Introduction to Experimental Uncertainty

Apparatus: tape measure and/or metersticks, stopwatch, golf ball, ping pong ball, plum bobs

Theory: Error analysis or experimental uncertainty is the study and evaluation of uncertainty in measurement. Experience has shown that no measurement, however carefully made, can be completely free of uncertainties. Since the whole structure and application of science depends on measurements, it is extremely important to be able to evaluate these uncertainties and to keep them to a minimum. In science, the word "error" does not carry the usual connotations of "mistake" or "blunder." "Error" in a scientific measurement means the inevitable uncertainty that exists in all measurements. As such, errors are not mistakes; you cannot avoid them by being very careful. The best you can hope to do is to ensure that the experimental errors are as small as reasonably possible, and to have some reliable estimate of the experimental errors. We shall use the term "error" exclusively in the sense of "uncertainty," and treat the two words as being interchangeable.

Random and Systematic Errors

Not all types of experimental uncertainties can be assessed by statistical analysis based on repeated measurements. For this reason, uncertainties are classified into two groups: the *random* uncertainties, which can be treated statistically; and the *systematic* uncertainties, which cannot. Experimental uncertainties that can be revealed by repeating the measurements are called random errors; those that cannot be revealed in this way are called systematic errors. To illustrate this distinction, let us consider an example. Suppose first that we time a revolution of a steadily rotating turntable. One source of error will be our reaction time in starting and stopping the watch. If our reaction time were always exactly the same, these two delays would cancel one another. In practice, however, our reaction time will vary. We may delay more in starting, and so underestimate the time of a revolution; or we may delay more in stopping, and so overestimate the time. Since either possibility is equally likely, the sign of the effect is *random*. If we repeat the measurement several times, we will sometimes overestimate and sometimes underestimate. Thus our variable reaction time will show up as a variation of the answers found. By analyzing the spread in result statistically, we can get a very reliable estimate of this kind of error. On the other hand, if our stopwatch is running consistently slow, then all our times will be underestimates, and no amount of repetition (with the same watch) will reveal this source of error. This kind of error is called *systematic*, because it always pushes our result in the same direction. Systematic errors cannot be discovered by the kind of statistical analysis that we will be discussing below.

The treatment of random errors is quite different from that of systematic errors. The statistical methods described in the following sections give a reliable estimate of the random uncertainties, and, as we shall see, provide a well-defined procedure for reducing them. On the other hand, systematic uncertainties are hard to evaluate, and even to detect. Unfortunately, in introductory physics laboratory courses, such evaluations are rarely possible; so the treatment of systematic errors is often awkward. For now, we will discuss experiments in which all sources of systematic errors have been identified and made much smaller than the required precision.

The Mean and Stand Deviation

Suppose we need to measure some quantity *x*, and have identified all sources of systematic error and reduced them to a negligible level. Since all remaining sources of uncertainty are random, we should be able to detect them by repeating the measurement several times. We might, for example, make the measurement five times and find the results

71, 72, 72, 73, 71

(where, for convenience, we have omitted any units). The first question that we address is as follows: given the five measured values, what should we take for our best estimate x_{best} of the quantity *x*? It seems reasonable that our best estimate would be the average or mean $\overline{x}$ of the five values. Thus

$$x_{best} = \overline{x} = \frac{71+72+72+73+71}{5} = 71.8$$

More generally, suppose we make *N* measurements of the quantity *x*, and find the *N* values $x_1, x_2, \ldots, x_N$. The ***best value*** or ***average*** is

$$\boxed{x_{best} = \overline{x} = \frac{\sum x_i}{N}}$$

The concept of the average is almost certainly familiar to most students. Our next concept, that of the ***standard deviation***, is probably less so. The standard deviation of the measurements $x_1, x_2, \ldots, x_N$ is an estimate of the average uncertainty of the measurements $x_1, x_2, \ldots, x_N$, and is arrived at as follows.

Given that the average $\overline{x}$ is our best estimate of the quantity *x*, it is natural to consider the difference $x_i - \overline{x} = d_i$. This difference, often called the ***deviation*** (or variance) of x_i from $\overline{x}$, tells us *how much the i^{th} measurement x_i differs from the average* $\overline{x}$. If the deviations are all very small, then our measurements are all close together and are presumably very precise. If some of the deviations are large, then our measurements are obviously not so precise.

To be sure we understand the idea of the deviation, let us calculate the deviations for the set of five measurements reported in the table below.

