A Quick Introduction to Data Analysis (for Physics)

Dr. Jeff A. Winger

1 What is data analysis?

Data analysis is the process by which experimental data is used to obtain a valid and quantifiable result. Part of this analysis **must** include an estimation of the **accuracy** of the result, i.e. how certain are you that you have the correct answer. We refer to this as the uncertainty in the result. Any experimental value that does not have an associated uncertainty is worthless since it indicates 100% uncertainty in the value. There are several aspects to data analysis which will be discussed in this document. Let's start with some general information and move on to more complex issues.

2 Measurement Uncertainty

Whenever an experiment is performed, every measurement has associated with it some inherent uncertainty. This can be due to the limits of our ability to measure a quantity with a particular device, or in our fitting of measured values to some model. For example, if I measure the length of something then the **precision** of the value depends on the device I use. I certainly can't expect to measure as precisely the thickness of a quarter with a meter stick as I could with a vernier caliper or micrometer. Hence, I must assign with any measurement some reasonable estimate of the uncertainty in the value, i.e. the range of values in which I would expect the actual value to exist. Hence, if I measure the length L of a rod, I will assign an uncertainty of σ_L to this value and say that there is a good probability that the actual length lies between $L - \sigma_L$ and $L + \sigma_L$. For simplicity we write this as $L \pm \sigma_L$.

How do you estimate measurement uncertainty? This depends on the device you use. If you have a scale to which you compare then you can estimate the value based on the relative position between tick marks on the scale. If a meter stick is being used, where the ticks are separated by 1 *millimeter*, then you can estimate the value to the nearest 0.1 mm. The uncertainty in this value can be reasonably assumed to be ± 0.3 mm. A similar rule of thumb can be used for many other scales. An exception is any device with a vernier scale where you would assume ± 0.5 of the lowest digit measured, e.g. 21.30 ± 0.05 mm with the last digit read from the scale was the 3. For digital scales, assume ± 0.5 of the last digit. Alternatively, you could make multiple measurements and use statistical analysis as described later. Remember, trailing zeros *are* significant.

3 Significant Figures

You have all dealt with significant figures to some extent in the past, but now I will introduce a proper prescription for how to limit the number of significant figures in a result. The basic concept of how to determine the number of significant figures is based on the precision of the number and not on the number of significant figures in the original numbers which were combined. We will use the following basic rules:

1. A value cannot have more significant figures than are allowed by the uncertainty in the value. E.g., if the averaged measured value is 1.2345 by the readings (each reading being made to this accuracy), but the uncertainty in the reading based on statistical analysis is 0.003, then the value must be limited in significant figures to 1.235.

2. We generally specify uncertainties in terms of the uncertainty in the last significant figure. For the previous example, the uncertainty is 3 in the last significant figure. When uncertainties are propagated, it will be necessary to round off the numbers to limit significant figures to meaningful numbers. For example, if the uncertainty is 100 in the last significant figure there can be very little meaning to the last digit, so this digit in both the number and the uncertainty should be dropped (i.e., 1.23456 ± 0.00100 should become 1.2346 ± 0.0010). The question then arises as to when to round up the numbers. As a rule of thumb, we will choose to keep significant figures until the uncertainty in the last digit is **25**. If it exceeds 25, then the uncertainty will be rounded and the last digit dropped.

These rules are not set in stone, but are a good guideline that you should follow. (These are not even the rules as given in any particular book, but I feel they are closer to the accepted practice for real scientific publications.) One last point before leaving this topic is to state that as you propagate uncertainties in an analysis, you should retain all the digits until the final answer is obtained. Only then should you begin rounding off the numbers. In this way you limit the possibility of overstating or understating your uncertainty in the final result due to how the uncertainties were rounded during the analysis.

4 Propagation of Uncertainty

In general, in doing experiments you will not measure a single quantity in order to have your final result. Instead, you will measure several quantities which need to be combined in order to obtain your final value. For example, to find the area of a rectangle, you would measure length and width, and multiply these together. Since both measurements have uncertainties associated with them, how do you find the uncertainty in the area. This process is often termed error analysis, but I don't like the use of the word "error" since there is no "error" in the measurement, just uncertainty. I will present the boring theory first, then give you some useful formulas.

