Data Analysis Primer

F.E. Wietfeldt
revised: January 5, 2015

This primer is a consise summary of what you need to know about data and error analysis for Advanced Lab. These concepts and methods are much more sophisticated than those you used in your freshman physics lab. This paper is brief, so if you are unfamiliar with this material I recommend you spend some time with the references listed at the end to supplement this information.

1 Understanding Errors in Physics Experiments

It has been said that physics is an "exact" science. That statement is not strictly correct. All theories and experimental results in physics are at some level approximations. Nothing can ever be measured with perfect precision. In physics one can come very close to the truth, but one can never reach it exactly; just as you can count as high as you like but you can never reach infinity. Understanding the size of these approximations, and the reasons for them, are of crucial importance in interpreting the results of all physics experiments. This subject is known as error analysis. In the following discussion some important terms that you should understand and remember are indicated in **bold face** when they are defined.

A physics experiment measures some quantity or property of nature. The quantity has some **true value**, in other words an actual value which is not, and cannot be, exactly known (in quantum mechanics you can consider the true value to be the exact expectation value of the quantity). The experiment will yield a **measured value** for the quantity. The difference between the measured value and the true value is called the **true error** of the measurement. No matter how carefully a scientist does an experiment, there will always be a true error. Because the true value is not known, the true error also will never be known. By carefully analyzing the technique and apparatus used for the experiment one can obtain an **estimated error**. The estimated error is sometimes also called the **experimental uncertainty**. The estimated error is chosen so that it is probably comparable to the true error. **Error analysis** is the process of calculating the estimated error for an experimental measurement.

Consider the following simple example. You use a tape measure to measure the width of your kitchen table, and you find it to be 65.3 cm wide. This number, 65.3 cm, is the measured value. The true value is the exact width of the table, which could be 65.32752319 cm, for example. The true error is then -0.02752319 cm. In real life we would not know the true value and true error. A good way to derive the estimated error in this example would be to study the tape measure. The smallest divisions on the tape are 1 mm apart, and when you measured the table you rounded to the nearest 1 mm. So the largest error you

could have made, assuming that you worked carefully and the tape measure was accurately constructed, is ±0.5 mm, or 0.05 cm. So you can properly conclude that the estimated error for this measurement is 0.05 cm. When formally stating the result you would report both the measured value and the estimated error: the table is 65.3 ±0.05 cm wide. By this you mean that the true value is *probably* between 65.25 and 65.35 cm. If you wanted a more precise measurement you could use a more refined device, such as an expensive vernier calipers that has a smallest division of 0.01 mm. In that case you would find that the table width is 65.328 ±0.0005 cm. By putting more effort (and money!) into the experiment you can reduce both the true error and the estimated error. You can get closer to the true value, but you can never actually reach it.

Note that this definition of experimental error does not include mistakes that the scientist may have made. In the above example, if you misread the tape measure and reported that its width is 55.3 cm, that would *not* be considered an experimental error. That is more properly called a **blunder**. It is important to understand the difference between an error and a blunder. If the scientist actually does something wrong, that is a blunder. The error, in our definition, is assumed to be due only to the fundamental limitations of the technique and apparatus used; it assumes that the scientist worked as carefully as possible and made no blunders. If you make a blunder in your experiment you must repeat it and do it right!

Another important distinction is the difference between the terms precision and accuracy of a measurement method or device. The **precision** refers to the smallest quantity that can be distinguished, such as the smallest division on the tape measure, or the estimated error of the device. The **accuracy** refers to how close a device's measurement is to the true value of the quantity being measured, in other words the size of the true error. A good device has an accuracy comparable to or better than its precision. If the above tape measure were poorly constructed, it might read off by a few millimeters, even though the smallest division is one millimeter. Its precision would be 0.05 cm, but it would not be accurate.

In every experiment there are two major classes of experimental errors to consider: statistical errors and systematic errors. The methods of error analysis for these two classes are quite different. They will be defined and discussed in the next two sections.

Important Note: A very common mistake students make is to simply state how far off from the "right" answer their result was. For example: in your lab suppose you measure Plank's constant to be 7.83×10^{-34} Js, and your physics textbook says that it should be 6.63×10^{-34} Js, therefore your experimental error is (7.83 - 6.63) / 6.63 = 18%. This is incorrect! Your experimental error is determined from your own data and apparatus using the methods outlined here – it is independent of any other measurement or the value found in books. While it is useful to compare your result to the accepted value, that difference does not determine your experimental error.

2 Statistical Errors

Statistical errors are errors caused by random fluctuations in either the measurement device or the measured quantity itself. For example, if you take several measurements of the temperature of a glass of water with a high-precision thermometer, the results will differ slightly. Even if the temperature of the water is perfectly stable, the rate of heat transfer from the water to the thermometer will fluctuate randomly, as will certain properties of the thermometer. After taking many repeated measurements you will obtain a distribution of results, each of which is slightly smaller than, or greater than, some average, or central value. Mathematically, the distribution of values one obtains after repeatedly measuring some random quantity y is described by a **probability distribution function** $P(y)$. This function describes the probability that one will obtain a particular result y_i in a single measurement. $P(y)$ should have the property that its integral over all possible y is equal to one:

$$\int_{-\infty}^{\infty} P(y)dy = 1 \tag{1}$$

When $P(y)$ has this property it is said to be **normalized**. This means that the *total* probability of obtaining some result between $-\infty$ and ∞ when you make a measurement of y is 100%. This makes sense, if you make a measurement you must always obtain some result. The probability of obtaining a particular result y_i will then be less than one. Three key concepts for the analysis of statistical errors are the mean, variance and standard deviation of a probability distribution function $P(y)$. The **mean** $\overline{y}$ is the average value of y for the distribution $P(y)$:

$$\overline{y} = \int_{-\infty}^{\infty} yP(y)dy \tag{2}$$

The **variance** σ_y^2 of $P(y)$ is defined as follows:

$$\sigma^2 = \int_{-\infty}^{\infty} (y-\overline{y})^2 P(y)dy, \tag{3}$$

where $\overline{y}$ is found from Eq. 2. The **standard deviation** σ_y of $P(y)$ is a measure of the width of the distribution. It is defined to be the square root of the variance:

$$\sigma = \sqrt{\sigma_y^2} \tag{4}$$

Eqs. 2, 3, and 4 assume that $P(y)$ is normalized.

