Data Analysis Problem Set

1. Be a Data Detective

There are many ways in which errors can creep into data, and you should always be alert to anything amiss which could invalidate the data. Here are some examples of suspicious-looking data sets. In each case, see if you can deduce what is wrong and suggest what can be done about it, if anything - for example, is it still possible to obtain a result with a standard error?

1. a) The time taken (in seconds) for 20 revolutions of a turntable, measured by stopwatch:

   15.03, 14.89, 14.96, 15.08, 14.21, 14.99

1. b) The frequency (in kHz) of a violin string measured by a digital frequency meter:

   0.6618, 0.6578, 0.6598, 0.6588, 0.6608, 0.6598

1. c) The length (in cm) of a cylindrical shaft, measured with a metre rule:

   49.1, 48.9, 49.2, 50.8, 49.1, 49.0

1. d) The current (in A) required in a small electromagnet to pick up a steel disk:

   5.16, 5.11, 4.92, 4.93, 4.83, 4.75

1. e) Millikan's photoelectric effect paper (*Phys Rev* **8** 355–388) contains several examples of how errors can creep in during the printing process. Here's part of the results table from p. 372, giving the stopping potential (in negative Volts) and the photocurrent (in mm deflection of an electrometer measuring the charge collected in a time of 30s). The data below are for illumination of a sodium surface with the 433.9 nm spectral line of mercury.

Potential	1.524	1.576	1.629	1.581
Photocurrent	4	10	20	44

   What do you think it said in Millikan's logbook?

1. f) A slightly more subtle printing error is contained in the previous entry in the same table, for the 404.7 nm spectral line:

Potential	1.367	1.419	1.471	1.524	1.576
Photocurrent	3	24	36	55	82

   The relevant page of Millikan's paper is given on the web site. Can you work out what has happened here? Probably the best way to sort it out is to try to read the data off the graph. This is surprisingly difficult, as the graph paper does not appear to have thick lines every ten squares, and Millikan has not used tick marks on the axis. Furthermore on the vertical scale he has 'multiplied by such factors as were necessary to give all the curves practically the same maximum ordinate'. For this curve he has divided the ordinate by two, whereas on the neighbouring 433.9 nm plot there is no scaling. When you have reconstructed the data from graph, compare with the values in the table.

2. Means and Standard Deviations

2. a) The following are readings of times in seconds measured with a stopwatch for ten cycles of a rapid oscillation. Examine them, and before actually carrying out any calculations, try to *estimate* rough values for their mean and standard deviation. Remember that some deviations are bigger and some smaller than the standard deviation (and in general, rather more are bigger than smaller, but with only 5 this may not help much).

3.21, 3.09, 3.12, 3.09, 3.18

Now *calculate* values for the mean $\bar{t}$, the best estimate of the population standard deviation σ_t and the standard deviation of the mean $\sigma_{\bar{t}}$.

2. b) Five further readings were also taken:

3.23, 3.15, 3.11, 3.24, 3.10

Use them to find further values of $\bar{t}$, σ_t and $\sigma_{\bar{t}}$. Is one set more reliable than the other? Are the two sets of data consistent?

2. c) Now treat all ten readings as a single data set. Find values of $\bar{t}$, σ_t and $\sigma_{\bar{t}}$ for the set. Which of the quantities has shown the most significant change (as compared with the values for five readings), and why?

3. How much precision do we need?

A digital voltmeter is used to determine the output (in Volts) of a power supply subject to small random fluctuations. Five readings were taken, as follows:

4.84468, 4.84392, 4.84616, 4.84499, 4.84257

Find the mean and the standard deviation of the mean.

We now wish to investigate whether we needed a voltmeter which gave so many significant figures — i.e., if these measurements had been made with an equally accurate but less precise voltmeter, would it have mattered? The following rows, obtained by successively rounding the original readings, show what the readings would have been with less precise meters. For each row, calculate the mean and the standard deviation. From your results, suggest how many significant figures we actually need in each reading.

(i)	4.84468	4.84392	4.84616	4.84499	4.84257
(ii)	4.8447	4.8439	4.8462	4.8450	4.8426
(iii)	4.845	4.844	4.846	4.845	4.843
(iv)	4.84	4.84	4.85	4.84	4.84

Obviously the number of decimal places we need depends on the size of the random fluctuation in the data. So we should re-phrase the question: how accurately do we need to record the random fluctuations in the data? The same issue crops up with giving results: how many decimal places do we need to give in the result and error? A good rule of thumb is that the error in the last place should lie between 2 and 19, or with a large data set, 3 and 29. That implies that the fluctuations should be significantly larger, maybe between 4 and 40.

4. ⋆ The Mean as a Minimum Variance Estimator

Suppose we have three independent measurements of the same quantity x, x_1, x_2, and x_3, with variances V_1,V_2 and V_3. Thus $\mathsf{x} = x_1 - \epsilon_1$ with $\langle \epsilon_1^2 \rangle = \mathsf{V}_1$ and similarly for 2 and 3. Consider a weighted mean

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3}{w_1 + w_2 + w_3}$$

where the w are the weights of the three measurements. Show that

$$\mathsf{x} = \bar{x} - \frac{w_1\epsilon_1 + w_2\epsilon_2 + w_3\epsilon_3}{w_1 + w_2 + w_3}.$$

Thus $\bar{x}$ is an unbiassed estimator for x for *any* weights w_i. So which is the best set of w? Show that multiplying all the w by a constant does not change $\bar{x}$, so in fact we are trying to find the best *ratios* of the w. The best set is the one with minimum variance.

