Statistics Primer

Introduction

The measurement process involves taking some data and extracting from it some fundamental parameter or set of parameters. In a laboratory course, you will use various detectors to collect your data, and then you will want to "fit" the data according to the theoretical expectation of that experiment in order to measure the quantity of interest. For example, you may want to measure the lifetime of a radioactive isotope given a series of measured times recorded by a particle detector. This Primer concerns itself with what we mean by "fit," and how you extract the uncertainty in your measured parameter.

It is assumed that you already know basic statistical concepts such as the standard deviation of a data sample. Moreover, you are still expected to read the Statistical Analysis write-up accompanying the PHY4803L course, which will give a much more in-depth treatment of the topics and which contains the assigned homework problems.

Probability Distribution Functions

All data is distributed according to some parent distribution defined by the laws of Quantum Mechanics and by the properties of the detector. Each data point is considered to be the result of random process, and a *probability distribution function* describes how individual measurements will be distributed. If $p(x)$ is the probability distribution function, then $p(x)dx$ represents the probability of getting a data point between x and $x+dx$. For example, in a world with perfect detectors, this probability distribution function would be the square of a wavefunction. The probability distribution is expected to be normalized so that

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

Below we review the most widely used distribution functions.

Gaussian:

The Gaussian distribution, also known as the Normal distribution, is probably the most familiar. Mathematically it is given by:

$$p(x) \equiv G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\ \sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

The reason the Gaussian distribution is so important is that the distribution of most measurements are close to Gaussian in shape, and in fact if you average enough measurements the resulting distribution is always a Gaussian no matter what the initial distribution is. This is known as the Central Limit Theorem. The important parameters of the Gaussian are:

μ = mean

σ = standard deviation

The full-width at half-maximum (FWHM) of a Gaussian is related to the standard deviation by: FWHM = 2.35σ.

The probability of a measurement to lie within one standard deviation of the mean for a Gaussian can be computed:

$$P(\leq 1\sigma) = \int_{\mu-\sigma}^{\mu+\sigma} G(x;\mu,\sigma)\, dx = 2\int_0^1 \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

$$= 2\ \mathrm{erf}\left(\frac{1}{\sqrt{2}}\right) = 0.68$$

$$\text{where } \mathrm{erf}(y) = \frac{1}{\sqrt{\pi}} \int_0^y e^{-t^2}\, dt$$

Thus, there is a 68% probability of getting a measurement within one standard deviation of the mean. Likewise, we can compute the probability of a measurement to lie within two standard deviations of the mean:

$$P(\leq 2\sigma) = \int_{\mu-2\sigma}^{\mu+2\sigma} G(x;\mu,\sigma)\, dx = 2\int_0^2 \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = 0.95$$

The 68% probability is now taken as the convention for error bars when reporting measurements with uncertainties: $x \pm \sigma$. The σ is defined and interpreted such that there is a 68% probability of the true parameter lying within this range.

A plot of a Gaussian distribution is shown in Fig. 1.

