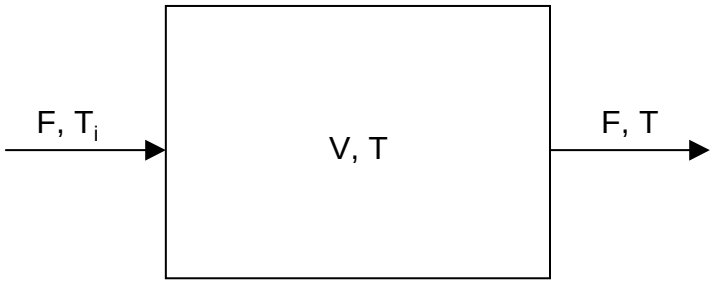

Figure 1 Schematic of mixing tank of volume V with volumetric flow rate F. The temperature of the fluid at the outlet of the tank, T, is the same as the temperature of the entire contents of the tank, assuming that the contents are perfectly mixed.

Table 5 Temperature data for mixing tank shown in Fig. 1 (Marlin, 1995).

t, min.	T, C
0	103.5
0.4	102
1.2	96
1.9	91
2.7	87
3.4	84
4.2	81
5	79
6.5	76
8.5	73

An energy balance over the tank, assuming that it is perfectly mixed, gives a differential equation for the rate of change of temperature with time,

$$\rho c V \frac{dT}{dt} = F \rho c (T_i - T) \tag{20}$$

where the density and heat capacity of the fluid are ρ and c. The initial condition for (20) is

$$T(0) = T_0 \tag{21}$$

We can simplify (20) by defining

$$\begin{aligned} \theta &= T - T_i \\ \theta_0 &= T_0 - T_i \\ \tau &= \frac{V}{F} \end{aligned} \tag{22}$$

The mean residence time of a particle of fluid in a perfectly mixed tank is τ. For our problem,

$$\tau = \frac{V}{F} = \frac{2.7}{0.71} = 3.8 \text{ min.} \tag{23}$$

The solution to (20) and (21) is

$$\ln\left(\frac{\theta}{\theta_0}\right) = -\frac{t}{\tau} \tag{24}$$

or

$$\frac{\theta}{\theta_0} = \frac{T - T_i}{T_0 - T_i} = e^{-t/\tau} \tag{25}$$

Now we are ready to test our model against the data. If we plot -ln(θ/θ0) vs. t, the slope of the data should be 1/τ and the intercept should be zero, provided our assumption of perfect mixing is correct. Using the Regression Analysis Tool in Excel and the data in Table 5 gives the linear least squares fit shown in Fig. 2.

The slope and intercept are 0.235 ± 0.003 min^{-1} and -0.02 ± 0.01. The Regression Analysis Tool in Excel estimates the uncertainties (it calls them standard errors) and allows you to specify the confidence level. I used a confidence level of 95 percent. The time constant we calculate from the least squares fit of the data is 4.3 min. versus the value of 3.8 min. we calculated from V and F. We conclude that the assumption of perfect mixing is probably adequate for modeling this system.