An important example of a probability distribution function is the **Gaussian function**:

$$G(y) = \frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2s^2}\right), \tag{5}$$

where μ and s are constants. By applying Eqs. 2, 3, and 4 one can show that the mean of $G(y)$ is $\overline{y} = \mu$, the variance is $\sigma^2 = s^2$, and the standard deviation is $\sigma = s$. The Gaussian function $G(y)$ is shown in Figure 1. The random fluctuations in most (but not all) measureable quantities in physics obey, to good approximation, the Gaussian function.

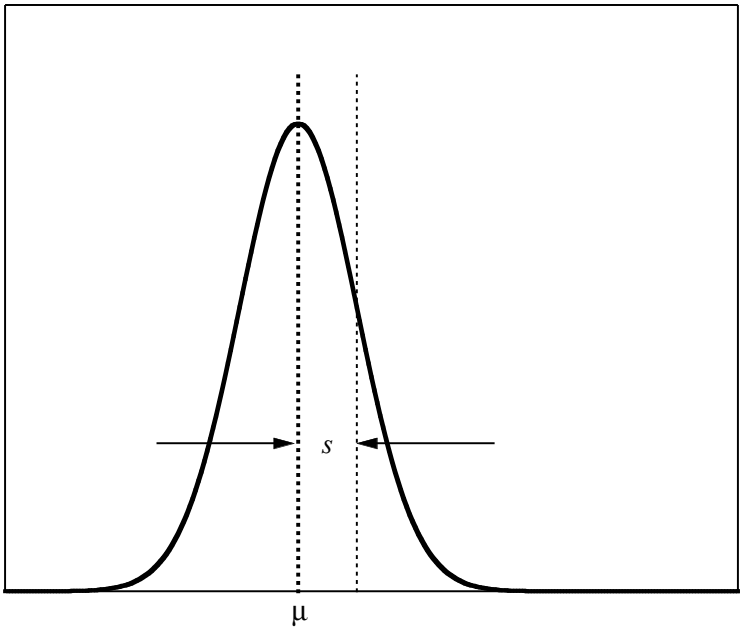

Figure 1: The Gaussian probability distribution function $G(y)$. The mean of the distribution is μ and the standard deviation is s.

Whenever you make an experimental measurement that has some statistical error, you should assume that there is some probability distribution function $P(y)$, called the **parent distribution**, that determines the probability of obtaining a particular result. Each repeated measurement will give a slightly different result. If you make many measurements, the distribution of results will begin to look just like $P(y)$. An illustration of this is shown in Figure 2. If you were to make in infinite number of measurements the distribution of results would look exactly like $P(y)$.

When you do an experiment and make a measurement, you may not know in advance what the parent distribution is. The most important attributes of the parent distribution are the mean, variance, and standard deviation. These can be estimated from the data sample. If you expect a statistical error it is always advisable to make as many measurements of your desired quantity as practical. If you make N measurements of y, the **sample mean** $\overline{y_s}$ is given by:

$$\overline{y_s} = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{6}$$

The **sample variance** σ_s^2 is defined to be:

$$\sigma_s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2, \tag{7}$$

and the **sample standard deviation** is:

$$\sigma_s = \sqrt{\sigma_s^2}. \tag{8}$$

The sample mean, sample variance, and sample standard deviation serve as estimates of the mean, variance, and standard deviation of the parent distribution. The more repeated measurements you make, the better these estimates will be (see Fig. 2). In the limit where you make an infinite number of measurements, you will obtain exactly the attributes of the parent distribution.

In certain kinds of experiments you may have additional information about the parent distribution that can help you estimate the statistical error. For example, radioactive decay is known to follow a *Poisson distribution*, in which the variance of the distribution equals the mean. In this case the estimated statistical error is simply the square root of the measured number of decays. In the absence of this kind of special information you must make repeated measurements. If one makes a set of N repeated measurements of the same quantity, by convention we take the measured value to be the sample mean $\overline{y_s}$, given by Eq. 6. The estimated statistical error is the **standard deviation of the mean** σ_{mean}:

$$\sigma_{mean} = \frac{\sigma_s}{\sqrt{N}}, \tag{9}$$

where σ_s is the sample standard deviation from Eq. 8. The standard deviation of the mean gives a measure of the uncertainty of the mean, in other words how close the sample mean is to the true mean of the parent distribution. The more repeated measurements you make (larger N), the smaller σ_{mean}, and therefore the estimated statistical error, will be (see Fig. 2).

Sometimes you will make repeated measurements of some physical quantity with different uncertainties, for example if the measurements are taken under different circumstances or with different tools. If so, instead of using Eqs. 6 and 9, you must calculate the sample mean using a *weighted average*:

$$\overline{y_s} = \frac{\sum_i \frac{y_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}, \tag{10}$$

and the weighted standard deviation of the mean is:

$$\sigma_{mean} = \sqrt{\frac{1}{\sum_i \frac{1}{\sigma_i^2}}}. \tag{11}$$

In this case, the uncertainties of the measured points σ_i must be determined independently.

3 Systematic Errors

Systematic errors result from the limited precision and imperfections in the devices and methods used to make the measurement. The true systematic errors are *not* random in nature, so the statistical methods described in section 2 are not generally useful in arriving at an estimated error. The systematic error is estimated by examining and understanding the limitations of the experimental devices and methods. There is often no "right" way to