Table 1. Calculation of Deviations

Trial number, i	Measured value, x_i	Deviation, $d_i = x_i - \overline{x}$
1	71	-0.8
2	72	0.2
3	72	0.2
4	73	1.2
5	71	-0.8
	$\overline{x} = 71.8$	$\overline{d} = 0.0$

Notice that the deviations are not (of course) all the same size; d_i is small if the *i*th measurement x_i happens to be close to $\overline{x}$, but d_i is large if x_i is far from $\overline{x}$. Notice also that some of d_i are positive and some negative, since some of the x_i are bound to be higher than the average $\overline{x}$, and some are bound to be lower. To estimate the average reliability of the measurements $x_1, ..., x_5$, we might naturally try averaging the deviations d_i. Unfortunately, as a glace at Table 1 shows, the average of the deviations is zero. In fact, this will be the case for any set of measurements $x_1, x_2, ..., x_N$, since the definitions of the average $\overline{x}$ ensures that d_i is sometimes positive and sometimes negative in just such a way that $\overline{d}$ is zero. Obviously, then, the average of the deviations is not a useful way to characterize the reliability of the measurements $x_1, x_2, ..., x_N$.

The best way to avoid this annoyance is to *square* all the deviations, which will create a set of positive numbers, and then average these numbers. If we then take the square root of the result, we obtain a quantity with the same units as *x* itself. This is called the standard deviation of $x_1, x_2, ..., x_N$, and is denoted by σ_x:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i)^2}$$

With this definition, the standard deviation can be described as the root mean square (RMS) deviation of the measurements $x_1, x_2, ..., x_N$. Unfortunately, there is an alternative definition of the standard deviation. There are theoretical arguments for replacing the factor *N* by (*N-1*) and defining the ***standard deviation*** σ_x of $x_1, x_2, ..., x_N$ as

$$\boxed{\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i)^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}}$$

For the five measurements of Table 1, we calculate σ_x:

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i)^2} = \sqrt{\frac{1}{5-1}\sum_{i=1}^{5}(x_i - \overline{x})^2} = \sqrt{\frac{1}{4}(0.64 + 0.04 + 0.04 + 1.44 + 0.64)} \approx 0.84$$

Thus the average uncertainty of the five measurements is about 0.84.

Suppose that we obtain the values $x_1, x_2, ..., x_N$ and compute $\overline{x}$ and σ_x. If we then make one more measurement (using the same equipment), there is a 68 percent probability that the new measurement will be within σ_x of $\overline{x}$. Now, if the original number of measurements *N* was large, then $\overline{x}$ should be a very reliable estimate for the actual value of *x*. Therefore we can say that there is *a 68 percent probability that a single measurement will be within* σ_x *of the actual value*. Clearly σ_x means exactly what we have used the term "uncertainty" to mean in the preceding sections. If we make one measurement of *x*, then the uncertainty associated with this measurement can be taken to be σ_x; and with this choice we are 68 percent confident that our measurement is within σ_x of the correct answer.

The number $\overline{x} \pm \sigma_x$ *will overlap the best estimate* x_{best} *approximately* 68% *of the time*
The number $\overline{x} \pm 1.5\sigma_x$ *will overlap the best estimate* x_{best} *approximately* 87% *of the time*
The number $\overline{x} \pm 2\sigma_x$ *will overlap the best estimate* x_{best} *approximately* 95% *of the time*

In the previous example of Table 1, our best estimate at one standard deviation is

$$x_{best} = \overline{x} \pm \sigma_x = 71.8 \pm 0.84 \cong 71.8 \pm 0.8$$
$$\Rightarrow 71.0 \le x_{best} \le 72.6$$

If we make one more measurement, one has a 68% confident level that it will be between 71.0 and 72.6.

Interpretation of the Standard Deviation

The first problem in discussing measurements that are repeated many times is to find a way to handle and display the many values obtained. One convenient method is to use a distribution or histogram. Suppose, for instance, that we were to make ten measurements of some length *x*. We might obtain the values (all in cm)

26, 24, 26, 28, 23, 24, 25, 24, 26, 25

Written in this way, these ten numbers convey fairly little information; and if we were to record many more measurements in this, the result would be a confusing jungle of numbers. Obviously a better system is called for. As a first step we can reorganize the numbers in ascending order,

23, 24, 24, 24, 25, 25, 26, 26, 26, 28.

Then we can record the different values of *x* obtained, together with the number of times each value was found in a table.

Table 2

Different values	23	24	25	26	27	28
Number of times found	1	3	2	3	0	1

The distribution of our measurements can be graphically displayed in a histogram. This is just a plot of the frequency *f(x)* against x_i, with different measured values x_i potted along the horizontal axis, and the frequency that each x_i was obtained indicated the height of the vertical bar drawn above x_i.

