Statistics and Data Analysis

Jeremiah Mans

November 2, 2017

For use with PHYS3605W (Modern Physics Laboratory) at the University of Minnesota

1 Communicating a Scientific Measurement

Science, as we learn in elementary school, is about the formulation and testing of hypotheses. In the examples of the "Scientific Method" which are presented to elementary school students, hypotheses are proposed, tested, and either proven or disproven. The real scientific process is rarely so simple or direct. Properly, a hypothesis is never "proven" or "disproven". Instead, the data support or exclude a given hypothesis to a well-defined extent. The essence of the scientific process is the measurement of a process or observable and *the determination of the uncertainty on that measurement.* Indeed, the uncertainty of a measurement is often rather more important than the specific value measured.

Scientific measurements, therefore, are usually presented complete with uncertainties. As a example, the mass of the neutral elementary particle called the Z boson is an important parameter in the Standard Model of particle physics. With careful experimental and theoretical work, the mass of the boson has been measured to be:

$$m_Z = 91.1876 \pm 0.0021 \text{ GeV} \tag{1}$$

The measurement, therefore, has an uncertainty of 2.1 MeV. As we will see below, an uncertainty presented in this manner means that there is 68% chance that the actual true mass of the Z is somewhere between 91.1855 GeV and 91.1897 GeV. This precision corresponds to a measurement precision of 0.002%, which is very precise for any measurement. As we will see below, issues can arise with any measurement which make even a 1% precision impossible.

The second important observation about scientific process is *communication.* The scientific process is not complete until the results of the experiment or calculation have been communicated to the community. In long tradition, the laurels for discovery of a new process, phenomenon or insight goes to the first to communicate the result in a proper scientific manner. It is not enough to make a brilliant measurement and put the results into a desk drawer – scientific results must be communicated to really constitute science rather than being a hobby. Scientific communication is different than press-releases and promotional materials, however. Scientific communication must be designed for clarity on the data and processes used for the result presented and must clearly establish the bounds of what is claimed. Press releases and promotional materials often are intentionally vague on these points!

1.1 Significant Figures

When expressing a measured quantity or a quantity which is derived from a measurement, the uncertainty should be clearly expressed. One aspect of expressing the uncertainty is the number of digits used for the presentation. This concept can be most clearly understood when numbers are expressed in scientific notation. If we recast the measurement of the Z boson mass into scientific notation, using eV as the base unit, we obtain

$$m_Z = 9.11876 \times 10^{10} \pm 2.1 \times 10^6 \text{ eV}$$

The mass of the Z boson is expressed with six significant digits and the error is expressed with two. A zero at least-significant end of a number can be a significant digit, but not a zero at the most-significant end.

Uncertainties should usually be presented with two digits of accuracy. It is foolish to imagine that one understands a uncertainty to a level better than 5%. Measured quantities should be presented to the same least-significant digit as the uncertainty. Since two digits of accuracy on the Z boson mass implies a precision down to 10^5 eV, the measurement is presented to the same least-significant digit.

The basic rules for combining quantities are as follows:

1. When multiplying two numbers, the result should have the same number of significant digits as the quantity with the least number of significant digits.

   Example:

   $$34.189 \times 2.1 = 72$$

2. When adding or subtracting, the number of decimal places of the result should be the smaller of the number of decimal places of any of the numbers combined.

   Example:

   $$120.34 + 7.4 = 127.7$$

3. Exact numbers have an unlimited number of effective significant digits. As an example, the diameter of a circle is defined as twice the radius so the diameter of circle of radius 3.71 cm is

   $$2 \times 3.71 \text{ cm} = 7.42 \text{ cm}$$

4. When performing a sequence of several multiplication steps, one should retain extra significant figures for the computation. Otherwise, repeated rounding effects can results in an incorrect computation. The final result should be presented with the correct number of significant digits.

The use of significant digits is not a substitute for proper error propagation, as discussed below, but is an important part of clear scientific communication.

2 Uncertainties and Propagation of Errors

The Z boson mass measurement discussed above was produced at the LEP collider in the 1990s. While the particle accelerator and detectors are complex devices, the measurement of the Z boson mass depended on an accurate measurement of the magnetic field required to keep the electrons moving in a circle.

The force required to keep a particle moving in a circle of radius R is

$$\frac{dp}{dt} = \frac{pv}{R}$$

We use the momentum p in this expression as it makes the expression relativistically correct.

The relativistically-correct force from a magnetic field in vacuum is

$$\frac{d\vec{p}}{dt} = q\vec{v} \times \vec{B}$$

Since the magnetic field must provide the force to keep the electron moving in a circle in the collider, we obtain

$$\begin{aligned} \frac{pv}{R} &= qvB \\ p_e &= qRB \end{aligned}$$

Since the electron momentum is very large compared with its mass, we can approximate its mass as

$$E_e = p_e c$$

When the electron and positron collided at LEP, their full energy went into producing the rest mass of the Z boson through the Einstein relation $E = mc^2$.

$$\begin{aligned} m_Z c^2 &= E_{\text{collision}} = 2E_e \\ m_z c^2 &= 2qRBc \\ m_z &= \frac{2qRB}{c} \end{aligned}$$

Of the expressions in the mass of the Z, c is speed of light, which has an exact defined value in the SI system of units. The charge of the electron q_e is a universal quantity is measured to be $(1.602176565 \pm 0.000000035) \times 10^{-19}$ C. The other two critical inputs are parameters of the accelerator – the effective accelerator radius and the magnetic field. These two were measured extremely precisely to determine the Z boson mass.

Given the inputs, it is quite straightforward to determine the best value for the Z mass – insert the best measured values into the equation! However, how do we determine the uncertainty for the Z mass measurement based on the inputs we have?

2.1 Mean, Median, Mode, and Variance

To understand this question, we need to step back and consider first the measurement of one of these quantities to derive a number of important concepts: the mean, median, mode and variance. There are two primary schools of thought in statistical analysis: the frequentist and the Bayesian. We will derive these quantities from the frequentist point of view, but they are valid and useful concepts in either statistical framework.

Statistical analysis assumes that a given data set can be described by some mathematical model of the likely observations for data points from the set. Repeated observations provide more information about this true "parent" distribution and the observed distribution will more closely match the parent distribution as the number of observations in the data set increases. The core of statistical analysis is to estimate the true values of a parameter using a limited set of data and to determine the uncertainty of that estimate.

To begin, let us consider the repeated measurement of the effective radius of the LEP collider. Suppose we measure the effective radius fifteen times and obtain the set of data shown in Table 1.

Table 1: Hypothetical measurements of the effective radius of the LEP collider in units of kilometers. The effective radius is smaller than the physical radius as the bending dipoles do not fill the full circumference of the accelerator.

3.10014 km	3.10010 km	3.10007 km
3.10010 km	3.10003 km	3.10009 km
3.10010 km	3.10004 km	3.09986 km
3.09997 km	3.09997 km	3.10011 km
3.09977 km	3.09976 km	3.10003 km

In a set of data, the *mean* is defined as the sum of the data divided by the number of data points. Colloquially, this quantity is also called the *average*.

$$\bar{x} = \frac{1}{N} \sum_i x_i \tag{2}$$

For our example data set, $\bar{x} = 3.10001$ km. When N becomes very large, we assume that the data set mean approaches the mean of the parent distribution which was the source of the data.

