

EXPERIMENTAL ERRORS AND DATA ANALYSIS

by J.C. de Paula

1. RANDOM AND SYSTEMATIC ERRORS

Every experimental result is subject to error.¹ One can attempt to minimize errors but cannot eliminate them completely. Experimental errors fall under two categories: random or systematic.

Random errors arise from natural limitations of making physical measurements. For example, repeated measurements of the same property often differ even if they are performed on a single instrument that is calibrated and operated properly. Such variations establish the *precision* of the measurement. The precision is also referred to as the *reproducibility*.

Systematic errors arise from blunders in the measuring process. For example, an instrument that is not operating properly is likely to give erroneous results. The measurement lacks *accuracy*. It is even possible that repeated measurements with this broken instrument will give reproducible results (high precision), but every one of them will deviate from the true value (low accuracy). Thus, accuracy and precision are not related to each other.

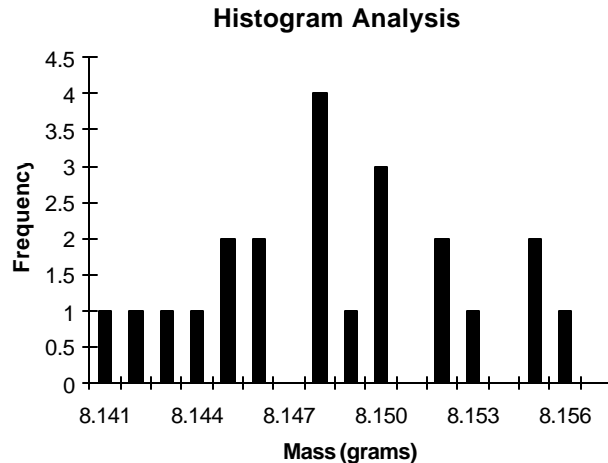
2. STATISTICS OF DATA ANALYSIS

The Normal Distribution. Let us turn our attention to precision, since every set of measurements will be subject to random errors, no matter what the degree of accuracy. Let us consider the data below, obtained by measuring the mass of a metal strip 24 times with the same instrument.

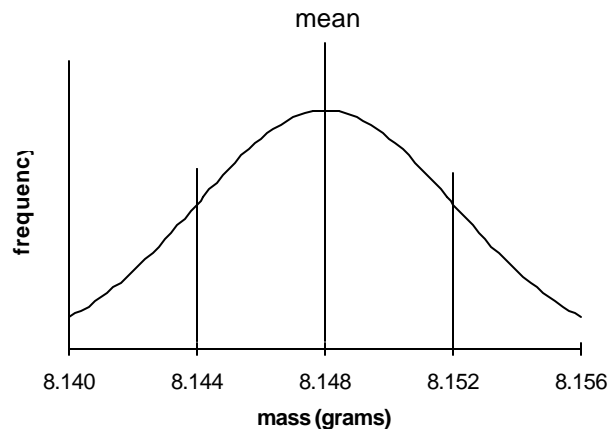
Mass Measurements (in Grams) for a Metal Strip			
8.150	8.145	8.148	8.145
8.155	8.155	8.156	8.152
8.142	8.143	8.144	8.148
8.150	8.153	8.152	8.150
8.140	8.149	8.146	8.146
8.148	8.145	8.148	8.150

¹ For a more thorough mathematical treatment of errors and error analysis, please consult: Shoemaker, D.P, Garland, C.W., and Nibler, J.W. (1989) in "Experiments in Physical Chemistry", 5th Edition, McGraw Hill, New York, Chapter II.

We can visualize the reproducibility of this set of measurements by plotting the data as a *histogram*, which shows how often (frequency) a given value was obtained in the set. The histogram for this experiment is shown below.