Consider the case of a quantity x which will be determined by the measurement of a set of variables $\{u, v, \ldots\}$, where each measured value has some associated uncertainty ($u \pm \sigma_u,\ v \pm \sigma_v,\ etc.$). We won't specify the way in which these individual uncertainties were determined since they can be statistical or instrumental. We write x in terms of a function of our measured variables and various constant factors

$$x = f(u, v, \ldots). \tag{1}$$

Now we will assume that a reasonable estimate for the true value of x comes from using the mean ("best") values for each of the measured quantities—i.e.,

$$\bar{x} = f(\bar{u}, \bar{v}, \ldots). \tag{2}$$

The uncertainty in x must somehow be related to the uncertainties in the variables $\{u, v, \ldots\}$. This relationship is done by looking at how the final answer changes as each variable is changed. In simplest terms this is related to the slope of the function as a variable changes. Doing a little math this becomes

$$\sigma_x^2 \simeq \sigma_u^2 \left(\frac{\partial f}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial f}{\partial v}\right)^2 + \ldots + 2\sigma_{uv}^2 \left(\frac{\partial f}{\partial u}\right)\left(\frac{\partial f}{\partial v}\right) + \ldots \tag{3}$$

The first two terms are related to the *variance* or uncertainties in the individual measurements $\{\sigma_u, \sigma_v, \ldots\}$ while the last term is related to the *covariances* σ_{uv}^2 of the data set, i.e. how the variables change with each other. Typically your experiment will involve only independent variables since each is measured independently of the others. By independent, we do not mean unrelated since the final result depends on both. Instead, we would consider the covariance only in cases where the two variables are measured simultaneously. For example, if we are finding a resistance by measuring current and voltage simultaneously, then we do need to worry about covariance. In cases with covariance, we will be making multiple measurements and doing statistical analysis so that

$$\sigma_{uv}^2 \equiv \frac{1}{N}\sum_{i=1}^{N}[(u_i - \bar{u})(v_i - \bar{v})], \tag{4}$$

which can be negative. If the two variables are actually independent then the covariance should be small.

4.1 Examples

In this section we will consider various cases where we assume a set of dependent variables $\{u, v\}$ which may have covariance and are combined with fixed constants $\{a, b\}$ which have no uncertainty. We will use all terms of Eqn. 3. If your data set does not contain covariance, then just use $\sigma_{uv}^2 = 0$.

1. **Adding/subtracting a constant.** The uncertainty does not change, but the percentage uncertainty will.

$$x = u \pm a, \qquad \sigma_x = \sigma_u, \qquad \frac{\sigma_x}{x} = \frac{\sigma_u}{u \pm a} \tag{5}$$

2. **Adding/subtracting variables.** You will note that the percentage uncertainty in subtracting large numbers can become quite large, possibly exceeding 100%. It is a good practice to avoid subtracting large numbers.

$$x = au \pm bv, \qquad \sigma_x^2 = a^2\sigma_u^2 + b^2\sigma_v^2 + 2ab\sigma_{uv}^2, \qquad \frac{\sigma_x}{x} = \frac{\sqrt{a^2\sigma_u^2 + b^2\sigma_v^2 + 2ab\sigma_{uv}^2}}{au \pm bv} \tag{6}$$

3. **Multiplying two numbers.** Notice that it becomes easier to work with the percentage uncertainties.

$$x = auv, \qquad \sigma_x^2 = (av\sigma_u)^2 + (au\sigma_v)^2 + 2a^2uv\sigma_{uv}^2, \qquad \left(\frac{\sigma_x}{x}\right)^2 = \left(\frac{\sigma_u}{u}\right)^2 + \left(\frac{\sigma_v}{v}\right)^2 + 2\left(\frac{\sigma_{uv}^2}{uv}\right) \tag{7}$$

4. **Dividing two numbers.** Notice the similarity to multiplication. Also notice that due to the negative sign that the covariance term can lead to a smaller uncertainty in the final result. This is actually true in any situation since the covariance can be negative.

$$x = \frac{au}{v}, \qquad \sigma_x^2 = \left(\frac{a\sigma_u}{v}\right)^2 + \left(\frac{au\sigma_v}{v^2}\right)^2 - 2\frac{a^2u}{v^3}\sigma_{uv}^2, \qquad \left(\frac{\sigma_x}{x}\right)^2 = \left(\frac{\sigma_u}{u}\right)^2 + \left(\frac{\sigma_v}{v}\right)^2 - 2\left(\frac{\sigma_{uv}^2}{uv}\right) \tag{8}$$

5. **Variable raised to a power.**

$$x = au^b, \qquad \sigma_x = abu^{b-1}\sigma_u, \qquad \frac{\sigma_x}{x} = b\frac{\sigma_u}{u} \tag{9}$$

6. **Exponential Functions.**

$$x = ae^{bu}, \qquad \sigma_x = abe^{bu}\sigma_u, \qquad \frac{\sigma_x}{x} = b\sigma_u \tag{10}$$

7. **Constant raised to the power of u.**

$$x = a^{bu}, \qquad \sigma_x = b\ln(a)a^{bu}\sigma_u, \qquad \frac{\sigma_x}{x} = b\ln(a)\sigma_u \tag{11}$$