$$\mu = \lim_{N \to \infty} \left(\frac{1}{N} \sum_i x_i \right) \tag{3}$$

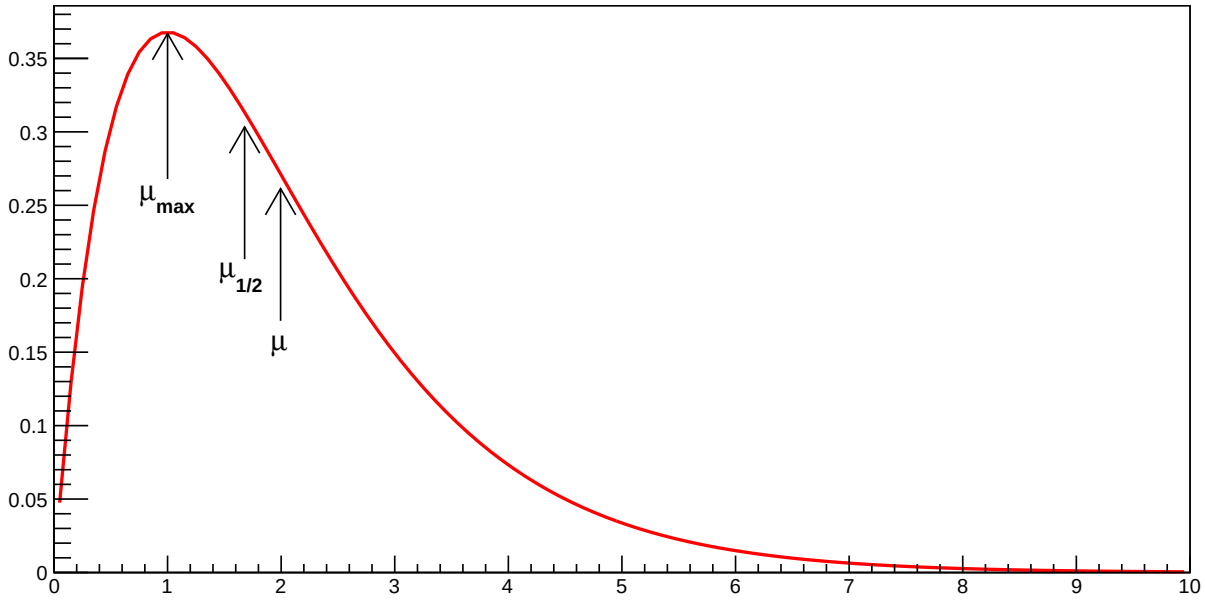


Figure 1: Mean, median, and mode for the probability density function $P(x) = xe^{-x}$.

The *median* of the data set is the measurement in the middle of the ordered list. Expressed differently, the median is the the value such that half the data points have a larger value and half have a smaller value. For a symmetric data set, the mean and median are often very similar, however for an asymmetric distribution the mean and median can be quite different as seen in Fig. 1. The median for our dataset is $\mu_{1/2} = 3.10004$ km.

The *mode* of a data set is the most probable value to observe. For a real-valued observable, often each observation is distinct as is the case for our data set. The better-defined quantity is the maximum of the parent distribution $\mu_{\max}$. For some distributions, the mode and the mean are the same, while for others they are quite different as we will see below when we study some specific probability density functions.

Once we have defined the mean, we can then study the variation of the measurements from that mean. We will use the *standard deviation* σ as our metric for deviations from the mean. In general, we would like to understand the spread of the measurements around the mean. This means that we would not like to have positive and negative variations cancel each other out. The simplest way to ensure this is to use the variance, which uses the sum of squares.

$$\sigma^2 = \lim_{N\to\infty}\left[\frac{1}{N}\sum_i (x_i - \mu)^2\right] = \lim_{N\to\infty}\left[\frac{1}{N}\sum_i x_i^2\right] - \mu^2 \tag{4}$$

This definition of the standard deviation is correct when the mean is known independently of these observations, which is a rare case. If the mean is determined from the same dataset, the standard deviation must be a bit larger to account for that fact. This is expressed by the equation below:

$$s^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2 \tag{5}$$

For our set of data, we observe $s = 0.00012$ km For large values of N, $\sigma \approx s$ and we will discuss the uncertainty using the symbol σ as is typical in most texts. For data measurement purposes, uncertainties can come directly from the measurement (for example, a distance measurement precision is limited by the apparatus in use) or from statistical considerations. As we will see below, multiple measurements with random variations can allow an effective precision better than that of the apparatus in some situations.

2.2 Error Propagation

Let's consider a general derived quantity x which is composed from several measured inputs u, v, w. The quantities have been measured multiple times, resulting in a data set of u_i, v_i, w_i values and average values for each variable $\bar{u}, \bar{v}, \bar{w}$. Each measured input has an uncertainty which is represented by σ_u for the u variable, σ_v for the v variable and so on. We can determine the total uncertainty for x by performing a Taylor expansion of the quantity around the average values of each of the measured inputs.

$$x_i - \bar{x} = (u_i - \bar{u})\left(\frac{\partial x}{\partial u}\right) + (v_i - \bar{v})\left(\frac{\partial x}{\partial v}\right) + (w_i - \bar{w})\left(\frac{\partial x}{\partial w}\right) \tag{6}$$

We can then calculate the standard deviation for x using Eq.4.

$$\begin{aligned}
\sigma_x^2 &= \lim_{N\to\infty}\frac{1}{N}\sum_i \left[(u_i - \bar{u})\left(\frac{\partial x}{\partial u}\right) + (v_i - \bar{v})\left(\frac{\partial x}{\partial v}\right) + (w_i - \bar{w})\left(\frac{\partial x}{\partial w}\right)\right]^2 \\
&= \lim_{N\to\infty}\frac{1}{N}\sum_i \left[(u_i - \bar{u})^2\left(\frac{\partial x}{\partial u}\right)^2 + (v_i - \bar{v})^2\left(\frac{\partial x}{\partial v}\right)^2 + (w_i - \bar{w})^2\left(\frac{\partial x}{\partial w}\right)^2\right. \\
&\quad \left. + 2(u_i - \bar{u})(v_i - \bar{v})\left(\frac{\partial x}{\partial u}\frac{\partial x}{\partial v}\right) + 2(u_i - \bar{u})(w_i - \bar{w})\left(\frac{\partial x}{\partial u}\frac{\partial x}{\partial w}\right) + 2(v_i - \bar{v})(w_i - \bar{w})\left(\frac{\partial x}{\partial v}\frac{\partial x}{\partial w}\right)\right]
\end{aligned}$$

The expression $\sum(u_i - \bar{u})^2$ is the same as the standard deviation or uncertainty of the u variable. This allows us to re-write the first terms of the standard deviation of x:

$$\begin{aligned}
\sigma_x^2 &= \sigma_u^2\left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2\left(\frac{\partial x}{\partial v}\right)^2 + \sigma_w^2\left(\frac{\partial x}{\partial w}\right)^2 \\
&\quad + \lim_{N\to\infty}\frac{1}{N}\sum_i \left[2(u_i - \bar{u})(v_i - \bar{v})\left(\frac{\partial x}{\partial u}\frac{\partial x}{\partial v}\right) + 2(u_i - \bar{u})(w_i - \bar{w})\left(\frac{\partial x}{\partial u}\frac{\partial x}{\partial w}\right) + 2(v_i - \bar{v})(w_i - \bar{w})\left(\frac{\partial x}{\partial v}\frac{\partial x}{\partial w}\right)\right]
\end{aligned}$$

To simplify the second set of terms, we need to define the *covariance*

$$\sigma_{uv}^2 = \lim_{N\to\infty}\frac{1}{N}\sum_i [(u_i - \bar{u})(v_i - \bar{v})]$$

With this notation, we can write the full error propagation equation as

$$\sigma_x^2 = \sigma_u^2\left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2\left(\frac{\partial x}{\partial v}\right)^2 + \sigma_w^2\left(\frac{\partial x}{\partial w}\right)^2 + 2\sigma_{uv}^2\left(\frac{\partial x}{\partial u}\frac{\partial x}{\partial v}\right) + 2\sigma_{uw}^2\left(\frac{\partial x}{\partial u}\frac{\partial x}{\partial w}\right) + 2\sigma_{vw}^2\left(\frac{\partial x}{\partial v}\frac{\partial x}{\partial w}\right) \tag{7}$$

If the inputs u and v are not correlated in any way, then σ_{uv}^2 will tend to zero. This is quite common and in this situation the error propagation equation simplifies to

$$\sigma_x^2 = \sigma_u^2\left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2\left(\frac{\partial x}{\partial v}\right)^2 + \sigma_w^2\left(\frac{\partial x}{\partial w}\right)^2 \tag{8}$$

Care must be used in assuming that the correlation is zero, however. In the case of the precision measurement of the Z boson mass, the radius value and the magnetic field value can have some correlation depending on how each is measured. If the inputs are to be assumed uncorrelated, this should stated clearly in any paper or write-up of the data analysis, including in a logbook.

3 Parameter Estimation and Linear Fitting

3.1 The Weighted Mean

In the definition of the mean as the best estimator of the true value of an observable, we assumed that each measurement had the same uncertainty. If this is not true, the simple average will give the incorrect result.