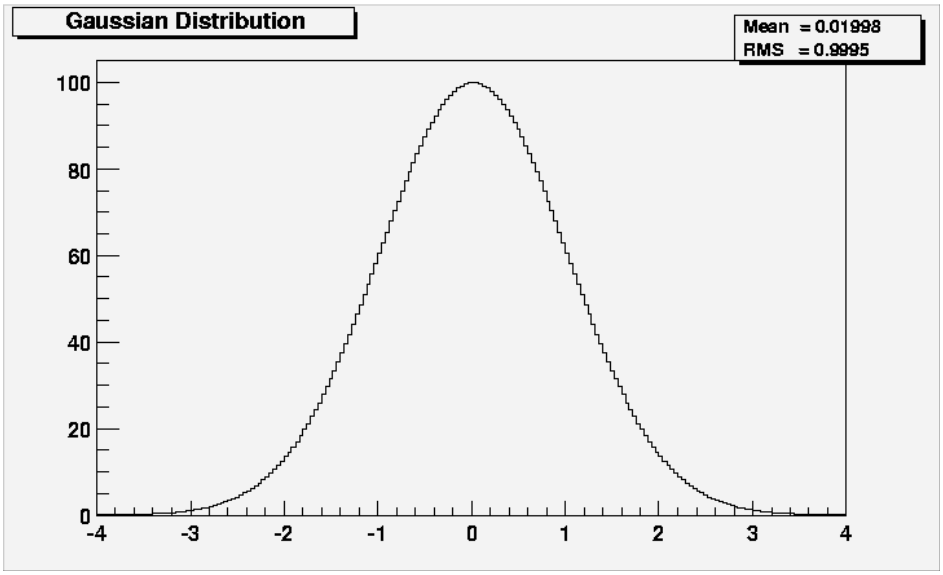

Figure 1: Histogram of a Gaussian distribution with $\mu = 0$ and $\sigma = 1$.

Poisson:

The Poisson distribution applies to counting experiments, where the measurement n is the number of occurrences of a certain event. For example, n might represent the number of γ-rays measured to have an energy between E and $E+\Delta E$. When such occurrences are truly random and do not depend on the history of the previous measurements, which is the case for radioactive decays, then the probability distribution for n follows the Poisson distribution:

$$P(n) = \frac{e^{-\mu}\mu^n}{n!}$$

One can explicitly check the normalization condition:

$$\sum_{n=0}^{\infty} P(n) = 1$$

For the Poisson distribution:

μ = mean

$\sigma = \sqrt{\mu}$ = standard deviation

Since the standard deviation is just the square root of the true mean, an estimate of the uncertainty in a measured number of events n is just taken to be $\sqrt{n}$.

For large enough μ (in practice only larger than about 5), the Poisson distribution approaches a Gaussian:

$$P(n) \to G\left(n;\ \mu, \sigma = \sqrt{\mu}\right)$$

A plot of a Poisson distribution is shown in Fig. 2.

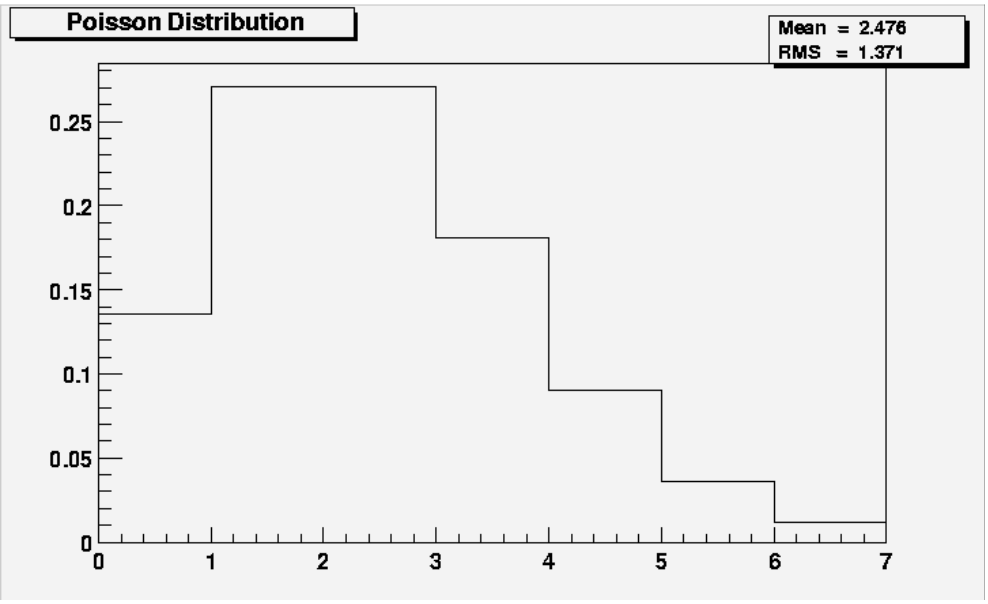

Figure 2: Histogram of a Poisson distribution with μ = 2. The mean is reported with an offset of 0.5 because the bin center, not the left edge, is taken by the plotting package.

Exponential:

The distribution of times between Poisson events follows the exponential distribution. For example, the distribution of times between nuclear disintegrations follows the Radioactive Decay Law, which is an exponential distribution:

$$p(t) = \frac{1}{\tau} e^{-t/\tau}$$

$\mu = \tau =$ mean lifetime

$\sigma = \sqrt{\tau} =$ standard deviation

A plot of an exponential distribution is shown in Fig. 3.