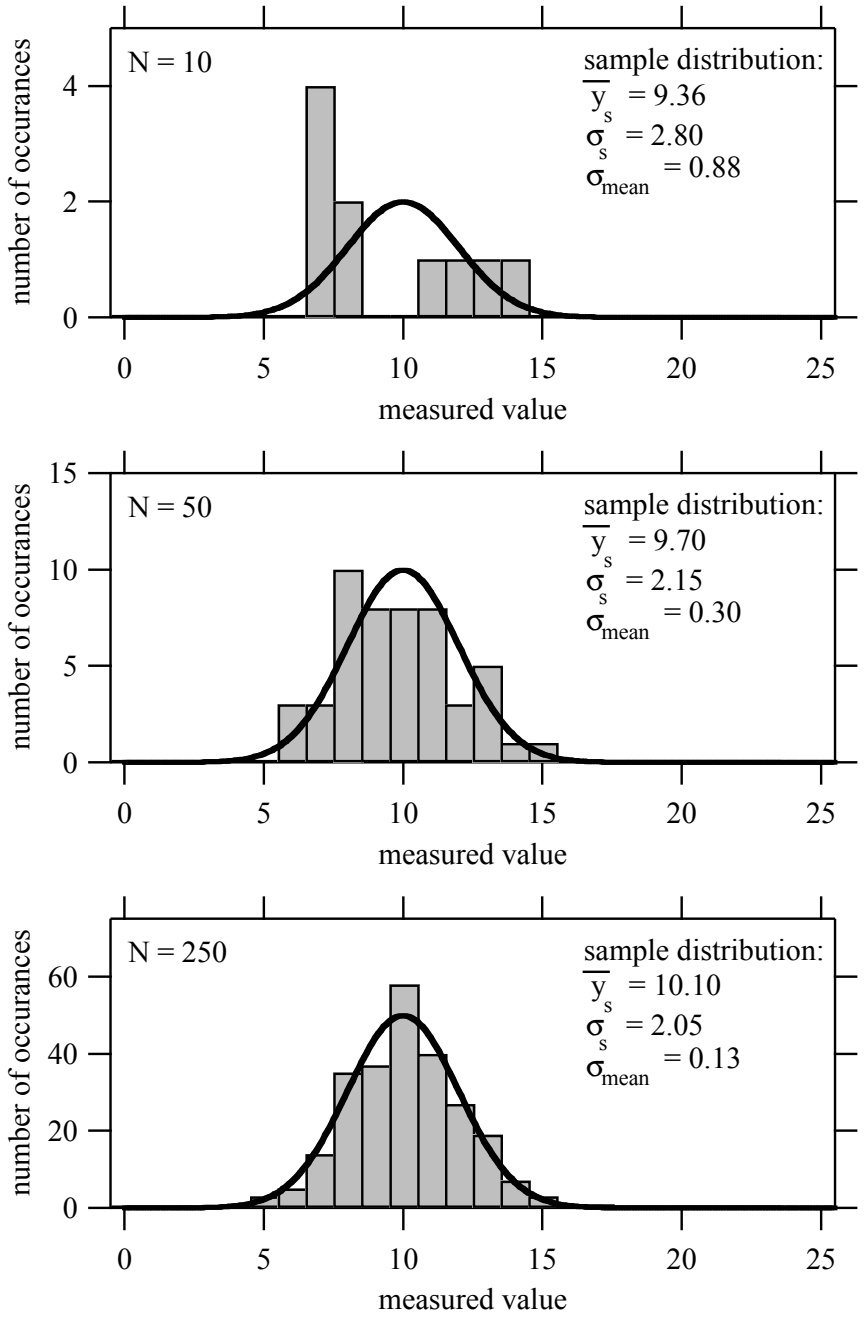

Figure 2: In the above plots, a physical quantity is measured N times, and the measured value is rounded to the nearest integer. The smooth curve is the parent probability distribution function $P(y)$, a Gaussian in this example, with $\mu = 10.0$ and $s = 2.0$. As the number of measurements N increases, the sample distribution more closely resembles, and better estimates, the parent distribution.

do this; it is something of an art. It takes a lot of experience and a thorough understanding of experimental techniques to become proficient at estimating systematic errors.

The approaches for estimating systematic errors can be broadly grouped into three categories:

1. **Instrument precision and calibration:** Every measurement device has a limit to its precision, which may determine its estimated systematic error. A good example this was the error in the tape measure measurement of the table described in section 1. The error resulted from the limited precision of the tape measure. Repeated measurements would not give different results; the true error would be the same for each measurement of the table. In this case it was easy to estimate the systematic error: the measured value was rounded to the closest small division on the tape (1 mm) so the estimated systematic error was half that, or 0.5 mm. This method is generally useful for many types of measurement instruments. It is also important to understand the accuracy of a device, which is governed by its calibration. A **calibration** is a procedure by which a measurement device is compared to an accurate standard, and adjusted so as to make the device read as accurately as possible. No device is perfectly accurate. The calibration accuracy of a test device will often be provided by its manufacturer, for example a voltmeter may be certified to an accuracy of 0.01%. If this information is not available, the scientist may have to do an auxiliary calibration experiment to determine the accuracy of one or more devices used in the main experiment. Usually the calibration accuracy of a device is superior to its precision limit, but not always. Both the calibration accuracy and precision limit of a device must be considered when estimating its systematic measurement error.

2. **Theoretical or computational modeling:** This approach is often used for highly complex experimental methods and apparatus, which can often be modelled mathematically or by computer simulation. By systematically varying parameters in the model, one can obtain estimates for systematic errors in the experiment. You will not need to use this method in Advanced Lab.

3. **Pseudorandom systematic errors:** In contrast to statistical errors, systematic errors are by nature not random. However in certain cases systematic errors may vary in complicated ways that are approximately random, enough so that the statistical methods described in section 2 can be used to estimate the error.

In Advanced Lab you will find that category 1 above is the most useful approach for estimating systematic errors in measurements.

4 Combining Errors

There may be simultaneously different sources of error for some measured quantity. For example, when the temperature of a cup of water is measured there will be a systematic error due to the precision limit and calibration accuracy of the thermometer, and at the

same time there will be a statistical error due to random fluctuations of heat transfer from the water to the thermometer. These two sources of error can be separately estimated using the methods of sections 2 and 3. We then need a way to combine these errors to arrive at the final experimental error in the measurement. In some experiments there may also be numerous sources of systematic errors that must be combined. If we can assume that the individual errors are **uncorrelated**, in other words the true errors caused by the different sources act independantly and are unrelated to each other, then we combine the errors by taking the **quadrature sum**:

$$\sigma_{tot} = \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \ldots}, \tag{12}$$

where $\sigma_1, \sigma_2, \sigma_3, \ldots$ are the individual errors. Note that we use the symbol σ here to denote the estimated error, and we used the same symbol to denote the standard deviation in section 2. Please be careful not to be confused by this. The estimated error and standard deviation are *not* the same thing. The standard deviation only applies to statistical errors. The quadrature sum formula can be used for both statistical and systematic errors, as long as they are uncorrelated. The procedure for combining *correlated* errors, in which the individual errors are dependent on each other, is much more complicated and will not be used in Advanced Lab. Whenever you have to combine errors in this course you may assume that they are *uncorrelated* and use Eq. 12.