This is clear if we consider three measurements of the temperature of a semiconductor sample using different techniques which provide different precision:

$$310.30 \pm 0.20 \text{ K} \quad 310.70 \pm 0.50 \text{ K} \quad 311.8 \pm 2.0 \text{ K}$$

Notice that all measurements are *consistent*: they agree within uncertainties. The simple mean gives a value of 310.9 K. This result appears to be *inconsistent* with the most precise measurement! We can determine the uncertainty using propagation of errors:

$$\begin{aligned} \bar{x} &= \frac{x_1}{3} + \frac{x_2}{3} + \frac{x_3}{3} \\ \sigma_{\bar{x}}^2 &= \sum \sigma_i^2 \left(\frac{\partial \bar{x}}{\partial x_i} \right)^2 \\ &= \sum \sigma_i^2 \left(\frac{1}{N} \right)^2 \end{aligned}$$

We find that the square root of the variance is 0.7 K, which is significantly larger than the uncertainty on the most precise measurement, higher than the middle measurement, and much lower than the uncertainty on either of the other two measurements.

We can do better by seeking a definition of the mean which minimizes the difference between the mean and the input measurements, when considered in units of the uncertainty of each measurement. We do this by defining a new quantity χ (chi):

$$\chi_i = \frac{x_i - \bar{x}}{\sigma_i}$$

We then minimize the sum of squares of χ to find the best definition of $\bar{x}$.

$$\begin{aligned} \chi^2 &= \sum \frac{(x_i - \bar{x})^2}{\sigma_i^2} \\ \frac{\partial \chi^2}{\partial \bar{x}} &= \sum -2 \frac{x_i - \bar{x}}{\sigma_i^2} \\ &= -2 \left[\sum \frac{x_i}{\sigma_i^2} \right] + 2 \left[\sum \frac{\bar{x}}{\sigma_i^2} \right] \\ 0 &= -2 \left[\sum \frac{x_i}{\sigma_i^2} \right] + 2\bar{x} \sum \frac{1}{\sigma_i^2} \\ \bar{x} \sum \frac{1}{\sigma_i^2} &= \sum \frac{x_i}{\sigma_i^2} \\ \bar{x} &= \frac{\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}} \end{aligned}$$

Each measurement is thus weighted by the reciprocal of its variance – measurements with larger variances are weighted less than measurements will small variances in the final result. For the variance itself, we can apply the standard process of error propagation and find the variance.

$$\begin{aligned} \sigma_{\bar{x}}^2 &= \sum \sigma_i^2 \left(\frac{\partial \bar{x}}{\partial x_i} \right)^2 \\ &= \left[\sum \frac{1}{\sigma_i^2} \right]^{-2} \sum \sigma_i^2 \left(\frac{1}{\sigma_i^2} \right)^2 \\ &= \left[\sum \frac{1}{\sigma_i^2} \right]^{-2} \sum \frac{1}{\sigma_i^2} = \left[\sum \frac{1}{\sigma_i^2} \right]^2 = \frac{1}{\sum \frac{1}{\sigma_i^2}} \end{aligned}$$

Applying this procedure to our measurements, we obtain 310.37 ± 0.18 K. The two higher-precision measurements dominate the measurement, while all measurements remain statistically consistent with the new weighted mean. The uncertainty has been improved by the addition of the new information, compared with the highest-precision single measurement alone.

3.2 Linear Fitting

To this point, we have considered only the measurement of individual quantities. However, often we are interested in understanding the behavior of an observable as a function of another parameter. For example, consider the data set below on the thermal conductivity of lead as a function of applied pressure.

Applied Pressure [kg/cm^2]	Thermal Conductivity
2000	1165.6 ± 3.1
4000	1205.0 ± 2.5
6000	1250.0 ± 2.5
8000	1271.3 ± 6.3
10000	1312.5 ± 8.8
11900	1347.5 ± 2.5

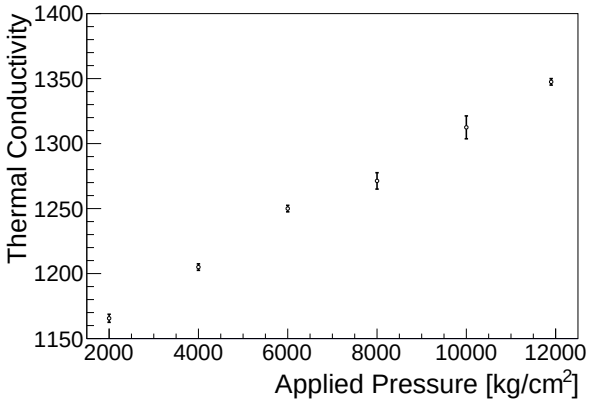


Figure 2: Data on the thermal conductivity of lead, adapted from Fig.4, p.102 of "The Effect of Pressure on the Thermal Conductivity of Metals", P. W. Bridgman, **Proceedings of the American Academy of Arts and Sciences**, Vol. 57, No. 5 (Apr., 1922)

If we plot the data, we observe what appears to be a linear relationship. It would appear to be reasonable to model the data using form such as

$$C = a + bp$$

where the parameters b (slope) and a (offset) should be determined from the data. We can determine the values of a and b using the χ^2 formalism which we developed for the weighted mean.

Let us consider an observable y and a control parameter x. We will assume that the fractional uncertainty on y is much larger than the fractional uncertainty on x. If this not the case, the fit is unlikely to give much useful information! For each of the N values of the control parameter x_i considered, we have an observation y_i with an uncertainty σ_i.

Let us assume we model the data using the function $f(x; a, b) = a + bx$. The parameters a, b are the two *parameters* of the function. If we included additional terms in our fit, we would have three or more parameters. Note that for a successful fit, we must have more data points than parameters so that we enough observations to uniquely determine all the parameters – generally we want several times more observations than parameters.

For our linear fit, we can write the χ^2

$$\chi^2 = \sum_i \frac{[y_i - (a + bx_i)]^2}{\sigma_i^2} \tag{9}$$

We determine each parameter by minimizing the χ^2 with respect to each parameter individually. This process yields a pair of linear equations:

$$\frac{\partial \chi^2}{\partial a} = 0 \quad = \quad \sum_i -2\frac{[y_i - a - bx_i]}{\sigma_i^2}$$

$$\frac{\partial \chi^2}{\partial b} = 0 \quad = \quad \sum_i -2\frac{[y_i - a - bx_i]}{\sigma_i^2} x_i$$

Collecting terms, we can reorganize the two equations to

$$\begin{aligned}\sum_i \frac{y_i}{\sigma_i^2} &= a\left[\sum_i \frac{1}{\sigma_i^2}\right] + b\left[\sum_i \frac{x_i}{\sigma_i^2}\right] \\ \sum_i \frac{y_i x_i}{\sigma_i^2} &= a\left[\sum_i \frac{x_i}{\sigma_i^2}\right] + b\left[\sum_i \frac{x_i^2}{\sigma_i^2}\right]\end{aligned}$$

To solve these equations, we can use the method of determinants. To express this simply, we define

$$\Delta = \begin{vmatrix} \sum_i \frac{1}{\sigma_i^2} & \sum_i \frac{x_i}{\sigma_i^2} \\ \sum_i \frac{x_i}{\sigma_i^2} & \sum_i \frac{x_i^2}{\sigma_i^2} \end{vmatrix} = \sum \frac{x_i^2}{\sigma_i^2} \sum \frac{1}{\sigma_i^2} - \left[\sum \frac{x_i}{\sigma_i^2}\right]^2$$

With this,

$$\begin{aligned} a &= \frac{1}{\Delta}\left[\sum \frac{x_i^2}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2}\right] \\ b &= \frac{1}{\Delta}\left[\sum \frac{1}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2}\right] \end{aligned}$$

The uncertainties on the parameters can be determined by error propagation to be

$$\begin{aligned} \sigma_a^2 &= \frac{1}{\Delta} \sum \frac{x_i^2}{\sigma_i^2} \\ \sigma_b^2 &= \frac{1}{\Delta} \sum \frac{1}{\sigma_i^2} \end{aligned}$$