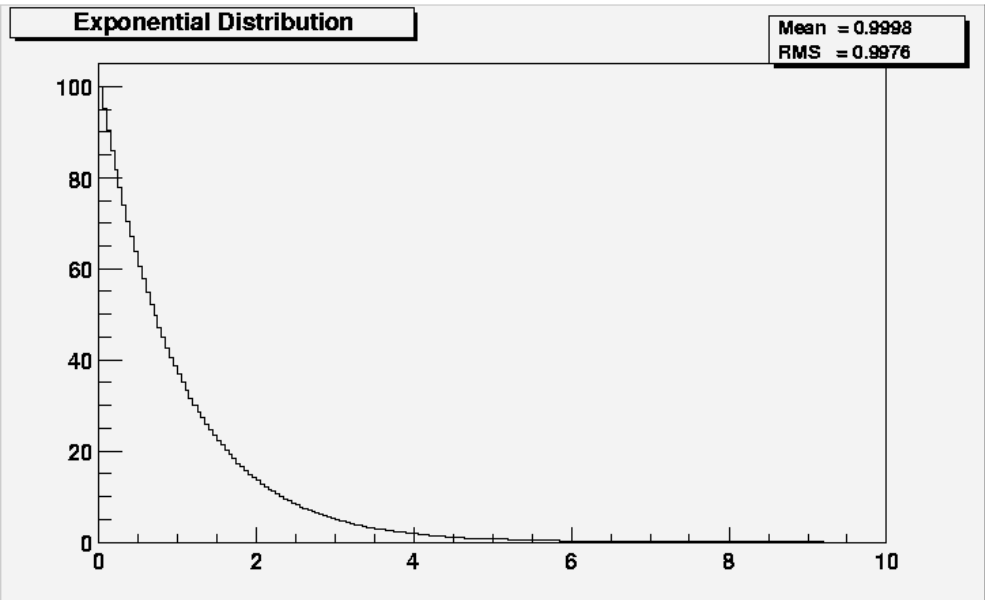

Figure 3: Histogram of an exponential distribution with $\tau = 1$.

Flat:

A special case not considered so far is the flat distribution, which just means that the data is distributed such that all values are equally likely between two limits. In particular:

$$p(x) = \text{constant} = \frac{1}{x_2 - x_1}; \quad x_1 \le x \le x_2$$

For a flat distribution, it is clear that the mean is just the center of the distribution. What is not so clear, but is left as an exercise for the reader, is that the standard deviation of a flat distribution is the total width divided by $\sqrt{12}$:

$$\mu = \frac{x_2 + x_1}{2}$$

$$\sigma = \frac{x_2 - x_1}{\sqrt{12}}$$

A plot of a flat distribution is shown in Fig. 4.

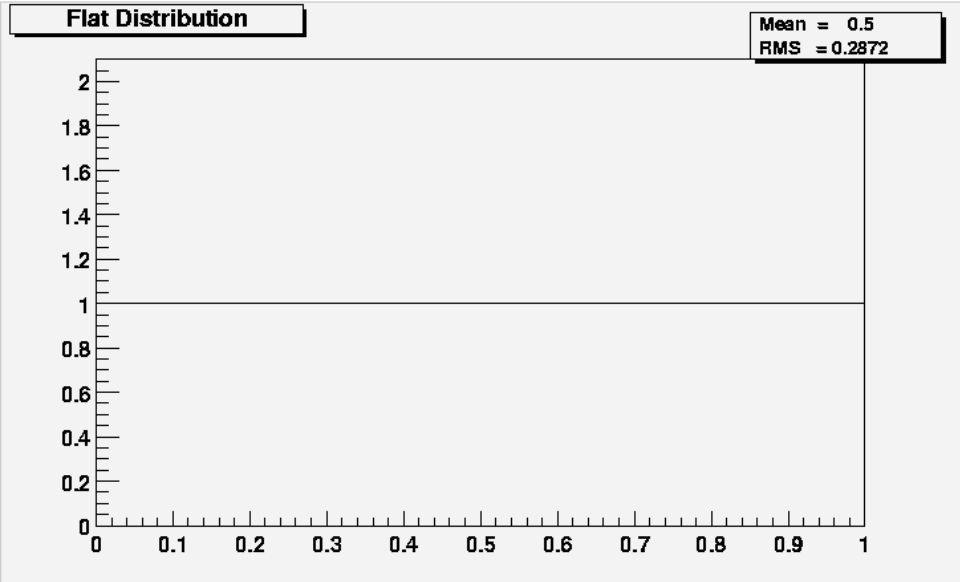

Figure 4: Histogram of a flat distribution between 0 and 1.