We see that the value that occurred most often was 8.148 (four times), and that values that deviated by too much or too little from 8.148 occurred infrequently. If we were to conduct a very large number of measurements on the metal strip, we would have obtained a histogram whose shape resembles a bell, as shown below:



This bell-shaped curve denotes the *normal probability distribution* for a large number of mass measurements. The exact shape of the normal distribution is characterized by two parameters: the *mean value*, \bar{x} , and the *standard deviation*, s :

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} \quad (1)$$

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}} \quad (2)$$

For a large number of measurements, the mean value is also the most probable value, as shown on the plot above. The standard deviation is a measure of the breadth of the curve: the larger the standard deviation, the broader the distribution. Put differently, *the less precise the measurement, the broader the distribution and the larger the standard deviation.*

The significance of the standard deviation is as follows: the probability of finding a result x_i falling between $\bar{x} - s$ and $\bar{x} + s$ is 68.30 %. Put differently, 68.30 % of all observations will fall within one standard deviation of the mean. For the example we have been considering, where $\bar{x} = 1.848$ grams and $s = 0.004$ grams, the odds are 16 out of 24 that *an individual mass measurement* of the metal strip will fall between 1.844 grams and 1.852 grams. With reference to the normal distribution depicted above, this statement says that the area under the bell-shaped curve between the limits denoted by the dashed lines represents 68.30 % of the total area.

Your audience may not be impressed by these odds. We can state the results differently, however. It turns out that, in a normal distribution, 95.5 % of the observations will fall within the limits of $2s$. In other words, the odds are 19 out of 20 that an individual measurement of the mass will be within $2s = 0.008$ grams of the mean value of 1.844 grams. You are 95.5 % confident that a new individual measurement of the mass will give a result in the range 1.848 ± 0.008 grams. This statement determines the *confidence interval* of your experiment. You are telling the audience that if a new measurement of the mass falls outside the range 1.840 to 1.856 grams, the authenticity of the new result is doubtful.

So far, we have dealt with the precision of a single measurement, based on the standard deviation of the entire normal distribution curve. How about the precision (or uncertainty) of the mean value \bar{x} ? The standard deviation of the mean value is given by

$$s_{\bar{x}} = \frac{s}{\sqrt{N}} \quad (3)$$

Using the 95.5 % confidence interval, the mean value should then be reported as $\bar{x} \pm 2s_{\bar{x}}$. Let us go back to our example, where $\bar{x} = 1.848$ grams and $2s_{\bar{x}} = 0.001$ grams. By reporting the average as 1.848 ± 0.001 grams, we are saying that if 20 different students were to measure the mass of the metal strip 24 times, 19 out of 20 of the measured *mean values* will fall within 1.847 and 1.849 grams.

Working with Small Data Sets. In real life, one seldom has so many measurements in a set as to see the bell-shaped curve in all of its glory. There is the possibility that your sample is not representative of the bell-shaped curve. In such instances, we first calculate a mean and a standard deviation according to Equations (1)-(2) and then calculate a *confidence limit*, λ . The confidence limit defines the range on both sides of the mean within which the true value can be expected to be found with a given level of confidence.

$$\lambda = \pm \frac{ts}{\sqrt{N}} \quad (4)$$

where the value of t depends on the confidence level. Note that Equations (3) and (4) differ only by the factor t . A brief table of t values follows²:

N	t (90%)	t (95%)	t (99%)
2	6.31	12.7	63.7
3	2.92	4.30	9.92
4	2.35	3.18	5.48
5	2.13	2.78	4.60
6	2.01	2.57	4.03
7	1.94	2.45	3.71
9	1.86	2.31	3.36
11	1.81	2.23	3.17
16	1.75	2.13	2.95
21	1.72	2.09	2.85
31	1.70	2.04	2.75
∞	1.64	1.96	2.58

If we report a result as " $\bar{x} \pm \lambda$ ", it is also necessary to report the confidence level and the number of measurements in the data set. It is not necessary to repeat units in the presentation of the result. For example, we write 1.848 ± 0.001 grams (95%, $N = 24$) and not as 1.848 grams ± 0.001 grams (95%, $N = 24$).