Often the aim of an experiment is to obtain an experimental value of a physical quantity that is not directly measured. Instead it is determined by combining other measured quantities. One then must have a method for combining the errors on different quantities to obtain an estimated error for the desired final quantity. For example, you may determine the resistance R of a resistor by simultaneously measuring the voltage drop V across it and the current I that passes through it. The resistance is then given by Ohm's Law:

$$R = \frac{V}{I} \tag{13}$$

If you have estimated errors for the voltage and current measurements, what is the estimated error for the resistance? We use the **error propagation formula** to combine the errors. If p is the desired quantity that has a functional dependence on some number of other measured quantities $a, b, c, \ldots$:

$$p = f(a, b, c, \ldots), \tag{14}$$

then the estimated error of p, which we will call σ_p is given by:

$$\sigma_p = \sqrt{\left(\frac{\partial f}{\partial a}\right)^2 \sigma_a^2 + \left(\frac{\partial f}{\partial b}\right)^2 \sigma_b^2 + \left(\frac{\partial f}{\partial c}\right)^2 \sigma_c^2 + \ldots}, \tag{15}$$

where σ_a, σ_b, $\sigma_c, \ldots$ are the estimated errors of $a, b, c, \ldots$. Again, this equation can be used as long as the different errors being combined are uncorrelated, which you can always assume in Advanced Lab.

In the above example with Ohm's Law we would have:

$$\sigma_R = \sqrt{\left(\frac{\partial R}{\partial V}\right)^2 \sigma_V^2 + \left(\frac{\partial R}{\partial I}\right)^2 \sigma_I^2} = \frac{\sqrt{\sigma_V^2 + R^2 \sigma_I^2}}{I} \tag{16}$$

Because of the quadratic nature of Eqs. 12 and 15, the combined error will be dominated by the largest of the individual errors. Very small errors tend to be insignificant and can usually be omitted with little effect on the result. If the desired quantity is a function of a single parameter, *i.e.* $p = f(x)$ then Eq. 15 reduces to:

$$\sigma_p = \frac{dp}{dx}\sigma_x, \tag{17}$$

a very useful formula to remember.

5 Regression Analysis and Data Fitting

It is often the case in experimental physics that one wants to measure a physical quantity y as a function of some independent variable x. These data can then be used to test an hypothesis that $y(x)$ obeys some particular functional form, such as a straight line, a parabola, or an exponential, and to determine the specific parameters of that function. The general procedure for doing this is called **regression analysis**.

In **linear regression** the hypothesis function is a straight line. For example, we might wish to test whether a particular spring obeys Hooke's Law, which states that the spring force is proportional to the displacement of one end of the spring from its relaxed position, keeping the other end fixed:

$$F = -kx. \tag{18}$$

In an experiment, we can displace the end of the spring by different amounts x_i, and measure the resulting force F_i using a strain gauge, making a total of N measurements. This gives us N **data points**, each corresponding to some displacement x_i and its associated measured force F_i. We can plot these points on a graph, putting the independent variable x on the horizontal axis and the dependent variable $F(x)$ on the vertical axis. Each data point can be plotted as a point on this graph, as shown in Fig. 3. We now wish to test whether the data conform to a straight line, and if so, extract the value of the spring constant k. Even if the spring *does* obey Hooke's Law exactly, our data points will not lie exactly on a straight line. Due to measurement errors the data points will tend to lie a little bit above or below the true line.

The task for linear regression analysis is to take the data points and their estimated errors, and determine the *best fit* straight line. The best fit line is the line that is *most likely* to be the true line. Because of the measurement errors in our data points, we will never be certain what the true line is, just as we can never know the true value of a measured quantity. The best we can do in physics is to come as close as possible, and then estimate the error.

The functional form of a straight line can be written:

$$y = mx + b \tag{19}$$

where m is the slope and b is the y-intercept. Linear regression will determine the *best fit* values of m and b for a set of data points, in other words the values of m and b that are

most likely to be the true values. It will also determine their estimated errors σ_m and σ_b, such that the differences between the best fit values and the true values are expected to be less than the estimated errors. This is the best we can do; it is what we want to obtain from our experimental data.

At the heart of regression analysis is the **chi-squared** function:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - y_{fit}(x_i))^2}{\sigma_i^2}. \tag{20}$$

Here x_i and y_i are the values of x and y for each data point and σ_i is the combined *uncommon* estimated error in y for each point. The sum is taken over all N data points. The function $y_{fit}(x)$ is the best fit function, for example the straight line of Eq. 19 using the best fit values of m and b. It is important to understand the distinction between y_i and $y_{fit}(x_i)$: y_i is the *measured* value of y for the ith data point, and $y_{fit}(x_i)$ is the value of the function $y_{fit}(x)$ for $x = x_i$, where x_i is the measured value of x for the ith data point.

We need to take a closer look at what we mean by " σ_i is the combined *uncommon* estimated error in y for each point." Each data point represents a measurement of y, with some measurement error. This error can be statistical, systematic, or a combination of both. For example, to estimate the statistical error we might make repeated measurements of y for each value of x_i. Then for each x_i we would calculate the sample mean y_s for the repeated measurements, and use it for y_i in our data point x_i, y_i. There may also be a systematic error in y that should be combined with the statistical error, using Eq. 12, for each y_i. *Now here is the catch.* Only the **uncommon** errors should be combined to make the σ_i to use for regression analysis. These are the errors such that the true errors in y are expected to be different for different data points. In contrast, **common** errors are errors such that the true errors in y are expected to be the same for each data point. Statistical errors are random in nature, so they are always uncommon. Systematic errors can be either common or uncommon. A good example of an uncommon systematic error is the error due to the precision limit of an instrument. For each data point, the measured value of y is rounded to the nearest small division of the instrument. Each data point has a different value of y, therefore the magnitude and direction of this rounding is different at each point; the errors are uncommon. A good example of a common systematic error is the calibration accuracy of an instrument. An inaccuracy in the instrument calibration will cause all measurements to be wrong by the same amount or the same percentage, *e.g.* if a voltmeter calibration is too high by 0.01%, then all voltages measured by it will be too high by 0.01%; this is a common error.