Constructing Estimators

Method of Maximum Likelihood

The probability of getting a measurement within the small interval x_1 to $x_1 + dx$ is $p(x_1)dx$. The probability of getting a certain set of N measurements is:

$$P = p(x_1)dx_1\ p(x_2)dx_2\ \cdots\ p(x_N)dx_N$$

Now if we assume a Gaussian probability distribution for each measurement, we have

$$P = \frac{dx_1}{\sqrt{2\pi}\sigma_1}e^{-(x_1-\mu_1)^2/2\sigma_1^2}\ \ \frac{dx_2}{\sqrt{2\pi}\sigma_2}e^{-(x_2-\mu_2)^2/2\sigma_2^2}\cdots\frac{dx_N}{\sqrt{2\pi}\sigma_N}e^{-(x_N-\mu_N)^2/2\sigma_N^2}$$

$$P = \frac{dx_1 dx_2\cdots dx_N}{\left(\sqrt{2\pi}\right)^N\sigma_1\cdots\sigma_N}\exp\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{\left(x_i-\mu_i\right)^2}{\sigma_i^2}\right]$$

Now what if we don't know μ_i and want to measure it? One technique to estimate it is to apply the *method of maximum likelihood*, which for Gaussian probability distributions amounts to maximizing the probability:

$$L = \frac{1}{\left(\sqrt{2\pi}\right)^N\sigma_1\cdots\sigma_N}\exp\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{\left(x_i-\mu_i\right)^2}{\sigma_i^2}\right]$$

This can be done by taking the partial derivatives: $\frac{\partial L}{\partial \mu_i} = 0$

Method of Chi-Square Minimization

For measurements whose parent probability distribution is Gaussian, the method of maximum likelihood is equivalent to minimizing the quantity:

$$\chi^2 \equiv \sum_{i=1}^{N}\frac{\left(x_i-\mu_i\right)^2}{\sigma_i^2}$$

Thus, the *method of chi-square minimization* involves solving: $\frac{d\chi^2}{d\mu_i} = 0.$

For example, suppose that all measurements are expected to be distributed with a common μ and σ (*i.e. N* measurements of the same quantity). Then the best estimate for μ can be found as follows:

$$\frac{d\chi^2}{d\mu_i} = 0 = \frac{1}{\sigma^2}\sum_{i=1}^{N} -2\left(x_i - \mu\right)$$

$$\Rightarrow 0 = \sum_{i=1}^{N}\left(x_i - \mu\right)$$

$$= \sum_{i=1}^{N} x_i - N\mu$$

$$\Rightarrow \mu_{\text{est}} = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{simple average}$$

Thus, the best estimate for μ is just the average of the sample. Now suppose that each σ_i is different. Then

$$\frac{d\chi^2}{d\mu_i} = 0 = \sum_{i=1}^{N} -\frac{2}{\sigma_i^2}\left(x_i - \mu\right)$$

$$\Rightarrow 0 = \left(\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\right) - \mu\left(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)$$

$$\Rightarrow \mu_{\text{est}} = \overline{x} = \frac{\left(\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\right)}{\left(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)} \qquad \text{weighted average}$$

Thus, the best estimate for μ is the weighted sample average.

The uncertainty on these estimates for the mean can also be derived, and can be shown to be:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}} \qquad \text{for simple average}$$

$$\sigma_{\overline{x}} = \sqrt{\frac{1}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}} \qquad \text{for weighted average}$$

Fitting

Suppose that each μ_i is different, and in fact can be expressed as a functional dependence: $\mu = F(x)$. Then the constructed chi-square is:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left[y_i - F(x_i)\right]^2}{\sigma_i^2}$$

Here our measurements y_i have an uncertainty σ_i and depend on the quantity x_i. The measurements y_i could be the amplitude measurements of a sinusoidal trace on an oscilloscope at various times x_i, for example. The expected functional dependence for the measurements is $F(x)$, and we wish to "fit" this function to our data in order to obtain some parameters.