8. **Logarithmic Functions.** This is a rare case where the fractional error form is more complex.

$$x = a\ln(bu), \qquad \sigma_x = \frac{ab\sigma_u}{u}, \qquad \frac{\sigma_x}{x} = \frac{b\sigma_u}{u\ln(bu)} \tag{12}$$

9. **Trigonometric Function.** The angle used in the trigonometric function has a major effect on the resulting uncertainty. Please note that the value of bu in these formulas as well as σ_u must be in radians.

$$x = a\cos(bu), \qquad \sigma_x = ab\sin(bu)\sigma_u, \qquad \frac{\sigma_x}{x} = b\tan(bu)\sigma_u \tag{13}$$

$$x = a\sin(bu), \qquad \sigma_x = ab\cos(bu)\sigma_u, \qquad \frac{\sigma_x}{x} = b\cot(bu)\sigma_u \tag{14}$$

5 Statistical Analysis

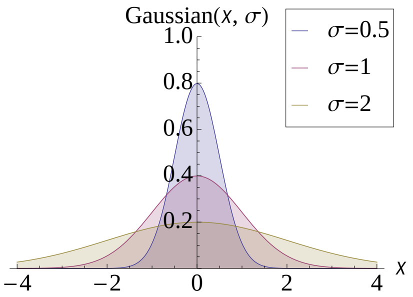

Figure 1: *Three normal (or Gaussian) distributions each with a mean of zero but with differing values of σ, the standard deviation. Larger standard deviations indicate a greater spread in the distribution.*

When measurements are made, the result may vary from one measurement to the next due to reading uncertainty in the measurement device or a variation in the system at the level of the precision of the measuring device. If we do not observe some variation in the results, it just means we are not using a device with a high enough level of precision for the measurement. Since the measurements are not always the same, there must be an associated uncertainty with this measurement. How do we estimate this uncertainty? As an example, let's consider the measurement of some quantity x, where the individual measurements of x are represented by x_i. We can bin our measured values (x_i) and plot the number of measurements within each bin as a histogram, then we will see that the results clump about some average value which we must assume is the correct value. (When I say "bin" I mean the following. We define values of x_j which are separated by $2\delta x$. If a value of x_i lies between $x_j - \delta x$ to $x_j + \delta x$, then we add one count to the x_j bin. The number of counts in each bin are what will be plotted.) The spread of the distribution about the average must somehow reflect the uncertainty in the measurements. If we take a very large number of measurements and let the bin width (δx) become very small, the distribution will become a smooth function, the *parent or limiting distribution* $p(x)$. Most measurements will form a symmetrical distribution which can be well represented by a *normal (Gaussian) distribution* (see Figure 1). There is no reason to get into all the details here, just suffice it to say that by using statistical analysis we can describe the center and width of the distribution from the measured values. To do this, I will consider how to determine the best value and its associated uncertainty for a set of N values given by $\{x_i\}_N$.

5.1 Mean Value

The *mean* (or average) of $\{x_i\}_N$is what we assume to be the best value. The mean value is defined as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{15}$$

5.2 Deviations

We use the concept of *deviation* to describe how the measured values vary about the mean value. Again there are lots of mathematical details, but the result is the definition of the *sample standard deviation*, s where:

$$s^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{(N-1)}\left[\sum_{i=1}^{N} x_i^2 - N\,\bar{x}^2\right] \simeq \sigma_x^2. \tag{16}$$

The use of $N-1$ in the formula relates to the fact that the average value was determined using the data. Be careful that this is the equation used by your calculator or spreadsheet. The sample standard deviation gives the variation in the distribution of measured values, but it does not give the uncertainty in the best value. Instead we would use the *standard deviation of the mean*, $\bar{s}$ where:

$$\bar{s}^2 \equiv \frac{1}{N(N-1)}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{(N-1)}\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2\right] = \frac{s^2}{N} = \sigma_{\bar{x}}^2 \simeq \sigma_\mu^2. \tag{17}$$

However, we need to use caution with this formula since we cannot assume that by actually doing an infinite number of measurements that we can obtain an infinitely small deviation. Other factors which can cause variability between the measurements will eventually limit our precision.

5.3 What does it all mean?

In our experiments, we will generate a data set $\{x_i\}_N$. From this we can extract the statistical results for $\bar{x}$, σ_x, and $\sigma_{\bar{x}}$. We can now assume that any future measurements of x using the same system must be distributed about the mean value with a characteristic width σ_x. In saying this, we can assume that any single measurement using this same system must yield a value and uncertainty given by $x \pm \sigma_x$. However, if we generate a data set $\{x_i\}_N$, then the uncertainty in the mean value is much more precisely known and is given by $\bar{x} \pm \sigma_{\bar{x}}$.