If we define the following "sum-symbols",

$$S_x = \sum \frac{x_i}{\sigma_i^2} \quad S_{xx} = \sum \frac{x_i^2}{\sigma_i^2} \quad S_{xy} = \sum \frac{x_i y_i}{\sigma_i^2}$$
$$S_y = \sum \frac{y_i}{\sigma_i^2} \quad S = \sum \frac{1}{\sigma_i^2}$$

we can then define these formulae in a more compact manner:

$$\begin{aligned} \Delta &= S_{xx} S - S_x^2 \\ a &= \frac{1}{\Delta}\left[S_{xx} S_y - S_x S_{xy}\right] \\ b &= \frac{1}{\Delta}\left[S S_{xy} - S_x S_y\right] \\ \sigma_a^2 &= \frac{1}{\Delta} S_{xx} \\ \sigma_b^2 &= \frac{1}{\Delta} S \end{aligned}$$

3.3 Interpreting Fit Results

For the lead thermal conductivity data, we find $a = 1133.9 \pm 2.6$ and $b = 0.01811 \pm 0.00035$ cm^2/kg. We can compare our fit to the measurements using both a table and a figure, as shown in Fig. 3. As can be seen, most of the points agree with the prediction, such that the χ^2 for each point is similar to or less than 1. The total χ^2 is 13.34. For a model which properly matches the data and uncertainties, we would typically find $\chi^2 = N - k$. The expected value of χ^2 is generally called the *number of degrees of freedom* in the fit. Each (x_i, y_i) pair provides a degree of freedom, while each fit parameter "consumes" one. For our lead thermal conductivity fit, we obtain $\chi^2/\text{ndof} = 3.34$ which suggests that the fit is not a great one, or that

Applied Pressure [kg/cm^2]	Thermal Conductivity Measured	Predicted	χ	χ^2
2000	1165.6 ± 3.1	1170.1	-1.45	2.09
4000	1205.0 ± 2.5	1206.3	-0.52	0.27
6000	1250.0 ± 2.5	1242.5	2.99	8.94
8000	1271.3 ± 6.3	1278.7	-1.18	1.40
10000	1312.5 ± 8.8	1315.0	-0.28	0.08
11900	1347.5 ± 2.5	1349.4	-0.75	0.57

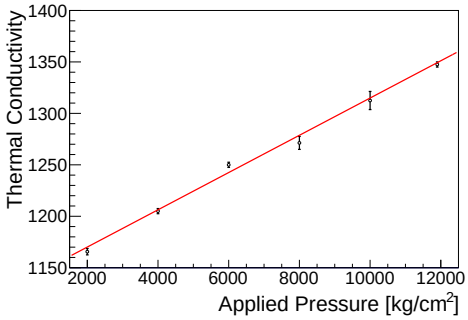


Figure 3: Results of the linear fit of thermal conductivity as a function of applied pressure.

the uncertainties on the measurements are underestimated. In contrast, if we observed $\chi^2/\text{ndof} < 0.1$, that would strongly suggest that the uncertainties are overestimated.

It is useful to examine the table of χ and χ^2 values to understand the data better. Looking more deeply, we observe that all of the χ values are less than zero except for a single point which has a χ much larger than zero. This is a symptom of a problematic fit. The χ values should be roughly symmetric around zero and have a typical magnitude near one.

The one point with a positive χ, at 6000 kg/cm^2, is an *outlier* with a $\chi^2 = 8.94$ compared to the next-largest value of 2.08. An outlier can be one of several things:

1. An outlier could be a data point which has an under-estimated uncertainty either due to an error on the experimenter's part or due to a difficulty with technique.

2. An outlier could be an invalid data point where the point, or its uncertainty estimate, were affected by an uncontrolled influence. In such a case, the outlying point should not necessarily be modeled by the same function as the other data points – the uncontrolled influence has changed the function for that point.

3. An outlier, particularly at the first or last data point, could be indicating that the model in use is not appropriate. For example, the true model could be an exponential rather than a linear function.

4. An outlier could be a part of the natural distribution of points, but an unusual one. The likelihood of this explanation decreases as χ^2 increases and is effectively ruled out for $\chi^2 > 10$.

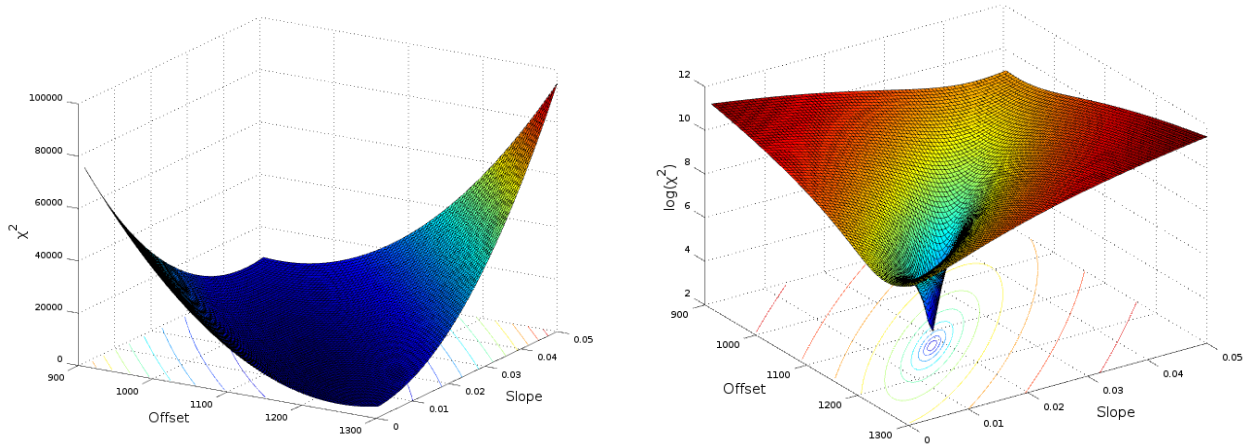
If an outlier appears to be an invalid data point, it may be removed from the analysis. In this case, a clear explanation must be put into the logbook and any paper indicating that some points were suppressed and what the justification was. This process should not be used to simply "prune" the data to decrease the χ^2 value – that could hide the fact that the real issue lies in the uncertainties.

If we prune the data point at 6000 kg/cm^2, we obtain new offset and slope values $a = 1130.6 \pm 2.8$ and $b = 0.01821 \pm 0.00035$ cm^2/kg. These are consistent with the previous values including all the data points, but the new $\chi^2 = 1.23$ and $\chi^2/\text{ndof} = 0.41$. In this case, the χ values are also much more symmetrically distributed around zero.

4 Non-linear fitting

4.1 Introduction to non-linear fitting

In Section 3.2, we discussed the case of linear fitting, where the parameters of the fit are all independent linear coefficients. However, sometimes we need to carry out a fit which is non-linear – where the parameters cannot be expressed as linear coefficients. In this case, we can still apply the χ^2 formalism, but we can't necessarily find a closed-form solution. Instead, it is typically necessary to carry out an iterative process to find a solution.



(a) (b)

Figure 4: Linear (a) and log (b) plots of the χ^2 surface for a linear fit to the data in section 3.2. The logarithmic plot highlights the best-fit region where the χ^2 reaches its minimum value. There is a significant correlation between the fit parameters, as clearly seen in the figure.

Consider a general problem where we have a set of observations $y_i \pm \sigma_i$ with control variables $\vec{x}_i$. We wish to find the best values of a set of parameters $(a_1, ..., a_n)$ for a function $f(\vec{x}; \vec{a})$. We can do this by minimizing the χ^2 with respect to the parameters.

$$\chi^2 = \sum_i \frac{[y_i - f(\vec{x}_i; \vec{a})]^2}{\sigma_i^2} \tag{10}$$

$$\frac{\partial \chi^2}{\partial a_j} = \sum_i -2 \frac{[y_i - f(\vec{x}_i; \vec{a})]}{\sigma_i^2} \frac{\partial f(\vec{x}_i; \vec{a})}{\partial a_j} \tag{11}$$

In the linear case, the partial derivatives $\frac{\partial f(\vec{x}_i;\vec{a})}{\partial a_j}$ are independent between the parameters (e.g. $\frac{\partial f(\vec{x}_i;\vec{a})}{\partial a_1}$ contains only a_1 and not $a_2, a_3, ...$ Thus, we can set the derivative to zero and obtain a set of linear equations to extract the best fit.