In this course, you will often be asked to report your results as a mean value of two or more measurements *with* the all important uncertainty. Without the uncertainty, you cannot communicate to the reader the degree of reproducibility of your experiment.

3. PROPAGATION OF ERRORS

It is bad enough that random errors cannot be eliminated and that is why we need to repeat experiments in the laboratory. Even worse is the fact that errors *propagate* as you perform your calculations. Let us consider the measurement of density as an example. In order to report the density of a substance, you must make two measurements - one of mass and one of volume - each of which carries a different precision. The question is, what is the error associated with the ratio of mass over volume (i.e., the density)? The uncertainty of the result of a mathematical operation between two measured values may be calculated with simple formulas, as shown below. The results apply to the propagation of both s and $s_{\bar{x}}$.

² Halpern, A. in "Experimental Physical Chemistry: A Laboratory Textbook", 2nd Edition, Prentice Hall, Upper Saddle River, NY, p. 6.

Addition and Subtraction. If you add or subtract the values $x \pm s_x$ and $y \pm s_y$, the uncertainty of the result is given by:

$$s_{x \pm y} = \sqrt{s_x^2 + s_y^2} \quad (5)$$

Products and Quotients. If you multiply or divide $x \pm s_x$ by $y \pm s_y$, the uncertainties are given by:

$$s_{xy} = \sqrt{s_x^2 y^2 + s_y^2 x^2} \quad (6)$$

$$s_{x/y} = \frac{1}{y^2} \sqrt{s_x^2 y^2 + s_y^2 x^2} \quad (7)$$

Natural Logs. The uncertainties of $\ln(x \pm s_x)$, $\ln[(x \pm s_x) + (y \pm s_y)]$, and $\ln\left(\frac{x \pm s_x}{y \pm s_y}\right)$ are given by:

$$s_{\ln x} = \frac{s_x}{x} \quad (8)$$

$$s_{\ln x+y} = \frac{(s_x^2 + s_y^2)^{1/2}}{x + y} \quad (9)$$

$$s_{\ln x/y} = \left(\frac{s_x^2}{x^2} + \frac{s_y^2}{y^2} \right)^{1/2} \quad (10)$$

An error analysis of the type described by Equations (5)-(10) is labor intensive but necessary. It allows you to determine the weak link in an experiment: the number that limits the reliability of your results. You should identify this specifically in your laboratory reports.

4. REJECTION OF DATA

Let us now consider systematic errors. Every now and then, you will come across an experimental result that simply does not seem to make sense. For example, after making only five measurements of our metal strip, you obtain the following results (in grams):

8.148 8.145 8.156 8.149 8.177

The last measurement seems a bit off, and you may be tempted to throw it out of the set on aesthetic grounds alone. However, **you must never throw out a result from a data set unless you have a statistical or chemical reason to do so.**

Statistically speaking, we are asking the question: does the measurement of 8.177 grams belong to the same normal distribution as the other four measurements? There are two ways to answer this question.

Let us first take the “common sense” approach. The last measurement may not belong to the same distribution as the other four because of a *systematic error* that you tracked down upon close examination of your laboratory notebook. That is, you know you goofed! No problem, it happens to all of us! In that case, there is no reason for you to keep the number in your data set, although you may want to make another measurement in its stead just in case. The bottom line is, however, that you still must document this blunder to your professor or supervisor. This is but one of the many reasons why you must keep complete and accurate records of your laboratory activities *in your laboratory notebook*.

A systematic error may not be apparent even after some detective work. In such cases, it is possible, though not always advisable, to use statistical methods to reject an observation. If we had *a very large data set*, then we could calculate $\bar{x} \pm 2s_{\bar{x}}$, and then determine if the measurement in question falls outside the confidence interval. However, our data set is very small ($N < 10$), so that the standard deviation alone is not a good criterion for rejection.

Statisticians have devised many rejection tests for the detection of non-random errors. We will describe only one - the Q test - which works well in cases where $3 < N < 10$.