Show that

$$\mathsf{V}_{\bar{x}} = \frac{w_1^2\mathsf{V}_1}{(w_1 + w_2 + w_3)^2} + \frac{w_2^2\mathsf{V}_2}{(w_1 + w_2 + w_3)^2} + \frac{w_3^2\mathsf{V}_3}{(w_1 + w_2 + w_3)^2}.$$

The minimum variance is when

$$\frac{\partial \mathsf{V}_{\bar{x}}}{\partial w_i} = 0.$$

Show that this implies that $w_1\mathsf{V}_1 = (w_1+w_2+w_3)\mathsf{V}_{\bar{x}}$ and hence that $w_1\mathsf{V}_1 = w_2\mathsf{V}_2 = w_3\mathsf{V}_3$. If we call this common product K then we have found that $w_i = K/\mathsf{V}_i$. The value of K is irrelevant, since it doesn't change $\bar{x}$ or $\mathsf{V}_{\bar{x}}$, so we can set it to be any convenient number.

Thus if the data are of equal accuracy then the best weights are all equal, $w_i = 1$, for example, and we recover the usual expression for the mean. Alternatively if the data are not of equal accuracy then the best weighted mean uses the reciprocal of the variance as weights: $w_i = 1/\mathsf{V}_i$.

5. Propagating Errors

A length of thick-walled glass tubing is measured to find its dimensions. The length l, internal diameter $2a$ and external diameter $2b$ were found to be: $l = 133.2(2)$ mm, $2a = 3.210(15)$ mm, $2b = 12.83(11)$ mm.

5. a) Calculate the fractional errors in these quantities.

5. b) Calculate the wall thickness, the area of the internal cylindrical surface and the volume of glass. In each case calculate also standard error and the fractional error. Discuss whether the resulting error is more simply related to the individual errors or fractional errors in each case.

5. c) The NIST Physical constant web-site, http://physics.nist.gov/cuu/Constants, gives a value for e/h as 2.417 989 454(60) $\times 10^{14}$ A J^{-1}. Compute the value and error of h/e. Compare the fractional errors in the two ratios.

5. d) The values of h and e given on the NIST website are actually derived from two more accurately known combinations, $J = h/e$ and $K = h/e^2$. (J is found from a technique known as the Josephson junction, and K is known as von Klitzing's constant, and is measured in the quantum Hall effect.) The two values are essentially independent. The fractional uncertainty in J is 2.5×10^{-8} and in K, 6.8×10^{-10}. Find expressions for h and e, and their fractional uncertainties and compare with the NIST web-site. Is the error in K contributing anything significant?

6. Straight Line Fit

The table below gives the data Millikan plotted on the frequency/voltage graph for the data set discussed above. Millikan does not give the stopping potential explicitly; the values given here are read off the graph on p. 371, which is available on the web-site. Similarly he does not give the frequencies explicitly; the values given here are derived from the wavelengths he quotes, and a contemporary value for the speed of light.

x	Frequency(PHz)	0.54894	0.69082	0.74077	0.82126	0.95913	1.18269
y	Stopping Potential	-2.046	-1.488	-1.296	-0.915	-0.383	0.518

Calculate $\bar{x}$ and $\bar{y}$, and then $A = \sum(x - \bar{x})^2$, $B = \sum(y - \bar{y})^2$ and $C = \sum(x - \bar{x})(y - \bar{y})$. The best fit slope is given by $m = C/A$. The sum of the squared y-deviations from the line is given by $R = B - C^2/A$, and the best estimate of the variance of the y-data $V(y) = R/(N - 2)$ for N data points. Use this to calculate the corresponding standard deviation and compare with the your own view of the accuracy of the extrapolation to zero photocurrent used by Millikan. Is the data as consistent with the straight line as you would expect? If it is, then the straight line fit is consistent, and we can trust the slope. The standard deviation on the slope is $\sqrt{V(y)/A}$. Give the result for h/e with its error, and compare with the current value quoted above.

Compare your results with those found using some standard analysis package (Origin or Excel, or your own calculator).

7. Final Teaser

The following table gives data on the radioactive background in the First Year General Physics Laboratory, in the form of sample counts n_i in various sampling periods t_i.

Sample Period	Sample Counts n_i											
10 s	3	2	2	1	1	2	1	1	2	4	1	2
20 s	4	1	1	3	4	5	1	1	3	3	2	3
30 s	1	8	4	3	3	3	10	4	6	4	4	6

Discuss the merits of the following methods for finding the mean background count rate R and its standard deviation σ_R. Are they: valid, or unbiassed but but sub-optimal, or systematically biassed ?

7. a) For each sample calculate a rate $r_i = n_i/t_i$, and find the mean and its standard deviation.

7. b) Add up all the counts $N = \sum_i n_i$ and divide by the total sample period $T = \sum_i t_i$: $R = N/T$ and since counts are Poisson-distributed, $\sigma_R = \sqrt{N}/T$.

7. c) As 7. a) but do a weighted mean and standard deviation, using $\sqrt{n_i}/t_i$ as the standard deviation of r_i, as in 7. b). The weights are then t_i^2/n_i, so $R = \sum t_i / \sum(t_i^2/n_i)$.

7. d) Find mean count rate of all data with same sampling period, and their standard deviations, and do weighted average using the inverse variances as weights.

7. e) As 7. d) but use the sample time as weights.

7. f) Add the rows of data to synthesize a set of 12 rates in 60s sample periods, and take mean and standard deviation of these.