You may have noticed that we have so far not considered the measurement error in the independent variable x. It is a measured quantity for each data point and as such it has some measurement error. Normally in regression analysis we bundle the error of x into the combined error of y, so that all of the error for a data point is taken to be in the dependent variable y, by using Eq. 15. With this idea we can write a general formula that can be used

to find the the combined uncommon estimated error σ_i for each data point:

$$\sigma_i = \sqrt{\sigma_i(\text{stat})^2 + \sigma_i(\text{uncommon sys})^2 + \left(\frac{\partial y}{\partial x}\right)^2 \sigma_i(x)^2} \tag{21}$$

Often the error in the independent variable x is relatively small and can be neglected (but not always – be careful).

The main premise of regression analysis is that the best fit function $y_{fit}(x)$ for a set of data points is the function that *minimizes* the chi-squared. This is known as the **principle of least squares**. We can calculate χ^2 using Eq. 20 for a variety of different functions $y_{fit}(x)$ with the same set of data points x_i, y_i and their errors σ_i, and each time we will obtain a different value for χ^2. The particular version of $y_{fit}(x)$ that gives the smallest value of χ^2 is considered the best fit function. The application of this principle is quite general, although there are some assumptions behind it that sometimes limit its use. These limitations won't be discussed here. For Advanced Lab you can always assume that the principle of least squares will determine the best fit function.

Focusing again on linear regression analysis, we use the straight line of Eq. 19 for $y_{fit}(x)$ in Eq. 20:

$$\chi^2(\text{linear}) = \sum_{i=1}^{N} \frac{(y_i - mx_i - b)^2}{\sigma_i^2}. \tag{22}$$

Now by differentiating χ^2 with respect to m and b, and setting the derivatives equal to zero, it can be shown that the following values of m and b give the minimum χ^2:

$$m = \frac{1}{\Delta}\left(\sum \frac{1}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2}\right) \tag{23}$$

$$b = \frac{1}{\Delta}\left(\sum \frac{x_i^2}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2}\right) \tag{24}$$

with:

$$\Delta = \sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2 \tag{25}$$

Each sum is taken over all N data points. Eqs. 23, 24, and 25 allow one to find the slope m and y-intercept b for the best fit straight line directly from the data points x_i, y_i and the combined uncommon errors σ_i. Notice that if the error σ_i is the *same* for each data point, then the σ_i cancel in Eqs. 23–25. In that case the best fit values are independent of the errors and the above expressions simplify to:

$$m = \frac{1}{D}\left(N \sum x_i y_i - \sum x_i \sum y_i\right) \tag{26}$$

$$b = \frac{1}{D}\left(\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i\right) \tag{27}$$

with:

$$D = N \sum x_i^2 - \left(\sum x_i\right)^2 \tag{28}$$

Remember that you can use Eqs. 26–28 *only* when the errors σ_i are equal for all points. If they are different then you must use Eqs. 23–25.

In order for linear regression to work, you always need more than two data points to fit. Obviously you cannot fit one point to a straight line. With two data points, you can always get a straight line by connecting the two points, regardless of their errors, but you will have no confidence that the data really have a linear relationship; any two points will always give a straight line. Only with three or more data points does linear regression actually tell you something. The **number of degrees of freedom** ν in regression analysis is defined to be the number of data points *minus* the number of parameters determined from the fit. In linear regression there are two parameters determined: m and b, so we have $\nu = N - 2$. For regression analysis to be valid, the number of degrees of freedom ν in the fit must be at least one; preferably much greater than one for a meaningful fit.

We can find estimated errors for the best fit parameters m and b by applying the error propagation formula (Eq. 15) to Eqs. 23–25. This gives:

$$\sigma_m = \sqrt{\frac{1}{\Delta} \sum \frac{1}{\sigma_i^2}} \tag{29}$$

$$\sigma_b = \sqrt{\frac{1}{\Delta} \sum \frac{x_i^2}{\sigma_i^2}} \tag{30}$$

where Δ is found using Eq. 25.

The principle of least squares allows us to do another important thing. We can use it to test the suitability of our proposed fit function for our set of data points and their errors. In the case of linear regression, we can test to see whether our data really do correspond to a straight line relationship between x and y. If the test fails, then we may want to consider using a different functional form, such as a parabola or an exponential, to fit our data. This test is called the χ^2 **test**. Looking at Eq. 20 we can see that, if the estimated errors are accurate, and $y_{fit}(x)$ is a good function to fit the data, then the expected deviation of a data point $y_i - y_{fit}(x_i)$ will be approximately equal to the estimated error σ_i for that point. So the value of χ^2 calculated using Eq. 20 for a good fit should be approximately equal to the number of data points. Furthermore, having more parameters to determine in the fit will tend to reduce the total χ^2, because more parameters allow the fit function $y_{fit}(x)$ to be adjusted to reduce the average deviation of the data. A more rigorous statement is that for a good fit, the total χ^2 should be approximately equal to the number of degrees of freedom ν. We can now define the **reduced chi-squared** χ^2_ν to be:

$$\chi^2_\nu = \frac{\chi^2}{\nu}, \tag{31}$$

and we can restate our good-fit criterian as: *for a good fit the reduced chi-squared* χ^2_ν *should be approximately equal to one.* The higher the number of degrees of freedom ν in the fit, the closer χ^2_ν should be to one. This idea can be quantified using the table in Appendix A (the last page of this paper). The table shows the probability $P(\chi^2_\nu, \nu)$ of obtaining a higher

reduced chi-squared when the best fit function $y_{fit}(x)$ is correct. To use this table, first find the number of degrees of freedom ν in your fit in the column at the left. Then move to the right along that row until you find the number that is closest to the reduced chi-squared from your fit. Then move up to the top row in that column, between the double lines. That number P is the probability of a higher χ^2_ν, *i.e.* a worse fit. If P is small, less than 0.05, then you had a **poor fit**. This means that either your data do not correspond to your chosen fit function *or* that your estimated errors were too small. If P is very large, greater than 0.95, then you had a **hyper fit** – your fit was too good. This usually means that your estimated errors were too large and should be reevaluated. If P falls in between 0.05 and 0.95 than you can conclude that you had a **good fit**. You can accept the results of your regression analysis only if you obtained a good fit.

There is something else we can do with the chi-squared. Occasionally we are faced with a set of data points for which we have no information with which to estimate the errors, but we have a lot of confidence in the form of the fit function. If we assume that the size of the error is the same for all points (not necessarily true, but reasonable since we have no other information), then we can fit the data to our fit function using regression analysis, and vary the size of the errors until the best fit reduced chi-squared equals one. This procedure gives a reasonable estimate for the error on the data points. The drawback of this is that it does not allow us to also test the goodness of fit to the assumed function.