In order to test the value of 8.177 grams, we must calculate the so-called Q_{calc} value for this observation. In general, the value of Q_{calc} is given by

$$Q_{\text{calc}} = \frac{\text{absolute value of the gap between the suspect value and the value closest to it}}{\text{range of values}} \quad (11)$$

To calculate Q_{calc} in our example, we display the data in increasing order of numerical value, then identify the suspect value and the value that is closest to it:

8.145	8.148	8.149	8.156	8.177
			<i>(value closest to suspect value)</i>	(suspect value)

Then,

$$Q_{\text{calc}} = \frac{|8.177 - 8.156|}{|8.177 - 8.145|} = 0.66$$

We now compare this Q_{calc} with a *critical value* Q_c . If $Q_{\text{calc}} > Q_c$, then the observation may be rejected. If $Q_{\text{calc}} < Q_c$, then we must keep the observation no matter how tempted we may

be to throw it out. The Q_c value depends on the confidence level and the number of observations in your set. A partial list follows:

Q_c (90% confidence)	0.94	0.76	0.64	0.56	0.51	0.47	0.44	0.41
N	3	4	5	6	7	8	9	10

Returning to our example, where $N = 5$, we see that $Q_{\text{calc}} = 0.66$ is indeed greater than $Q_c = 0.64$. Hence, we are justified in rejecting the observation. However, you must indicate in your report that the Q test was used at a 90 % confidence interval.

5. SIGNIFICANT FIGURES

You should review the topic of significant figures by consulting any General Chemistry textbook.³ Briefly, the number of significant digits is dictated by the precision of the instrument used to obtain the data. If we report a mass measurement as 1.840 grams, we state that we looked for data out to the third decimal place. There are four significant figures in the result. A report of 1.84 grams means that we did not look as far as the third decimal place; we stopped looking at the second decimal place. There are three significant figures in the result. Making a measurement of 1.840 grams requires a better balance than making a measurement of 1.84 grams.

The following rules are worth remembering:

- Zeros to the left of a non-zero digit are not significant.
- Zeros to the right of a non-zero digit are significant.
- When adding or subtracting two numbers, the result has as many digits to the right of the decimal point as the operand with the fewer significant figures to the right of the decimal point.
- When multiplying or dividing two numbers, the result has as many significant figures as the operand with the fewer significant figures.

In this course, you should report all experimental values with the correct number of significant figures. Also, you should use the rules above to make sure that your final answers (after calculations) have the correct number of significant figures.

6. LINEAR REGRESSION ANALYSIS: CONCEPTS

In many instances, simply plotting X-Y data does not reveal as much information as fitting the data pairs to an analytical expression. Theoreticians help experimentalists by describing a meaningful physical model with a mathematical expression. This expression may be simple, such as $PV = nRT$, or complex.

³ For example: Oxtoby, D. and Nachtrieb, N. (1996) in "Principles of Modern Chemistry", 3rd Edition, Saunders College Publishing, Fort Worth, pp. A-6 - A-7.

The experimentalist is lucky if the model can be reduced to a simple first-order polynomial of the form

$$y = mx + b \quad (12)$$

where y is an experimental value observed at a certain set of conditions described by x . This is, of course, the equation of a straight line with slope m and y-intercept b . By finding the best values of m and b that correlate y with x for all observations, the experimentalist can derive useful chemical and physical insight from the data.

Qualitatively, the fitting process may be summarized as follows:

- choose the equation that describes the model being tested. For example, assume a first-order polynomial:

$$y_{\text{theo}}(x) = mx + b \quad (13)$$

where y_{theo} is the theoretical value of the property y of the system for a set of parameters m and b (to be determined).