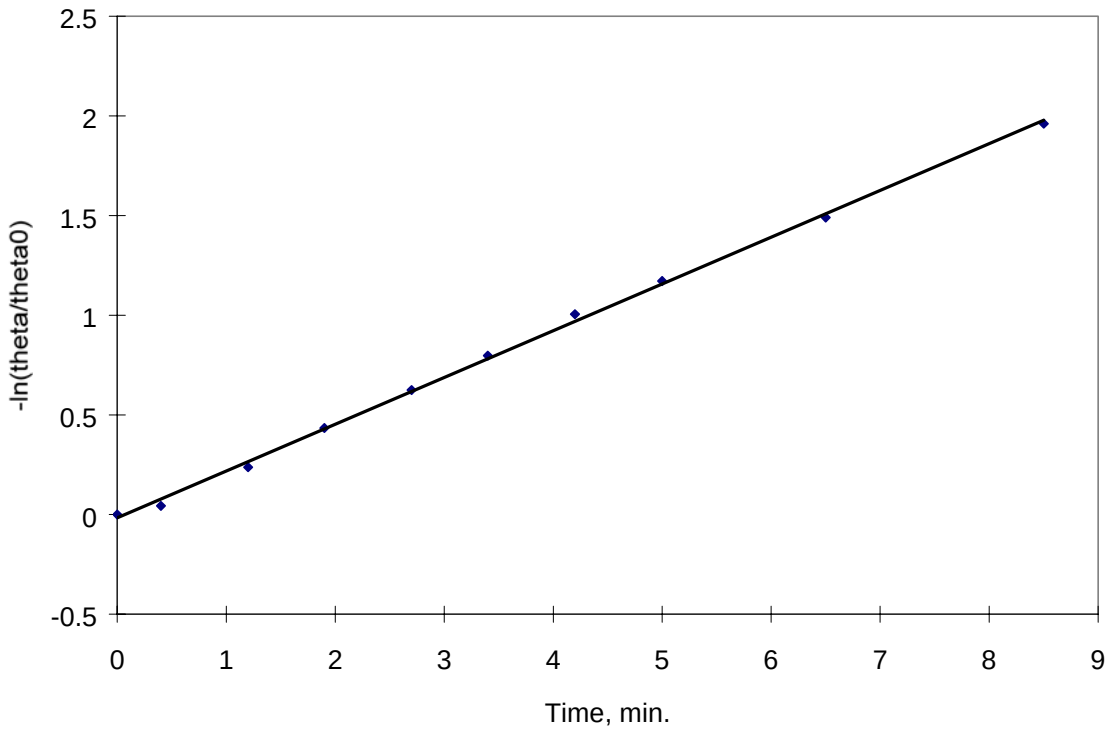

Figure 2 Linear least squares fit of Eqn. 19 to the data of Table 4.

6.0 Nonlinear Least Squares Fitting

The design of a PID controller for a process makes use of correlations that generally require three input variables: the steady state process gain (K_P), the process time constant (τ), and the process dead time (α). We can obtain these variables by performing experiments on the process and then fitting a simplified model to the data. Typical data from an experiment on a reactor are shown in Fig. 3 (Marlin, 1995).

In the experiment, the input is the valve position (% open) and the output is the temperature (C) of the reactor at some location. For the results shown in Fig. 3, the valve position was stepped from 30 to 38 percent open. The temperature of the reactor rose about 6 degrees in response to this change. Our tuning parameters, K_P, α, and τ, are obtained by fitting the following model to the data,

$$y(t) = K_P \delta \left(1 - e^{(t-\alpha)/\tau}\right), \; t \geq \alpha \tag{26}$$