The simplest function is the linear case: $F(x) = a + bx$. The best estimate for *a* and *b* can be solved analytically by taking partial derivatives of χ^2 with respect to these parameters. When the uncertainties σ_i are all the same (or unknown) the solution is independent of their values. This technique of determining linear coefficients is also known as *Linear Regression,* and packages like Microsoft Excel can solve for the linear parameters.

When the fitting function is not linear (such as $F(x) = Ae^{-x} + B$), one cannot solve for the parameters analytically. Instead, χ^2 must be minimized numerically to solve for the best estimates of the parameters. Programs like Microsoft Excel have "Solver" tools to minimize an expression by varying a set of parameters.

In order to ascertain the "goodness of fit", you should determine the probability of obtaining the value of χ^2 you found at the minimum. There are tables that catalog the Chi-square probability as a function of the number of degrees of freedom (number of data points minus the number of fit parameters). A good rule of thumb, however, is that the reduced Chi-square (χ^2/ndof, where "ndof" is the number of degrees of freedom) should be about 1. This means that on average, each data point is approximately one standard deviation away from the fitted function. In fact, if your errors are defined with the 68% probability definition, about 2/3 of your data points should lie within 1σ of the fitted curve, and 1/3 will lie outside. If your reduced Chi-square comes out much larger than 1, then the fitted hypothesis is not really describing your data, at least according to the errors you assigned to your data. If the reduced Chi-square comes out much smaller than 1, then the uncertainties assigned to your data are much larger than the actual spread of the data about the fitted curve.

Some packages, like the linear regression tool in Microsoft Excel, ignore any error bars and just fit linear functions to your data (x_i, y_i) by minimizing the quantity $\sum_{i=1}^{N}\left[y_i - F(x_i)\right]^2$. Then a "standard error" will be reported that represents the spread σ in your data necessary so that the reduced Chi-square comes out to be exactly 1. Clearly if not all measurements are equal, and some have larger errors than others, one should construct and minimize χ^2 directly to avoid biasing your fit.

Once you have found the best estimate of your fitted parameters by minimizing χ^2, you should also report the uncertainties in these quantities. This can be a complicated procedure. The recipe for the finding the errors on your fit parameters is the following:

1. Construct and minimize χ^2
2. Record your parameters $a_{\text{min}}, b_{\text{min}}, \ldots$ at χ^2_{min}
3. Fix one parameter at what you guess to be 1σ away from its minimum
4. Minimize χ^2 again by varying all parameters *except* the fixed one
5. Repeat steps 3 and 4 until the following condition is reached: $\chi^2_{\text{new}} = \chi^2_{\text{min}} + 1$

The difference in the value of the fixed parameter from its value at the minimum is taken to be its uncertainty. You should do this separately for positive and negative uncertainties for nonlinear fits because the error bars may be asymmetric. Of course the errors you obtain on your fitted parameters only make sense if the reduced Chi-square is about 1.

This iterative approach can be somewhat tedious in programs like Microsoft Excel. When many fits need to be done, one may wish to consider a statistics package that has all this machinery built in, such as the "Physics Analysis Workstation" (PAW) package in high-energy physics, or the commercial programs Origin and SigmaPlot.

Low Statistics Fitting to Counting Experiments

In rare counting experiments, such as the muon lifetime measurement, we sometimes have to work with low statistics in order to extract the quantity of interest. In such cases, we should be careful to avoid potential biases in our fitting technique.

For example, in a counting experiment, we may wish to fit an exponential curve to data binned in time in order to extract the radioactive lifetime of the process following the Chi-Square minimization procedure described above. Since the number of entries in each bin follows a Poisson distribution (the number of entries must be an integer, after all), we typically take the square root of the number of entries as an estimate of the uncertainty, σ_i, used in the χ^2 calculation. But what do we do if the number of entries is zero? In that case, our recipe breaks down. Even for bins with just a couple entries, our recipe gives too much weight to bins with low statistics that will bias our measurement.