- define the deviation $d(x)$ as the difference between the theoretical value and the observed value of y at each x :

$$d(x) = y_{\text{obs}}(x) - y_{\text{theo}}(x) \quad (14)$$

- The $d(x)$ values will be either negative or positive, depending on whether the fit overestimates or underestimates y at each x . The square of $d(x)$, however, will always be positive. Moreover, $d(x)^2$ will be very small if the fit of the data to the equation is good at a particular value of x . Hence, **the best fit to the data is that in which the sum of all $d(x)^2$ is as small as possible.**
- Using differential calculus, it is possible to minimize the sum of $d(x)^2$ with respect to the fitting parameters, thereby arriving at the best-fit values of m and b . This method of fitting is thus called the *least-squares method*. It should be noted that the sum of $d(x)^2$ is seldom zero for real experimental data (i.e., the fit is seldom "perfect") because every experimental measurement carries uncertainties.

There are several software packages that are capable of performing the protocol described above. You may have used programs such as SigmaPlot, Harvard Graphics, Kaleidograph, CricketGraph, etc. The academic computer clusters around campus have CricketGraph installed on their Macintosh computers. You are more than welcome to use CricketGraph for some of the plotting tasks associated with this course. We will not show you how to use CricketGraph, however. We *will* show you how Microsoft EXCEL may be used to perform linear fits.

Assessing the Quality of a Fit. A good fitting routine not only calculates the best slope and intercept for a set of X-Y data; it also gives you a sense of how good the fit is. The slope and intercept values should be accompanied by a standard deviation value. The standard deviation lets you know how much confidence you should give to the value of the fit parameters. This uncertainty arises from the fact that the calculated line did not go through all of your experimental points. In

other words, the fit was not "perfect." For these reasons, you should always report the slope and intercept as the calculated value and their associated standard errors as

$$\begin{aligned} &\text{slope} \pm \text{std. error} \\ &\text{intercept} \pm \text{std. error} \end{aligned}$$

The *correlation coefficient* (R^2) serves as a general gauge of the quality of the fit. Strictly speaking, R^2 measures the extent to which the fitting equation (in this case $y = mx + b$) correlates x and y values in the data set. A perfect fit has $R^2 = 1$, but this is rarely encountered. Good fits of experimental data typically return R^2 values of 0.90 or above. If $R^2 < 0.90$, chances are that the model summarized by the fitting equation does not describe the data appropriately. In this case, you should look for a different model or repeat the experiment (or both.)

7. LINEAR REGRESSION: EXCEL'S REGRESSION TOOL

Microsoft EXCEL's REGRESSION function performs automatically the calculations outlined above for any (x,y) data set you provide. We will show you how to set up a worksheet for a regression calculation. You can then save this worksheet as a "template" for future use.

Data Input:

- Enter the text string *Linear Regression Template* into cell A1 of a new worksheet.
- Enter titles for the X and Y axes into cells A3 and B3, respectively;
- The array of X values occupies the A column, beginning with cell A4;
- The array of Y values occupies the B column, beginning with cell B4.

Setting up the Regression Calculation:

- From the menu bar, choose OPTIONS, then ANALYSIS TOOLS A window will come up with many computational choices, which are listed alphabetically. Scroll down the list and double-click on REGRESSION.
- The REGRESSION window will appear.
- Place the cursor on the box to the right of Y INPUT RANGE, then go back to the worksheet and select with the mouse the region of column B that contains the Y data (starting with cell B4 and down to the last cell containing experimental data). Note that the first and last cells of the region appear in the appropriate box of the REGRESSION window.
- Place the cursor on the box to the right of X INPUT RANGE, then go back to the worksheet and select with the mouse the region of column A that contains the X data (starting with cell A4 and down to the last cell containing experimental data). Note that the first and last cells of the region appear in the appropriate box of the REGRESSION window.

- Place the cursor on the box to the right of SUMMARY OUTPUT RANGE, then go back to the worksheet and select cell G3 with the mouse. The summary output range comprises the statistical analysis of the fit (see *Section A.1*)
- Click on the box next to RESIDUALS.
- Place the cursor on the box to the right of RESIDUAL OUTPUT RANGE, then go back to the worksheet and select cell C3 with the mouse. The residual output range contains the predicted Y value and the residuals for each data pair.
- Click on OK to start the calculation.