5.4 Weighted Averages

Now lets consider the case where the uncertainties in the individual data points are not equal. In such cases, we weight the values (x_i) by their uncertainties (σ_i) so that the weighted average is defined by

$$\bar{x} = \frac{\sum\limits_{i=1}^{N} w_i x_i}{\sum\limits_{i=1}^{N} w_i} \qquad\qquad w_i = \frac{1}{\sigma_i^2}. \tag{18}$$

The uncertainty in $\bar{x}$ can be estimated using

$$\sigma_{\bar{x}}^2 = \frac{\sum\limits_{i=1}^{N} w_i \left(x_i - \bar{x}\right)^2}{(N-1)\sum\limits_{i=1}^{N} w_i} = \frac{1}{(N-1)}\left[\frac{\sum\limits_{i=1}^{N} w_i x_i^2}{\sum\limits_{i=1}^{N} w_i} - \bar{x}^2\right], \tag{19}$$

which includes a consideration of the statistical variation. However, the value used should never be less than

$$\sigma_{\bar{x}}^2 = \frac{1}{\sum\limits_{i=1}^{N} w_i}. \tag{20}$$

6 Linear Regression

Linear regression, often called a least-squares fit, refers to the process of fitting a straight line to a set of N simultaneously measured data points specified by $\{x_i, y_i\}_N$. In this data set it is assumed that we measured values of y_i, the dependent variable, at specific values of x_i, the independent variable. What is desired is to fit the data to a theoretical model to determine something. Generally, the "fit parameters", as will be described shortly, are the quantities of interest. We can use any function to describe the data, but we will only consider the case for a straight line (note that a_1 is the intercept and a_2 is the slope):

$$y(x) = a_1 + a_2 x. \tag{21}$$

Although most spreadsheets and calculators can do this for you, it is better if you understand how it is done and how to determine the uncertainties in the fit parameters. Linear regression uses the Maximum Likelihood method to determine the best fit parameters (a_1, a_2) to describe the data. We start by defining a quantity called *chi-squared* (χ^2) which describes the variation between the data set and the fit line.

$$\chi^2 = \sum_{i=1}^{N}\left[\frac{(y_i - y(x_i))^2}{\sigma_i^2}\right] = \sum_{i=1}^{N}\left[\frac{1}{\sigma_i^2}(y_i - a_1 - a_2 x_i)^2\right], \tag{22}$$

where the σ_i values are just the uncertainties in the y_i values. By finding the minimum value of χ^2 through adjusting the fit parameters, we determine the most probable solution. I won't bore you with the details here, but taking derivatives with respect to each fit parameter yields two simultaneous equations

$$\sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} = a_1 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} + a_2 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \tag{23}$$

$$\sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} = a_1 \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} + a_2 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \tag{24}$$

which can be solved (think about linear algebra) to obtain formulas for the fit parameters.

$$\begin{aligned} a_1 &= \frac{1}{\Delta}\left(\sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i\, y_i}{\sigma_i^2}\right) \\ a_2 &= \frac{1}{\Delta}\left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i\, y_i}{\sigma_i^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}\right) \\ \Delta &= \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - \left(\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}\right)^2 \end{aligned} \tag{25}$$

with uncertainties given by

$$\sigma_{a_1} = \sqrt{\frac{1}{\Delta} \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}} \qquad \sigma_{a_2} = \sqrt{\frac{1}{\Delta} \sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \tag{26}$$

Now what happens if you don't know the individual σ_i or you want to estimate them from the data. Let's assume that all the σ_i are the same and just let $\sigma_i \to \sigma_y$ in all our equations. (You can use $\sigma_i = 1$ if you so desire.) Now this will not cause a problem with the fit, but will give you bad estimates for the uncertainties in the fit parameters. This occurs because Δ depends on σ_y^{-4} while the numerator in the uncertainty expressions only depend on σ_y^{-2}. Therefore, you will need to use the data to measure the variance of the data from the fit function to obtain an estimate for σ_y. This is done using the formula

$$\sigma_y = \sqrt{\frac{1}{N-2} \sum_{i=1}^{N} (y_i - y(x_i))^2}. \tag{27}$$

The formulas for the fit parameters and uncertainties then reduce to

$$\begin{aligned} a_1 &= \frac{1}{\Delta}\left(\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i\, y_i\right) \\ a_2 &= \frac{1}{\Delta}\left(N \sum_{i=1}^{N} x_i\, y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i\right) \\ \Delta &= N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 \\ \sigma_{a_1} &= \sigma_y \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{\Delta}} \\ \sigma_{a_2} &= \sigma_y \sqrt{\frac{N}{\Delta}} \end{aligned} \tag{28}$$