Now let's look further at our example of linear regression analysis. We wish to test whether our spring obeys Hooke's Law (Eq. 18), and if so measure the spring constant k. We measure the force F using a strain gauge for a variety of displacements x. The results of our measurements are presented in Table 1. There are a total of eight data points. Now let's calculate the estimated errors. We measured the displacement using a meter stick with a smallest division of 1 mm, so the estimated systematic error in x is 0.05 cm. This is an uncommon error. The manufacturer's data sheet for the strain gauge states that the precision limit is ± 0.2 Newtons (an uncommon error), and the calibration accuracy is $\pm 3\%$ (a common error). The combined uncommon error for each data point is found using Eq. 21:

$$\sigma_i = \sqrt{(0.2)^2 + \left(\frac{dF}{dx}\right)^2 (0.05)^2} = \sqrt{0.04 + \left(\frac{6}{20.3}\right)^2 (0.0025)} = 0.201, \tag{32}$$

the same error for each point in this case. In the above I approximated:

$$\frac{dF}{dx} \approx \frac{\Delta F}{\Delta x} = \frac{6}{20.3} \tag{33}$$

between the first and last points, which is reasonable because we can see by eye that the data are approximately linear. It turns out that the measurement errors in x are negligible here; the combined error is dominated by the precision limit of the strain gauge. The data points are plotted in Fig. 3. The error σ_i on each point is shown as an **error bar**, a vertical line centered on the data point, extending $\pm\sigma_i$ above and below it, with a total length of $2\sigma_i$.

Since the error on each point is the same, we can use Eqs. 26–28 to calculate the slope m

x_i (cm)	F_i (N)	σ_i
0.0	0.0	0.20
2.5	0.8	0.20
4.8	1.2	0.20
8.7	2.4	0.20
12.4	3.6	0.20
14.9	4.0	0.20
18.8	5.6	0.20
20.3	6.0	0.20

Table 1: Data from a measurement of spring force F vs. displacement x.

and y-intercept b for the best-fit straight line:

$$D = (8)(1246.28) - (82.4)^2 = 3180 \tag{34}$$

$$m = \frac{1}{3180}\left((8)(359.96) - (82.4)(23.6)\right) = 0.294 \tag{35}$$

$$b = \frac{1}{3180}\left((1246.28)(23.6) - (82.4)(359.96)\right) = -0.078 \tag{36}$$

This best-fit line is shown as a solid line in Fig. 3. Notice that the line passes through most, but not all, of the error bars. This is normal. The true value of F for each point should usually lie within the error bar for that point, but it is OK for it to lie slightly (not far) outside the error bar sometimes. In fact, a good rule of thumb in regression analysis is that, with a good fit, the best fit function will pass through approximately two-thirds of the error bars on the data points.

We can use Eqs. 29-30 to find the estimated errors on the slope and offset:

$$\Delta = (198.88)(30982) - (2048.4)^2 = 1.966 \times 10^6 \tag{37}$$

$$\sigma_m = \sqrt{\frac{198.88}{1.966 \times 10^6}} = 0.010 \tag{38}$$

$$\sigma_b = \sqrt{\frac{30982}{1.966 \times 10^6}} = 0.126 \tag{39}$$

and we can calculate the chi-squared and reduced chi-squared using Eqs. 22 and 31:

$$\chi^2 = 4.426 \tag{40}$$

$$\chi^2_\nu = \frac{4.426}{6} = 0.738 \tag{41}$$

There are six degrees of freedom in this fit: eight data points minus two fit parameters m and b. Using the table in Appendix A, we see that for six degrees of freedom, our reduced chi-squared is closest to 0.762, so the probability of a worse fit is a bit higher than 60%. We

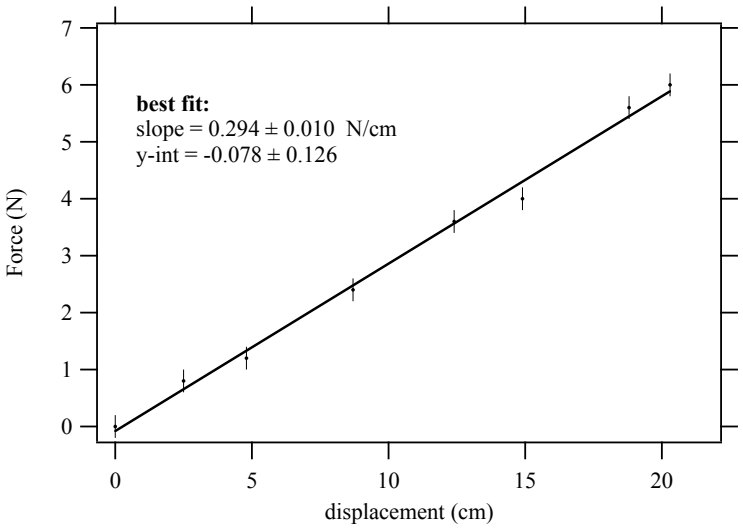

Figure 3: Experimental data, with error bars, of force vs. displacement for a spring, along with the best fit straight line (solid line).

can conclude that this was a good fit. Hooke's law is a straight line with a y-intercept of zero. The y-intercept of our best fit was $b = -0.078 \pm 0.126$, and zero falls within this error. We can conclude that our spring *does* obey Hooke's law $F = -kx$ at this level of precision. Finally, we have determined the spring constant k from our fit: $k = -m = 0.294 \pm 0.010$ N/cm. We are not quite done though, we still have the common calibration error of the strain gauge, which was not included in the fit, to account for. This error of $\pm 3.0\%$ will affect all the force measurements proportionally, so it is an additional error on k. We can add this to the estimated error from the fit by summing in quadrature (Eq. 12):

$$\sigma_k = \sqrt{(0.010)^2 + (0.030 \cdot k)^2} = 0.013 \tag{42}$$

So our experiment and linear regression analysis of the data gave us a final measured value of $\boxed{k = 0.294 \pm 0.013 \text{ N/cm.}}$ The true value of the spring constant probably lies within this range.