where δ is the change in the valve position and y(t) is the resulting change in

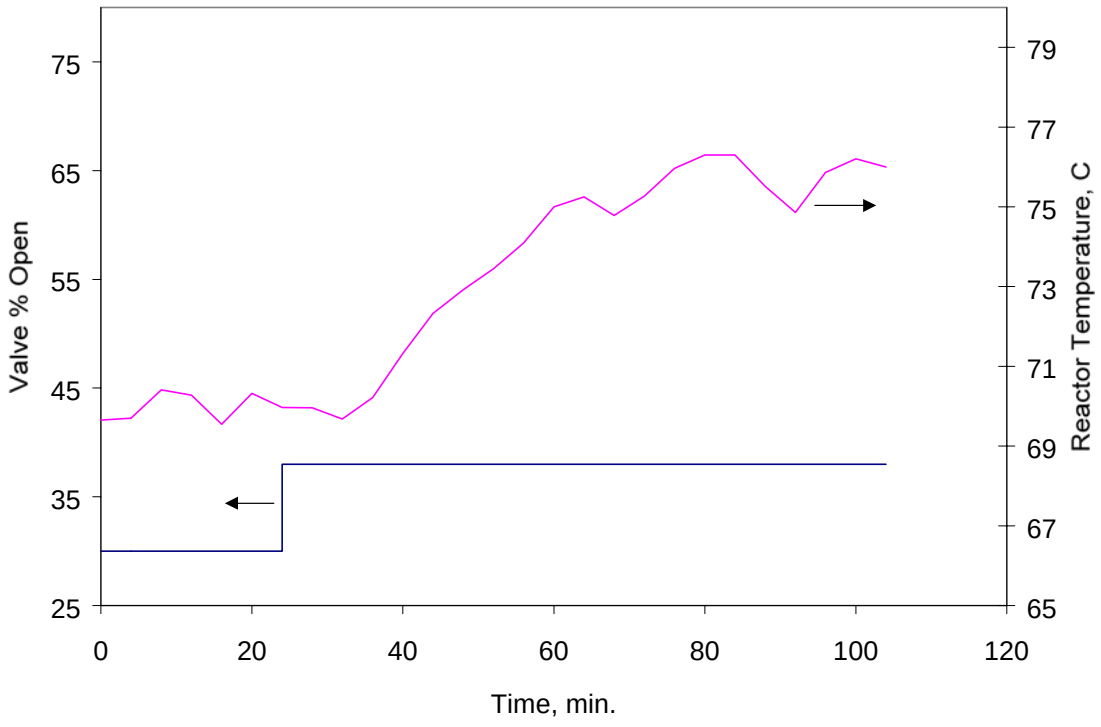

Figure 3 Reactor data used to obtain K_P, α, and τ.

temperature. For our experiment, δ is 8 percent open. We need to obtain a least squares fit of (26) to the data in Fig. 3. We can accomplish this using statistical software packages. We can also use the Solver Tool in Excel to obtain the fit shown in Fig. 4, where a time of zero minutes corresponds to the time at which the valve position is changed in Fig. 3. The spreadsheet used to perform the least squares calculations is shown in Fig. 5.

The calculation involves minimizing the sum of the squared residuals. In the spreadsheet you calculate the quantity

$$sum\ of\ squared\ residuals = \sum \left(y_i - y_{i,the} \right)^2 \qquad (27)$$

and minimize its value using the Solver. The Solver adjusts the values of K_P, α, and τ to find the minimum. In (27), the y_i are the measured changes in the reactor temperature and the $y_{i,the}$ are the changes calculated from (26). The Solver determines values of K_P, α, and τ of 0.764 C/% open, 12.4 min., and 16.3 min. This approach provides a convenient way of performing nonlinear least squares fitting but it provides no estimates of the uncertainties on K_P, α, and τ. Equation 12 provides a way of estimating these uncertainties if the uncertainties on the individual temperatures are known.