However, in the non-linear case, the partial derivative for a_1 will generically include $a_2, a_3, ...$ so the equations are generally not separable for solution. Instead, we will need to apply an iterative technique to find the best values, beginning with an initial guess which may be random.

To begin, it is useful to visualize the χ^2 surface for a fit. Figure 4 shows the χ^2 surface for the fit of the lead conductivity data to a linear model. An iterative fitting process will traverse the χ^2 surface and seek the minimum value of the χ^2. For a simple surface with a single minimum, many algorithms will work effectively. For a more-complex surface with multiple minima (as discussed below), algorithms may "get stuck" in non-optimal local minima and require additional input or help to achieve an effective fit.

In every case, once we find the minimum point we can also find the uncertainties on each of the parameters. The uncertainty on each parameter is defined as the change to the parameter required to increase the total χ^2 by 1.0 from the minimum value. For a well-behaved minimum, the χ^2 surface will be parabolic at its minimum, so the errors will be symmetric (have the same magnitude for positive and negative variations). However, the errors should be evaluated separately for the positive and negative variations, since this condition may not apply to the data or the fit functions in any particular situation.

4.2 Grid Search

The first technique which we'll consider is the grid search. In this technique, we start with an initial guess of the parameters and an initial step size δ_i for each of the parameters. As the parameters do not typically

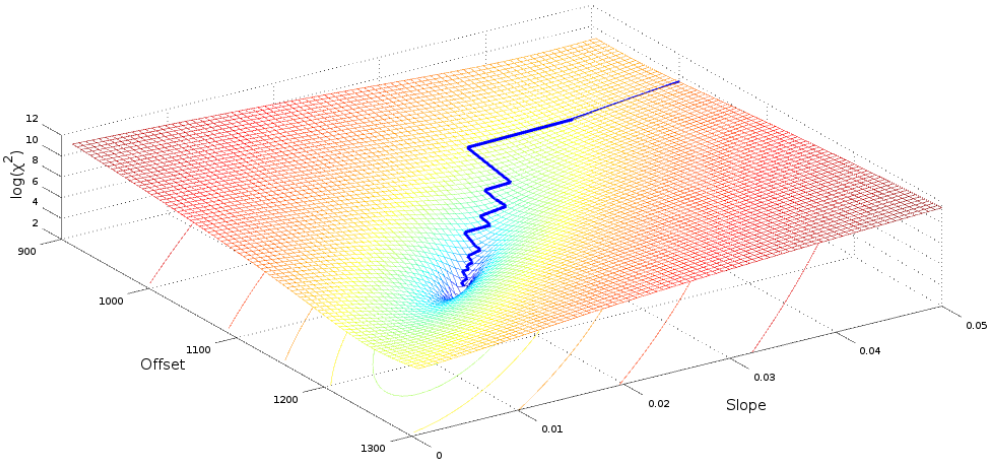


Figure 5: Grid search progress on the linear fit of the lead conductivity versus pressure dataset. Due to the correlation between the fit parameters, the grid search must proceed by small steps from one side of the valley to the other.

have the same range of reasonable values (and usually have different units), the initial step sizes for the different parameters are typically quite different.

The grid search does not require that we are able to calculate the closed-form derivatives of $f(\vec{x};\vec{a})$. Instead, the search proceeds by determining the χ^2 for each of the cases of $a_1 + \delta_1$, $a_1 - \delta_1$, $a_2 + \delta_2$, and so on. Each cases is evaluated independently and the shift which produces the biggest decrease in χ^2 is chosen. If all the shifts produce increases in the χ^2, it is likely that the algorithm is getting closer to the global minimum. In this case, the step size is typically reduced (divided by 2 for example) and a new set of more-closely-spaced points is checked. The algorithm can be typically terminated either after a defined number of steps or (better) when the χ^2 decrease becomes small (e.g. much less than 1.0).

The algorithm may also apply constraints to the parameters to keep them from moving outside a defined range – for example, a parameter may only make sense if it is positive or two parameters should not be too close to each other.

An example of the convergence of this algorithm for the lead conductivity dataset is shown in Fig. 5. The trace of the fit process shows how the fit converges by first taking a series of steps decreasing the slope parameter from the initial guess and then begins alternating steps for the slope and offset. Because of the correlated effects of the slope and offset in this fit, the minimum "valley" has a long and narrow aspect and the fit is forced to take many steps to "walk" down the valley. This characteristic is typical of many fitting problems and becomes more severe as the number of fit parameters increases.

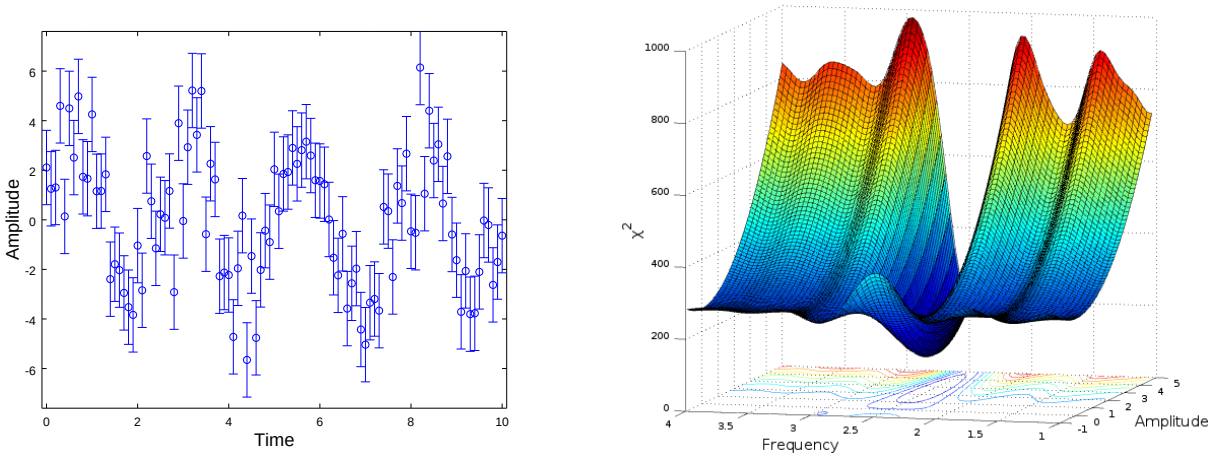
4.3 A more difficult example

The linear fit case discussed so far is something of a "fake" example, as we know how to solve this problem analytically. Consider the data shown in Fig. 6(a). The data has a sinusoidal form and we may attempt to fit it using the form

$$a_1 \sin(a_2 t)$$

If we calculate the χ^2 surface for such a fit, we obtain the surface shown in Fig. 6(b). This fit is much less well-behaved than the linear fit we considered earlier. Rather than the one unique minimum with a continous (if sometimes slow) decent possible towards that minimum, in this case we observe multiple minima. The best minimum is in fact "protected" by high ridges, which will tend to repel fitting attempts. Such a surface is typical when fitting a periodic function like the sine. Higher and lower integer multiples of the best fit frequency (harmonics) will appear as lesser valleys in χ^2 space.

If we attempt a grid search for such a fit, the results will depend strongly on the initial seeded parameters and the step sizes. Figure 7 demonstrates this effect. For some values of the initial parameters and step



(a) (b)

Figure 6: A periodic dataset (a) and the χ^2 surface associated with fitting it to $a_1 \sin a_2 t$.

sizes, the fit will end up trapped in a local minimum valley, as shown by the red path in the figure. For other initial conditions, the grid search will successfully find the global minimum and provide useful fit results.

This raises the question of how the scientist is to best determine these initial conditions for success. There are two basic principles to follow:

1. Use plots of the data to determine a rough starting point for each parameter. This is fairly straightforward in the case of the noisy sine data – the amplitude and frequency can be roughly determined from the plot and used to initialize the fit. The step size should indicate initial uncertainty on that rough determination of the parameters.

2. Repeat the fit multiple times using randomly-shifted parameters within a reasonable range of the initial guess. If every fit returns back to the same point, it is more-likely that it is the global minimum. While carrying out this process, keep track of the minimum χ^2 value observed.