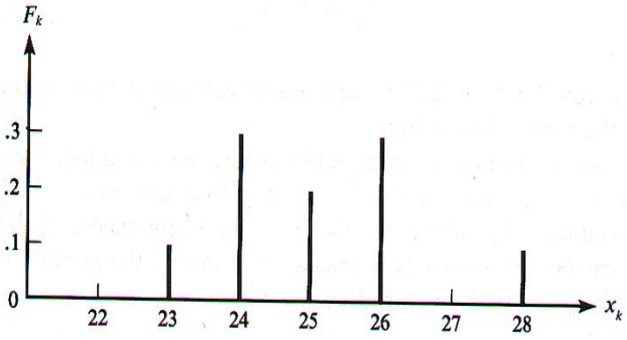

Figure 1

If one increases the number of measurements, then the histogram begins to take on a bell-shaped curve that is called ***Gaussian*** or ***Normal***.

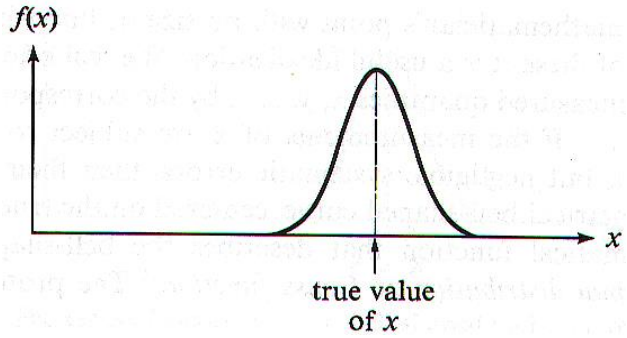

Figure 2

If the measurement under consideration is very precise, then all the values obtain will be very close to the actual value of *x*; so the histogram of results will be narrowly peaked, like the solid curve in Fig. 3. If the measurement is of low precision, then the values found will be widely spread out, and the distribution will be broad and low, like the dashed curve in Fig. 3.

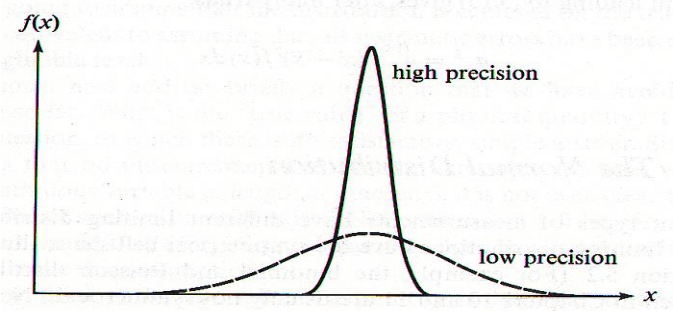

Figure 3

The mathematical function that describes the bell-shaped curve is called the ***normal distribution*** or ***Gaussian function***. The general form of this function is

$$f(x) \propto e^{-x^2/2\sigma^2},$$

where σ is a fixed parameter that we will call the *width parameter*. When $x = 0$, the Gaussian function is equal to one. The function is symmetric about $x = 0$, since it has the same value for x and $-x$. Note that as x moves away from zero, *the bell shape is wide if σ is large and narrow if σ is small*:

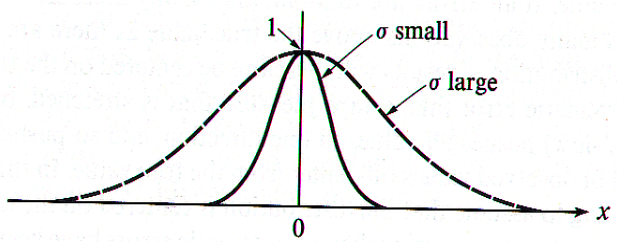

Figure 4

To obtain a bell-shaped curve centered on some other point, say $x = \bar{x}$, we merely replace x by $x - \bar{x}$ in the Gaussian function:

$$f(x) \propto e^{-(x-\bar{x})^2/2\sigma^2}.$$

The Gaussian function *f(x)* for measurement of some quantity x tells us the probability of obtaining any given value of x. Specially, the integral $\int_a^b f(x)dx$ is the probability that any one measurement gives an answer in the range $a \le x \le b$. In the figure below, the shaded area between $\bar{x} \pm \sigma_x$ is the probability of a measurement falling within one standard deviation of $\bar{x}$. (Note that we could have also found the probability for an answer falling within $2\sigma_x$ of $\bar{x}$, within $3\sigma_x$ of $\bar{x}$, etc.)