Numerical Output

- EXCEL will write the summary and residual output onto the worksheet, providing far more information about the fit than you actually need. Here is a breakdown of what you need to extract from the output:
- The y-intercept and its standard deviation are in cells G19 and H19, respectively.
- The slope (denoted as $x1$ by EXCEL) and its standard deviation are in cells G20 and H20, respectively.
- The predicted Y-values and the residuals are in columns D and E, respectively.
- The correlation coefficient of the fit (R^2) is in cell H6.

Plotting Your Results:

- Select the range A3:B200 and, while holding the control key, select the range D3:D200. This operation will allow you to plot on the same graph the observed and predicted Y values versus X. Note that there are blank cells in your range. This was done so that, when you use this worksheet in the future, any new data points will be plotted automatically.
- Now follow the usual procedures for plotting: click on the chart wizard button, place the chart anywhere on the worksheet, select a scatter plot with *markers* only, keep the legend, and give appropriate titles to the chart and axes.
- Inside the chart, double click on the PREDICTED Y series. Change the format to line only (no marker). Your plot now consists of markers for the observed data and a line for the fit.

Saving Your Work:

- Under FILE, choose SAVE AS
- Click on OPTIONS.
- Under FILE FORMAT, choose TEMPLATE, then OK.
- Name the FILE LINEAR REGRESSION TEMPLATE and save it to a floppy disk, *not the hard disk*.
- By turning the workspace into a template, the options you have chosen to format the data input region, the summary output regions, and the plot will be saved for future fits on any computer. The only requirement is that the other computer be equipped with EXCEL version 4.0 or above. Lower versions of EXCEL will not accept the template.

Using the Linear Regression Template. When you open the LINEAR REGRESSION TEMPLATE, you will find data and a fit summary from a previous session. The text and numbers in columns A and B may be overwritten. However, **do not overwrite or otherwise change columns C-E (the residual output range) or the summary output range (starting with column G).** You may move and resize the graph around the worksheet to improve visualization of the data.

Here are specific instructions on how to use the LINEAR REGRESSION TEMPLATE.

- Enter new labels for the X and Y axes in cells A3 and B3, respectively.
- Enter your X and Y data as columns beginning with cells A4 and B4, respectively. As you do this, the data points on the graph will change. Note that the straight line on the graph no longer reflects the best fit to the new data. We need to perform a regression analysis on the new data.
- From the menu bar, choose OPTIONS, then ANALYSIS TOOLS Scroll down the list and double-click on REGRESSION.
- The REGRESSION window will appear.
- Place the cursor on the box to the right of Y INPUT RANGE, then go back to the worksheet and select with the mouse the region of column B that contains the Y data (starting with cell B4 and down to the last cell containing experimental data).
- Place the cursor on the box to the right of X INPUT RANGE, then go back to the worksheet and select with the mouse the region of column A that contains the X data (starting with cell A4 and down to the last cell containing experimental data).
- Do not change any of the other settings, then click on OK.
- EXCEL will warn you with a dialog box that the fitting routine will overwrite the contents of the output and residual summary ranges. Click OK to allow the calculation to continue.
- Once completed, the results of the calculation will be reflected by: (i) new data in the summary ranges, (ii) new predicted values in column C, and (iii) a new straight line on the accompanying graph.
- **VERY IMPORTANT:** rename the file now with a very descriptive name (e.g., "J. de Paula - Data for Expt. 1") and save onto a floppy or zip disk.
- If you wish to print the data, summary, and plot on a single sheet of paper, choose FILE - PRINT.
- If you wish to print the graph only (the preferred format for laboratory reports), double click on the graph. A new window will come up containing the graph only. Now choose FILE - PRINT. To return to the worksheet, choose WINDOW - <WORKSHEET NAME>.