5 Probability Density Functions

A proper probability density function has the characteristic that it is normalized so that

$$\int_{-\infty}^{\infty} P(x)\,dx = 1$$

The partial integral over a finite range is called the cumulative density function.

$$\mathrm{CDF}(x_1) = \int_{-\infty}^{x_1} P(x)\,dx$$

5.1 The Poisson Distribution

The Poisson distribution is applicable in situations where a process has a single well-defined average rate or number and where the time between occurrences or the number of occurrences in subsequent time intervals is uncorrelated. This situation is very common, but is particularly well-suited to radiative processes such as photons arriving from an incoherent source or radioactive decay processes. Other famous examples of Poisson distributions are the number of Prussian soldiers killed each year by horse kicks, the number of cars arriving at a traffic light in a fixed time interval, and the number of yeast cells used in brewing a batch of Guinness beer.

The Poisson distribution has only one parameter, the average rate μ.

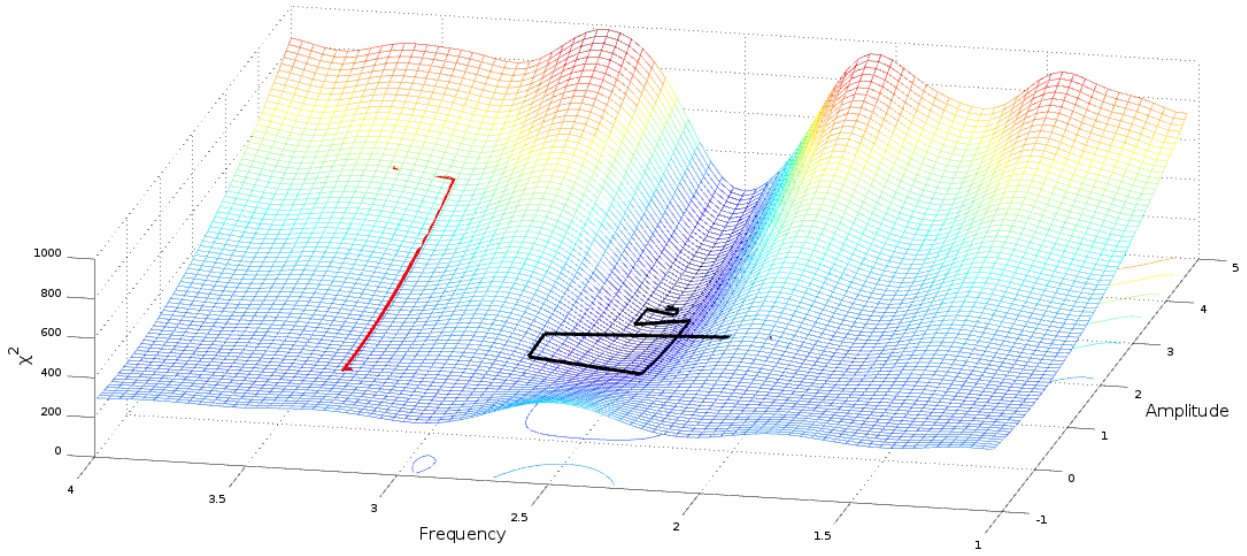


Figure 7: Two examples of the grid search process on the "noisy-sine" dataset. Depending on the initial values of the parameters and the step size, the fit may become trapped in a local minimum (red line) or find the global minimum (black line).

$$P(x;\mu) = \begin{cases} \frac{\mu^x e^{-\mu}}{x!} & x \in (0,1,2,3,...) \\ 0 & x < 0, \text{non-integer} \end{cases} \tag{12}$$

The average rate μ is a real parameter, but x can take only non-negative integer values. The variance of the Poisson distribution is also μ, so the standard deviation of a Poisson distribution is $\sqrt{\mu}$. If we plot the Poisson distribution for $\mu = 0.1, 1.0, 5.0, 10.0$ (Fig. 8), we observe that the distribution is highly asymmetric for small values of μ and becomes increasing symmetric as μ increases. For large values of μ, the Poisson distribution becomes nearly indistinguishable from the Gaussian distribution discussed below.

5.2 The Exponential Distribution

If the Poisson distribution is the correct distribution for counting the number of events in a fixed interval, the exponential distribution properly describes the time between individual events. The probability density function for the distribution is defined by a single parameter λ, the average rate of events.

$$P(t;\lambda) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{13}$$

The exponential distribution has a simply-expressed cumulative distribution function as well.

$$\text{CDF}(t_1;\lambda) = 1 - e^{-\lambda t_1} \tag{14}$$

The time t must be equal to or greater than zero. For the exponential distribution, we find that the standard deviation is equal to λ.

To directly connect the Poisson distribution to the exponential, we must include the time interval used for the Poisson measurement as the integration period τ. In this case, the rate parameter is simply

$$\lambda = \frac{\mu}{\tau}$$

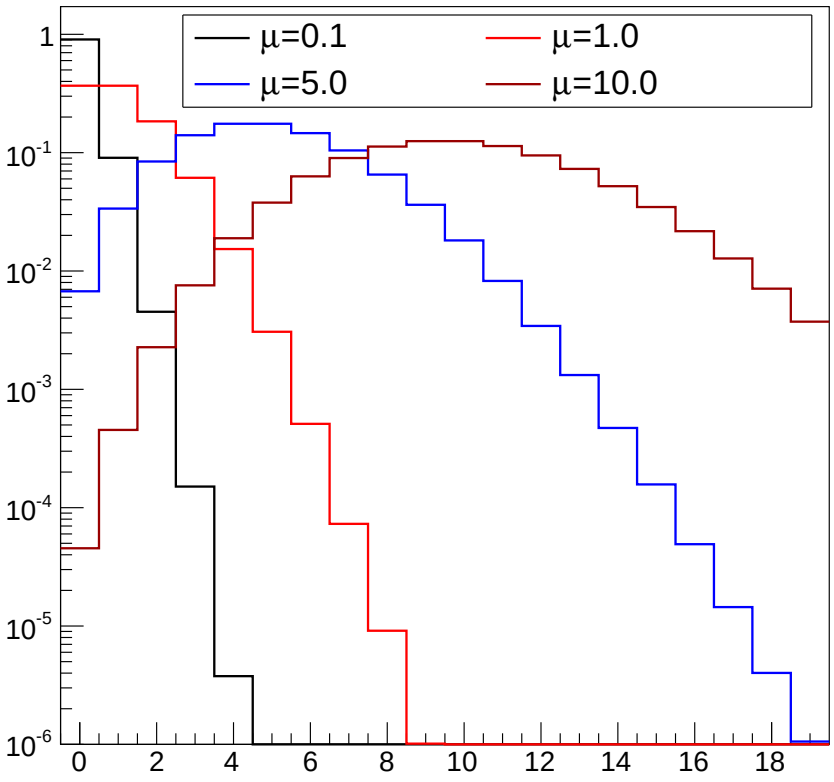


Figure 8: The Poisson distribution for $\mu = 0.1, 1.0, 5.0, 10.0$.

5.3 The Gaussian Distribution

The Gaussian distribution, sometimes called the "bell-curve" or "normal distribution", is a very common probability density function. In contrast to the Poisson and exponential distributions, the Gaussian has two parameters: μ and σ. The probability density function is

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2} \tag{15}$$

The probability density function is defined for all values of x, with no restrictions for the variable to be positive or integer. The cumulative distribution function does not have a simple closed form, but instead has a specific functional name: *the error function*.

The use of terms like "normal distribution" and "the error function" seems to indicate that the Gaussian is a particularly important distribution. This is because the Gaussian naturally appears in any situation where an observable is the sum of multiple uncorrelated random variables *independent of the PDF of the variables*. This behavior is known as the Central Limit Theorem. Strictly speaking, the variables must all have the same random distribution for the Central Limit Theorem to hold, but in practice this restriction is not important and the Gaussian appears in many situations.