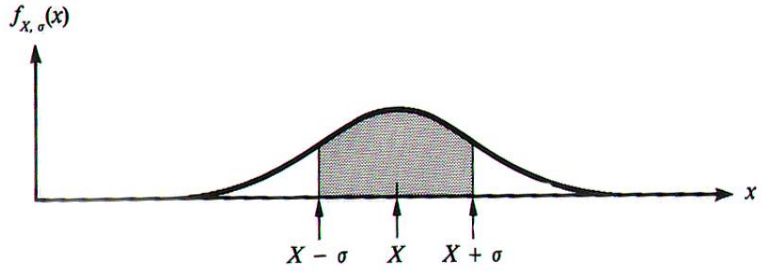

Figure 5

Standard Error (SE): the best estimate of the uncertainty of the mean

Let's summarize what we have done up to now: If $x_1, x_2 \ldots, x_N$ are the results of N measurements of the same quantity x, then, our best estimate for the quantity x is their mean, $\bar{x}$. The standard deviation σ_x characterizes the average uncertainty of the separate measurements $x_1, x_2 \ldots, x_N$. The standard deviation tells us the spread or uncertainty in our measured values, but it does not tell us how close our mean is likely to be to the true mean. To understand this, consider the following observation: if many different lab groups performed the same set of measurements, each group would obtain a different mean, and a different standard deviation. However (and this is the crucial point), the spread between the means of the different groups should be less than the spread between the measurements of each individual group. For example, suppose that five different lab groups obtained the following means ($\bar{v}$) and standard deviations (σ_v) for a set of measurements of the speed of a fired projectile:

Group 1: $\bar{v}$ = 10.5 m/s; σ_v = 0.5 m/s

Group 2: $\bar{v}$ = 10.6 m/s; σ_v = 0.4 m/s

Group 3: $\bar{v}$ = 10.8 m/s; σ_v = 0.6 m/s

Group 4: $\bar{v}$ = 10.3 m/s; σ_v = 0.3 m/s

Group 5: $\bar{v}$ = 10.2 m/s; σ_v = 0.5 m/s

The standard deviation of the means of the five different lab groups is 0.2 m/s, which is less than the spread (standard deviation) of the measurements of any individual group.

The standard deviation of the set of means is called the ***standard error*** **(SE)** and is the best estimate of the uncertainty of the mean from any individual group. Graphically, the frequency distribution of such a set of means would nearly always be a bell-shaped ***normal distribution***.

Statistical theory tells us that even when we have only a single set of measurements, we can estimate the uncertainty in the mean by calculating the ***standard error*** (**SE**), defined as:

$$\text{SE = uncertainty of the mean} = \frac{\sigma}{\sqrt{N}}$$

Therefore, we can state our final answer for the measurement of some quantity x as

$$\left(value\ of\ \ x\right) = x_{best} \pm SE = \bar{x} \pm \frac{\sigma_x}{\sqrt{N}}$$

As an example, we can consider the 5 measurements reported in Table 1. We saw that these measurements had a mean of $\bar{x}$ = 71.8 and the standard deviation is σ_x = 0.84. Therefore, the standard error is

$$SE = \sigma_x / \sqrt{5} = 0.38$$

and our final answer, based on these 5 measurements, would be that the lengths are

$$x = \bar{x} \pm \sigma_x = 71.8 \pm 0.4$$
$$\Rightarrow 71.4 \le x_{best} \le 72.2$$

An important feature of the standard error is the factor √N in the denominator. The standard deviation σ_x represents the average uncertainty in the individual measurements x_1, x_2…, x_N. If we were to make some more measurements using the same technique, we expect that the standard deviation σ_x would not change appreciably. On the other hand, the standard error $\sigma_x/\sqrt{N}$ would slowly decrease as we increase *N*. This is just what we would expect. If we make more measurements before computing an average, we should naturally expect the final result to be more reliable, and this is just what the denominator √N guarantees. This provides one obvious way to improve the precision of our measurements.

Summary:

If we measure a quantity *x* several times, then our best estimate for *x* is the mean $\bar{x}$, and the standard error, **SE**, is our best measure of the uncertainty in the mean.

In lab, we can use the following criteria to determine whether an experimental measurement of *x* is consistent with the theoretical prediction, x_{thy}:

If $\boxed{\bar{x} - 2SE \le x_{thy} \le \bar{x} + 2SE}$, then the experimental result and the theoretical prediction are said to be consistent. That is, if the theoretical prediction falls within two standard errors of the mean experimental value, then the experimental result is considered consistent with the model.