Another type of regression analysis, related to linear regression, is **log-linear regression**. This method is useful to fit data to an exponential function:

$$y(x) = C\mathrm{e}^{\lambda x}, \tag{43}$$

with x as the independent variable and c, λ the two constants to be determined by the fit. We begin by taking the natural logarithm of both sides of Eq. 43:

$$\ln y = \ln C + \lambda x. \tag{44}$$

We see that we now have a straight line function that can be fit using linear regression. The y-intercept is $\ln C$ and the slope is λ. The data points need to use are $\ln y_i, x_i$. We also need

to convert the errors σ_i on y_i into errors $\sigma(\ln y)_i$ on $\ln y_i$, using Eq. 15:

$$\sigma(\ln y)_i = \frac{d}{dy}(\ln y)\sigma_i = \frac{\sigma_i}{y_i}. \tag{45}$$

We can now fit Eq. 44 to our data using linear regression as before, obtain the slope m and y-intercept b, and we equate λ with m and C with e^b.

We use a similar idea in **log-log** regression. This is useful to fit data to a power function:

$$y(x) = ax^n, \tag{46}$$

where we wish to find the coefficient a and the power n. Again we take the natural log of both sides:

$$\ln y = \ln a + n \ln x \tag{47}$$

which gives us a straight line function to fit using linear regression. In this case the y-intercept is $\ln a$ and the slope is the power n. We fit the data points $\ln y_i, \ln x_i$ with the errors on $\ln y$ calculated using Eq. 45.

The above discussion shows how linear regression is used to fit data to a straight line, an exponential, or a power function. In general, regression analysis can be used to fit a set of data points to any desired function. It is always just a matter of finding the function that minimizes χ^2 in Eq. 20. For more complicated functions this can become quite difficult, and is usually best accomplished by a numerical computer calculation. If you find that you need to use non-linear regression analysis in Advanced Lab, you should discuss it with your instructor or TA first to decide how it can best be done.

6 Significant Figures

The number of **significant figures** refers to the precision in stating a number. For example $\pi = 3.14$ has 3 significant figures, $\pi = 3.14159$ has six, and $\pi = 3.141592654$ has 10. Numbers on both sides of the decimal point are counted: 138.234 has six significant figures. Leading zeros don't count, but trailing zeros are *assumed* to be significant, *e.g.* 0137.67500 has 8 significant figures. You should write trailing zeros in a number only if they are significant, for example to round 23.67983 to 5 significant figures you write 23.680, not 23.68 and not 23.68000.

When stating an experimental result it is conventional to give the error to two significant figures, and the measured value to the same absolute precision. For example you may write that you measured the mass of an object to be $m = 237.6 \pm 3.8$ g. You should not write $m = 237.563728 \pm 3.8$ g because the extra significant figures in the measured value are meaningless given the size of the estimated error.

7 Outlying Data Points

Sometimes one or two data points don't seem to fit in with the rest. These points are called **outlying data points**, or **outliers** for short. A good working definition of an outlier is a data point that, if you include it when fitting your data, causes the reduced chi-squared to become very large and results in a poor fit. The following is the accepted procedure for dealing with outliers:

1. First, think carefully about the conditions under which the outliers were taken. This is a situation where thoughtful and detailed lab note-taking really pays off. Review your notes and try to determine the reason(s) for the anomalous points. Were any settings changed then? Did you note any unusual events or conditions? If you can explain the outliers as a result of an unusual event or poor experimental conditions then you may eliminate those points from your data set. You must be very honest about this though! Don't stretch too hard to find a reason for the outliers.

2. If you cannot explain the outliers, then you must live with them. Usually there are good reasons for outliers, even if you can't find them. Sometimes they result from interesting physics that you haven't considered or don't know about. So don't worry too much – you didn't necessarily do anything wrong. The important thing is to be honest. If you present your data in a written report or oral presentation be sure to include any outliers in your tables and plots and point them out. Maybe a reader or audience member knows something that you don't and can offer a plausible explanation.

3. If you are fitting your data to a particular function, consider the possibility that you are using the wrong function. This is especially true if the data seem to follow a trend that your fitting function cannot accommodate.

4. If most of your data seem to behave according to your assumed fitting function, except for one or two outliers, then it is acceptable to exclude the outliers from your fit. You must still, however, put them on your plot and make clear in your report which points were considered outliers and excluded from the fit.

5. Outliers should be rare. If you have many of them you probably did something wrong in your experiment and you should repeat it.

8 Further Reading:

1. Bevington, *Data Reduction and Error Analysis for the Physical Sciences*, McGraw Hill, 1969, QA 278.B48; Chapters 1–6.

2. Lyons, *Statistics for Nuclear and Particle Physicists*, Cambridge, 1986, QC 776.L96; Chapters 1–5.

3. Dowdy and Weardin, *Statistics for Research*, Wiley, 1983, QA 276.D66; Chapters 1–7.

Appendix A

From Bevington, *Data Reduction and Error Analysis for the Physical Sciences*, McGraw Hill, 1969

Table C-4 χ^2 distribution. Values of the reduced chi-square $\chi_r^2 = \chi^2/\nu$ corresponding to the probability $P_\chi(\chi^2, \nu)$ of exceeding χ^2 vs. the number of degrees of freedom ν