This behavior is so strange that I encourage you to study it yourself using a Java simulation available here: `http://www.chem.uoa.gr/applets/AppletCentralLimit/Appl_CentralLimit2.html`

5.4 The Log-Normal Distribution

The Gaussian or normal distribution is the limiting distribution when considering uncorrelated *additive* effects. When the effects are *multiplicative*, then Gaussian distribution applies to the logarithm of the variable

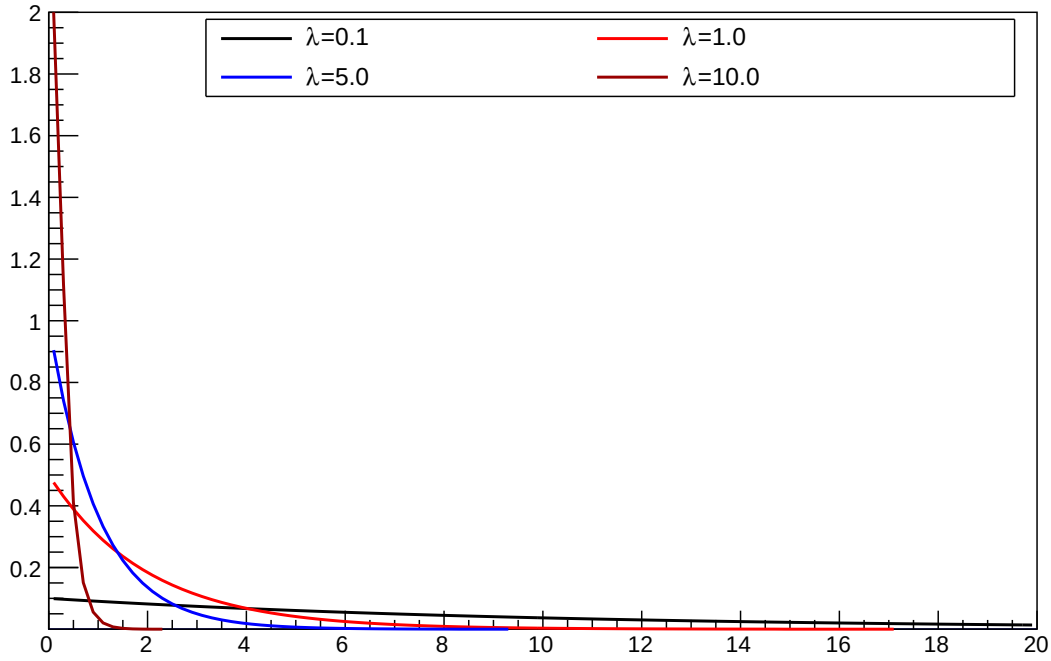


Figure 9: The exponential distribution for $\lambda = 0.1, 0.5, 1.0, 5.0$.

rather than to the variable itself. This modification of the Gaussian is call the "log-normal" distribution and the distribution function is given below.

$$P(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\left(\frac{\ln(x)-\mu}{\sqrt{2}\sigma}\right)^2} \tag{16}$$

The formula is nearly identical to that of the Gaussian, but the x is replaced by $\ln(x)$ and x appears in the prefactor. This means that the μ does not refer to the mean of the distribution, but rather the logarithm of the mean, and σ is similarly in logarithmic units. A μ of zero thus implies a median value of $e^0 = 1$. The log-normal distribution is shown in Fig. 11 for several different values of σ.

6 Hypothesis Testing

The preceding section has focused on the question of how to make a measurement of a quantity and properly determine the uncertainty on the quantity given the uncertainties of the measurements which went into the calculation. Sometimes, however, we want to use statistics to understand whether a given process is present in the data – for example, to discover a new physical phenomenon. In this case, we need to distinguish between two hypothesis:

1. The *null* hypothesis (H_0), which asserts that observation can be fully explained without any new phenomena. The null hypothesis does not imply that *nothing* is observed, but that nothing "new" is observed. Any deviation from the expected observation can be explained by natural statistical variation.

2. The *alternative* hypothesis (H_a), which asserts that a new phenomenon or effect is present along with the same phenomena which comprise the null hypothesis.

In particle physics, the null hypothesis might be that the observed data can be fully explained by existing known particles while the alternative hypothesis is that a specific new particle is present along with the known

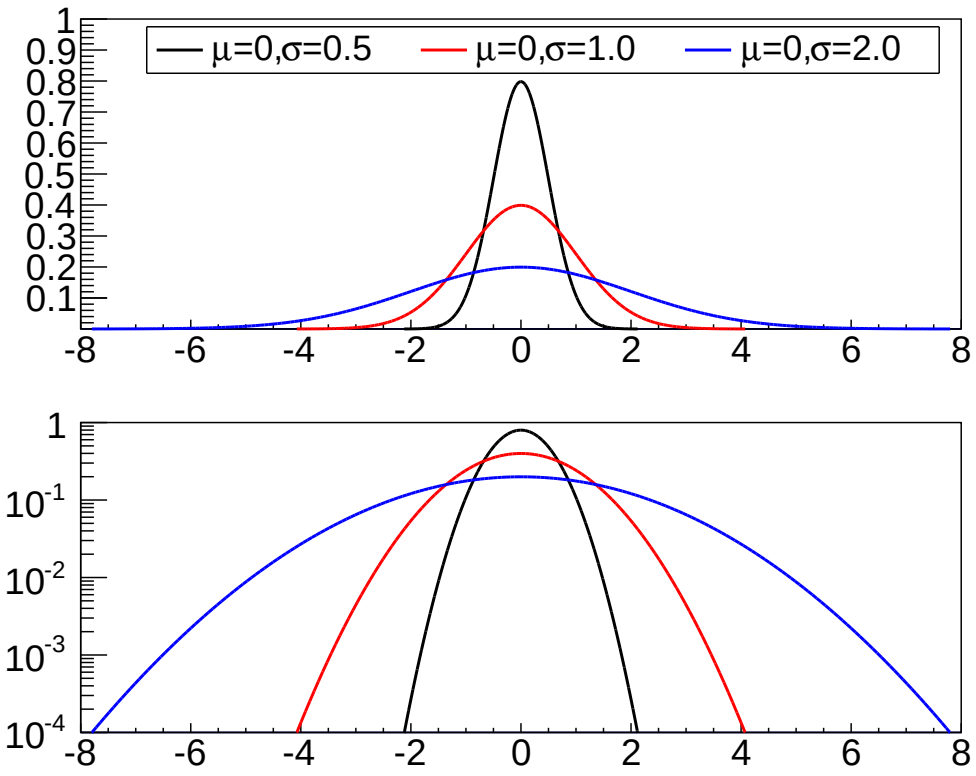


Figure 10: The Gaussian distribution in linear and logarithmic scales for $\mu = 0$ and $\sigma = 0.5, 1.0, 2.0$.

processes. In medicine, the null hypothesis might be that all variation in outcomes in a trial of a new medicine is due to the placebo effect, while the alternative hypothesis would be that the new medicine has a beneficial effect in addition to the placebo effect.

After identifying the two hypotheses, one must identify a test statistic to which the null hypothesis can be compared. In particle physics one can often compare the number of observed events to the integral Poisson distribution.

Let us consider the example of a particle physics experiment looking for neutrinos. If the expected number of events in the null (or background-only) hypothesis is μ_b and n_d events are observed in the data, the test statistic would be the probability of observing n_d events *or more* given the expectation of μ_b.

$$p(n_d; \mu_b) = \sum_{i=n_d}^{\infty} \frac{\mu_b^i e^{-\mu_b}}{i!}$$

Given the difficulty of computing the sum to infinity directly, one generally uses the fact that the Poisson is normalized to simplify the calculation to:

$$p(n_d; \mu_b) = 1 - \sum_{i=0}^{n_d-1} \frac{\mu_b^i e^{-\mu_b}}{i!}$$

The result of the computation is called a *p-value*. A small p-value for the null hypothesis indicates that that the null hypothesis is disfavored or excluded. Similarly, a small p-value for the alternative (signal-plus-background) hypothesis would disfavor or exclude the new process or phenomenon. Typical criteria for exclusion might be a confidence level of 99%, which implies a p-value of 0.01 or smaller.