Procedure:
Forms groups of 2 or 3 people.

Determine a method for measuring a 2.00 m height.
Estimate your error in measuring this height (might you be off by 1 cm, 2 cm, 5, cm, 10 cm, 20 cm, etc?).
Did you measure to a wall or simply hold a meterstick upright?
Did you have parallax view when lining up the ball with your 2.00 m mark?
Is it possible your meterstick was slightly angled?
How much error is introduced by a 2.5° angle to the vertical? Hint: draw a triangle and do SOH CAH TOA…
This estimate of the error in measuring the height is called the *absolute error* (δh) in measuring the height.
We often write this as $h = 2.00 \pm$ (my # for δh) m
The *percent error* in measuring height is given by

$$\%\text{ error in } h = \frac{\delta h}{h} \times 100\%$$

Each person in each group should record 15 times for a golf ball to fall 2.00 m.
Tip: do not write these on paper; immediately type them into Excel as you record them!
Each person will also record 15 times for a ping pong ball to fall 2.00 m.
You should end up with either 30 or 45 times for the golf ball AND 30 to 45 times for the ping pong ball.

When recording the times, think about the absolute error in measuring the time.
When using a stopwatch, usually human reaction time is a limiting factor in getting a quality measurement.
Typical human reactions times range between 0.15 and 0.3 s.
Efforts to synchronize the release of the ball (and carefully observe the impact) might improve on this error a bit.
Make a judgement call based on how you did the experiment to estimate absolute errand percent error for time.

Think: Is error in measuring height significant compared to error in measuring time?
Notice: The reason we convert absolute errors to percent errors is it helps us compare the quality of different types of measurements. Later we can also compare percent *errors* to percent *difference*. Perhaps these errors will help us understand any discrepancies between data and theoretical predictions.

Get the average (t_{avg}) and standard deviation (σ_t) for your group data.
Repeat this data collection for a ping pong ball.
Give this information to your instructor to display at the front of the class.
Once all students have done this, record t_{avg} and σ_t for all other groups in the class.

At this point you should have a table of your group data with and average and standard deviation.
You should also have a second table with each group's average and standard deviation.
Use this second table to determine the *average of the averages* and the standard error $\left(SE_{class} = \frac{\sigma_t}{\sqrt{N}}\right)$.
In this formula, N represents the number of groups in the class (read the lab handout for more on this).

Conclusions Questions

1) Do you believe your measurement for height is a significant contributor to any discrepancies I this experiment? Support your claim by comparing the % error in measuring height (include the numerical value) to the % error in measuring time (include a numerical value).
2) Regarding only your group's data, we expect approximately 68% of your measurements to fall within one standard deviation of your mean. This means 68% of your measured values should be between the values of $\bar{t} - \sigma_t$ and $\bar{t} + \sigma_t$. What percentage of your measured values actually fall in this range? Does your percentage agree well with the 68% figure we expect? You might consider making a second copy of the data and sorting it (or using the COUNTIF function in Excel) to make this step easier.
3) Approximately 95% of your measurements to fall within two standard deviations of your mean. This means 95% of your measured values should be between the values of $\bar{t} - 2\sigma_t$ and $\bar{t} + 2\sigma_t$. What percentage of your measured values actually fall in this range? Does your percentage agree well with the 95% figure we expect?
4) One way to describe errors in our experiment is to discuss accuracy and precision.
   The *accuracy* of our experiment is quantified with a percent difference.
   The *precision* of our experiment is quantified with by converting the single group standard deviation and the entire class SE to percents (methods discussed on the data sheet).
   We say the experiment is in good agreement with theory as long as the percent difference is less than (or *approximately* equal to) the percent precision. Otherwise, there is not good agreement.
   State which of our four cases (golf ball$_{\text{your group}}$, golf ball$_{\text{entire class}}$, ping pong$_{\text{your group}}$, ping pong$_{\text{entire class}}$) are in good agreement and support your claim by stating numerical values of % precision & % difference.
5) For each type of ball, are your predicted times in good agreement with the theoretical value? Said another way, does the theoretical result fall within the 2SE confidence interval? Said yet another way, is the following mathematical statement true for each type of ball?
   $$\bar{t} - 2SE < t_{th} < \bar{t} + 2SE$$
6) Usually students notice a small difference in the average time for a ping pong ball to fall compared to the golf ball to fall. Is this difference significant? Support your claim by considering both your group data and the entire class data. Hint: consider whether or not the range of golf ball data includes the average for the ping pong ball data (and vice versa). Use numerical values (i.e. the 68% or 95% confidence limits) to support your claim.