ν \ P	0.99	0.98	0.95	0.90	0.80	0.70	0.60	0.50
1	0.00016	0.00063	0.00393	0.0158	0.0642	0.148	0.275	0.455
2	0.0100	0.0202	0.0515	0.105	0.223	0.357	0.511	0.693
3	0.0383	0.0617	0.117	0.195	0.335	0.475	0.623	0.789
4	0.0742	0.107	0.178	0.266	0.412	0.549	0.688	0.839
5	0.111	0.150	0.229	0.322	0.469	0.600	0.731	0.870
6	0.145	0.189	0.273	0.367	0.512	0.638	0.762	0.891
7	0.177	0.223	0.310	0.405	0.546	0.667	0.785	0.907
8	0.206	0.254	0.342	0.436	0.574	0.691	0.803	0.918
9	0.232	0.281	0.369	0.463	0.598	0.710	0.817	0.927
10	0.256	0.306	0.394	0.487	0.618	0.727	0.830	0.934
11	0.278	0.328	0.416	0.507	0.635	0.741	0.840	0.940
12	0.298	0.348	0.436	0.525	0.651	0.753	0.848	0.945
13	0.316	0.367	0.453	0.542	0.664	0.764	0.856	0.949
14	0.333	0.383	0.469	0.556	0.676	0.773	0.863	0.953
15	0.349	0.399	0.484	0.570	0.687	0.781	0.869	0.956
16	0.363	0.413	0.498	0.582	0.697	0.789	0.874	0.959
17	0.377	0.427	0.510	0.593	0.706	0.796	0.879	0.961
18	0.390	0.439	0.522	0.604	0.714	0.802	0.883	0.963
19	0.402	0.451	0.532	0.613	0.722	0.808	0.887	0.965
20	0.413	0.462	0.543	0.622	0.729	0.813	0.890	0.967
22	0.434	0.482	0.561	0.638	0.742	0.823	0.897	0.970
24	0.452	0.500	0.577	0.652	0.753	0.831	0.902	0.972
26	0.469	0.516	0.592	0.665	0.762	0.838	0.907	0.974
28	0.484	0.530	0.605	0.676	0.771	0.845	0.911	0.976
30	0.498	0.544	0.616	0.687	0.779	0.850	0.915	0.978
32	0.511	0.556	0.627	0.696	0.786	0.855	0.918	0.979
34	0.523	0.567	0.637	0.704	0.792	0.860	0.921	0.980
36	0.534	0.577	0.646	0.712	0.798	0.864	0.924	0.982
38	0.545	0.587	0.655	0.720	0.804	0.868	0.926	0.983
40	0.554	0.596	0.663	0.726	0.809	0.872	0.928	0.983
42	0.563	0.604	0.670	0.733	0.813	0.875	0.930	0.984
44	0.572	0.612	0.677	0.738	0.818	0.878	0.932	0.985
46	0.580	0.620	0.683	0.744	0.822	0.881	0.934	0.986
48	0.587	0.627	0.690	0.749	0.825	0.884	0.936	0.986
50	0.594	0.633	0.695	0.754	0.829	0.886	0.937	0.987
60	0.625	0.662	0.720	0.774	0.844	0.897	0.944	0.989
70	0.649	0.684	0.739	0.790	0.856	0.905	0.949	0.990
80	0.669	0.703	0.755	0.803	0.865	0.911	0.952	0.992
90	0.686	0.718	0.768	0.814	0.873	0.917	0.955	0.993
100	0.701	0.731	0.779	0.824	0.879	0.921	0.958	0.993
120	0.724	0.753	0.798	0.839	0.890	0.928	0.962	0.994
140	0.743	0.770	0.812	0.850	0.898	0.934	0.965	0.995
160	0.758	0.784	0.822	0.860	0.905	0.938	0.968	0.996
180	0.771	0.796	0.833	0.868	0.910	0.942	0.970	0.996
200	0.782	0.806	0.841	0.874	0.915	0.945	0.972	0.997

Table C-4 χ^2 distribution (*continued*)

ν \ P	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.708	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.916	1.204	1.609	2.303	2.996	3.912	4.605	6.908
3	0.982	1.222	1.547	2.084	2.605	3.279	3.780	5.423
4	1.011	1.220	1.497	1.945	2.372	2.917	3.319	4.617
5	1.026	1.213	1.458	1.847	2.214	2.678	3.017	4.102
6	1.035	1.205	1.426	1.774	2.099	2.506	2.802	3.743
7	1.040	1.198	1.400	1.717	2.010	2.375	2.639	3.475
8	1.044	1.191	1.379	1.670	1.938	2.271	2.511	3.266
9	1.046	1.184	1.360	1.632	1.880	2.187	2.407	3.097
10	1.047	1.178	1.344	1.599	1.831	2.116	2.321	2.959
11	1.048	1.173	1.330	1.570	1.789	2.056	2.248	2.842
12	1.049	1.168	1.318	1.546	1.752	2.004	2.185	2.742
13	1.049	1.163	1.307	1.524	1.720	1.959	2.130	2.656
14	1.049	1.159	1.296	1.505	1.692	1.919	2.082	2.580
15	1.049	1.155	1.287	1.487	1.666	1.884	2.039	2.513
16	1.049	1.151	1.279	1.471	1.644	1.852	2.000	2.453
17	1.048	1.148	1.271	1.457	1.625	1.823	1.965	2.399
18	1.048	1.145	1.264	1.444	1.604	1.797	1.934	2.351
19	1.048	1.142	1.258	1.432	1.586	1.773	1.905	2.307
20	1.048	1.139	1.252	1.421	1.571	1.751	1.878	2.266
22	1.047	1.134	1.241	1.401	1.542	1.712	1.831	2.194
24	1.046	1.129	1.231	1.383	1.517	1.678	1.791	2.132
26	1.045	1.125	1.223	1.368	1.496	1.648	1.755	2.079
28	1.045	1.121	1.215	1.354	1.476	1.622	1.724	2.032
30	1.044	1.118	1.208	1.342	1.459	1.599	1.696	1.990
32	1.043	1.115	1.202	1.331	1.444	1.578	1.671	1.953
34	1.042	1.112	1.196	1.321	1.429	1.559	1.649	1.919
36	1.042	1.109	1.191	1.311	1.417	1.541	1.628	1.888
38	1.041	1.106	1.186	1.303	1.405	1.525	1.610	1.861
40	1.041	1.104	1.182	1.295	1.394	1.511	1.592	1.835
42	1.040	1.102	1.178	1.288	1.384	1.497	1.576	1.812
44	1.039	1.100	1.174	1.281	1.375	1.485	1.562	1.790
46	1.039	1.098	1.170	1.275	1.366	1.473	1.548	1.770
48	1.038	1.096	1.167	1.269	1.358	1.462	1.535	1.751
50	1.038	1.094	1.163	1.263	1.350	1.452	1.523	1.733
60	1.036	1.087	1.150	1.240	1.318	1.410	1.473	1.660
70	1.034	1.081	1.139	1.222	1.293	1.377	1.435	1.605
80	1.032	1.076	1.130	1.207	1.273	1.351	1.404	1.560
90	1.031	1.072	1.123	1.195	1.257	1.329	1.379	1.525
100	1.029	1.069	1.117	1.185	1.243	1.311	1.358	1.494
120	1.027	1.063	1.107	1.169	1.221	1.283	1.325	1.446
140	1.026	1.059	1.099	1.156	1.204	1.261	1.299	1.410
160	1.024	1.055	1.093	1.146	1.191	1.243	1.278	1.381
180	1.023	1.052	1.087	1.137	1.179	1.228	1.261	1.358
200	1.022	1.050	1.083	1.130	1.170	1.216	1.247	1.338