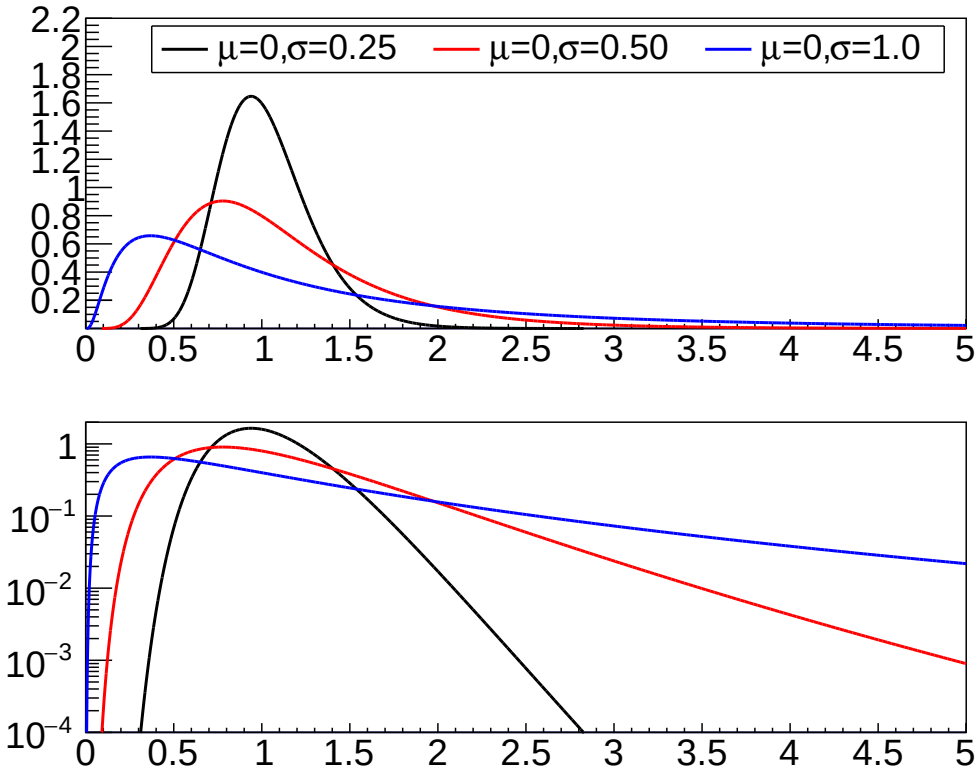


Figure 11: The log-normal distribution in linear and logarithmic scales for $\mu = 0$ and $\sigma = 0.25, 0.5, 1.0$.

7 Monte Carlo Techniques

Monte Carlo techniques are named after the famous casino of the same name in Monaco. These techniques are based on the use of random numbers to allow the estimation of integrals and the simulation of physical processes. These techniques can be used for data analysis as well as for determining uncertainty estimates to a greater precision than the Gaussian or normal error techniques discussed above.

7.1 Random Numbers

Monte Carlo techniques depend strongly on an unbiased source of random numbers. The algorithm used to produce random numbers is usually called a *pseudo-random number generator* (PRNG). Often the "pseudo" is dropped in conventional usage. Most commonly, random number generators produce floating point numbers in the range 0...1 and analytic algorithms are used to extend this range, as discussed below.

For an *unbiased* random number generator, subsequent random numbers produced by the algorithm do not have a significant pattern. This does not mean that the generator is not predictable. In fact, any useful generator must reproduce the same sequence of random numbers when initialized (or *seeded*) in the same way. The design of unbiased, computationally-efficient random number generators is an important topic in scientific computation. One of the most popular current algorithms is the " Mersenne twister", which is heavily used in many software packages [1]. There are many very bad random number generators available, however, including the defaults provided by the operating system in most cases. To see the visual impact of a bad random number generator compared with a good one, see Fig. 12.

Once armed with a generator of uniform random deviates (a random number generator which generates uniformly in the range 0...1), the next step is to consider random numbers following other distributions. There are two basic techniques which can be used. The first technique is to map the uniform deviate onto

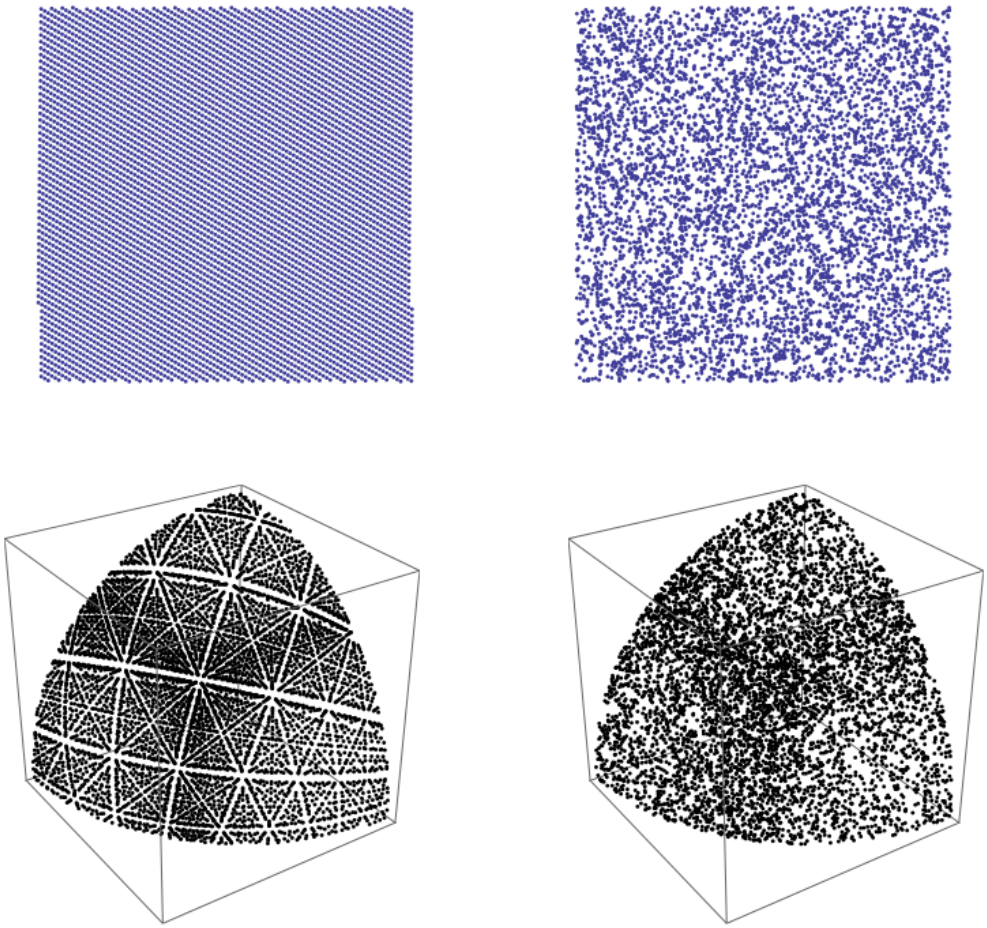


Figure 12: Correlations between subsequent random numbers as revealed in two dimensions and three dimensions for a poor home-brew random number generator (left) and the Mersenne twister (right) [1].

the function in question. This technique requires that the inverse of the integral of the function be known. For example, consider the probability density distribution of a parabola with a minimum at 0, defined between -1 and 1.

$$P(x) = \frac{3}{2}x^2 \; -1 \leq x < 1$$

The cumulative distribution function in this case is

$$C(x) = \frac{1}{2}x^3 + \frac{1}{2}$$

The uniform random number generator will provide values *of* the cumulative distribution function, so we must invert the function to determine the x value associated with a given C:

$$x = \sqrt[3]{2C - 1}$$

With this transformation, we can convert a sequence of uniformly-distributed random numbers into a parabolic distribution of random numbers. Since the argument of the cube root may be negative, we may need to take some care in a numerical program to avoid the results being interpreted as complex numbers – there should be a negative, real-valued solution in all cases.

The second technique is the accept/reject technique. It depends on the range of the function and the maximum amplitude to be known, but puts no requirements on the integral or the invertability of the integral.

In its simplest form, we may consider the case of a function $f(x)$ where the range is $[a, b)$ and the maximum value is C. We can generate random numbers in this range by repeatedly generating pairs of uniform random deviates u_1 and u_2. For each pair, we check if $f\left((b-a)\,u_1 + a\right) < C * u_2$. If it is, $(b-a)\,u_1 + a$ is taken as a random number from the distribution of $f(x)$, otherwise the process must be repeated with two additional random numbers. If $f(x)$ has a maximum value much larger than its typical value, this process will be very inefficient. In this case, it may be useful to find a bounding function $g(x)$ as described in [1].

References

[1] H. G. Katzgraber, "Random Numbers in Scientific Computing: An Introduction", *ArXiv e-prints* (May, 2010) `arXiv:1005.4117`.