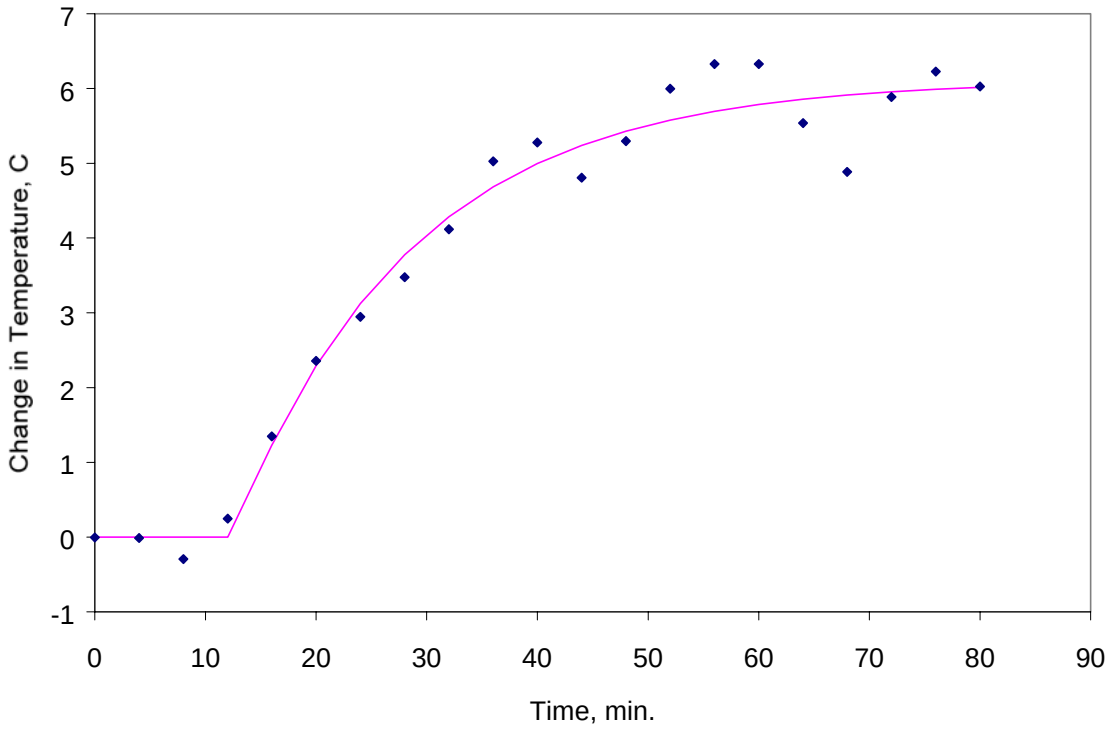

Figure 4 Least squares fit of (20) to the reactor data shown in Fig. 3.

The calculation involves minimizing the sum of the squared residuals. In the spreadsheet you calculate the quantity

$$sum\ of\ squared\ residuals = \sum \left(y_i - y_{i,the} \right)^2 \qquad (27)$$

and minimize its value using the Solver. The Solver adjusts the values of K_P, α, and τ to find the minimum. In (27), the y_i are the measured changes in the reactor temperature and the $y_{i,the}$ are the changes calculated from (26). The Solver determines values of K_P, α, and τ of 0.764 C/% open, 12.4 min., and 16.3 min. This approach provides a convenient way of performing nonlinear least squares fitting but it provides no estimates of the uncertainties on K_P, α, and τ. Equation 12 provides a way of estimating these uncertainties if the uncertainties on the individual temperatures are known.

Data

Initial T, C	69.974
Delta, %	8

Guessed quantities (fitting parameters)

Kp, C/%	0.764
tau, min.	16.280
theta, min.	12.348

Time, min	Adjusted time, min	Input, % open	Output, C	Output', C	Theory	Resid	Resid^2
0		30	69.65	-0.324			
4		30	69.7	-0.274			
8		30	70.41	0.436			
12		30	70.28	0.306			
16		30	69.55	-0.424			
20		30	70.32	0.346			
23.999		30	69.97	-0.004			
24	0	38	69.97	-0.004	0.000		
28	4	38	69.96	-0.014	0.000		
32	8	38	69.68	-0.294	0.000		
36	12	38	70.22	0.246	0.000		
40	16	38	71.32	1.346	1.229	0.118	0.014
44	20	38	72.33	2.356	2.293	0.063	0.004
48	24	38	72.92	2.946	3.125	-0.179	0.032
52	28	38	73.45	3.476	3.776	-0.300	0.090
56	32	38	74.09	4.116	4.286	-0.169	0.029
60	36	38	75	5.026	4.684	0.342	0.117
64	40	38	75.25	5.276	4.996	0.281	0.079
68	44	38	74.78	4.806	5.239	-0.433	0.187
72	48	38	75.27	5.296	5.430	-0.133	0.018
76	52	38	75.97	5.996	5.579	0.417	0.174
80	56	38	76.3	6.326	5.696	0.631	0.398
84	60	38	76.3	6.326	5.787	0.540	0.291
88	64	38	75.51	5.536	5.858	-0.322	0.103
92	68	38	74.86	4.886	5.914	-1.027	1.056
96	72	38	75.86	5.886	5.958	-0.071	0.005
100	76	38	76.2	6.226	5.992	0.235	0.055
104	80	38	76	6.026	6.018	0.008	0.000
				(sum of residuals)^2 =			2.653

Figure 5 Excel spreadsheet used to fit (26) to the data in Fig. 3.

7.0 References and Suggested Reading

Box, George E. P., William G. Hunter, and J. Stuart Hunter. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York, NY: John Wiley & Sons, 1978.

Gonick, L. and W. Smith. *The Cartoon Guide to Statistics*. New York, NY: HarperPerennial, 1993.

Kline, S. J. and F. A. McClintock. "Describing Uncertainties in Single-Sample Experiments," *Mechanical Engineering*, January 1953, 3.

Lyons, Louis. *A Practical Guide to Data Analysis for Physical Science Students*. Cambridge: Cambridge University Press, 1991.

Marlin, Thomas E. *Process Control: Designing Processes and Control Systems for Dynamic Performance*. New York, NY: McGraw-Hill, 1995.

Young, H. D. *Statistical Treatment of Experimental Data*. New York, NY: McGraw-Hill, 1962.