One way to reduce the bias is to realize that for a Poisson distribution, it is the square root of the true mean, not of the observed number of events *N*, that represents the standard deviation. But how do we determine the true mean? The final result of our fit probably represents the best estimate, since it smoothes out the statistical fluctuations. So in that case, a better approach to fitting data with a Poisson distribution would be to take the square root of the fitted function for bin *i* as the uncertainty σ_i for that bin when computing the Chi-Square:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left[y_i - F(x_i) \right]^2}{F(x_i)}$$

This minor change removes the difficulty with empty bins, but problems are still encountered if during the iterative minimization process the fit becomes zero or negative. Also, since the Chi-Square minimization technique was derived assuming that the data is described by a Gaussian approximation (which is true for Poisson distributions only for large *N*), biases will still exist for low statistics.

The best approach to handling low statistics is to return to the method of maximum likelihood, using the Poisson distribution to describe the probability of each bin:

$$L_{\text{Poisson}} = \prod_i e^{-\mu_i} \frac{\mu_i^{n_i}}{n_i!}$$

This can be compared to the likelihood function constructed from Gaussian distributions:

$$L_{\text{Gauss}} = \frac{1}{\left(\sqrt{2\pi}\right)^N \sigma_1 \cdots \sigma_N} \exp\left[-\frac{1}{2} \sum_{i=1}^{N} \frac{\left(x_i - \mu_i\right)^2}{\sigma_i^2} \right]$$

Note that the Chi-Square distribution can be written as:

$$\chi^2_{\text{Gauss}} = -2 \ln L_{\text{Gauss}} + C$$

where C is a constant. We can define an effective Chi-Square from the Poisson likelihood in a similar manner:

$$\begin{aligned} \chi^2_{\text{Poisson}} &\equiv -2 \ln L_{\text{Poisson}} \\ &= -2 \sum_i \left(n_i \ln \mu_i - \mu_i - \ln n_i! \right) \end{aligned}$$

If we change our notation such that x_i labels the bin along the *x* axis, y_i labels the bin contents, and $\mu_i = F(x_i)$ is the fitted function, then:

$$\chi^2_{\text{Poisson}} = -2 \sum_{i=1}^{N} \left[y_i \ln F(x_i) - F(x_i) - \ln y_i! \right]$$

Thus, maximizing the likelihood function composed of Poisson distributions is equivalent to minimizing the Chi-Square function constructed above. However, we can simplify the expression even more, since only *F(x)* will change during the minimization procedure and not the data points. So we can drop the last term in the sum:

$$\chi^2_{\text{Poisson}} = -2\sum_{i=1}^{N}\left[y_i \ln F(x_i) - F(x_i)\right]$$

Since we constructed this effective Poisson Chi-Square in analogy to the Chi-Square function for Gaussian distributions, we can determine the uncertainty on a fitted parameter in the exact same manner as described in the previous section, where we let $\chi^2 \to \chi^2 + 1$ while holding the parameter fixed. The only thing we cannot do is interpret the goodness of fit from value of the Chi-Square because we ignored some constants in the translation from the likelihood function to the Chi-Square function for Gaussian variables that we did not do in the translation for Poisson variables.

An example of the possible reduction in the bias of the best fit to a sample of low statistics data is shown in Fig. 5, which is a histogram of time coincidence data taken from a 2-day run of a muon lifetime experiment. Two fits were performed to the hypothesis that the data is described by a radioactive decay law plus a background: $F(x) = A\exp(-t/\tau) + B$. One fit uses the Gaussian Chi-Square method, and the other uses the Poisson likelihood fit. For this 2-day run, both fits give a lifetime of $\tau = 1.95 \pm 0.1$ μs, which compares well to the known lifetime of 2.2 μs. But with just half of the data (where the number of bins with 0 or 1 entries increases), the Gaussian method shows a bias toward lower lifetimes (typically 1.7 μs), whereas the Poisson method is unaffected except for a larger statistical error. Also, notice in the figure that the two methods disagree on the size of the background even for 2 days of running: 2 entries per bin for the Gaussian method, and 2.9 entries per bin for the Poisson method. The actual background can be estimated by just taking the average of the bin contents for time coincidences larger than 10 μs, where the contribution from the exponential is negligible. The result is 2.9 entries per bin, in agreement with the Poisson method.

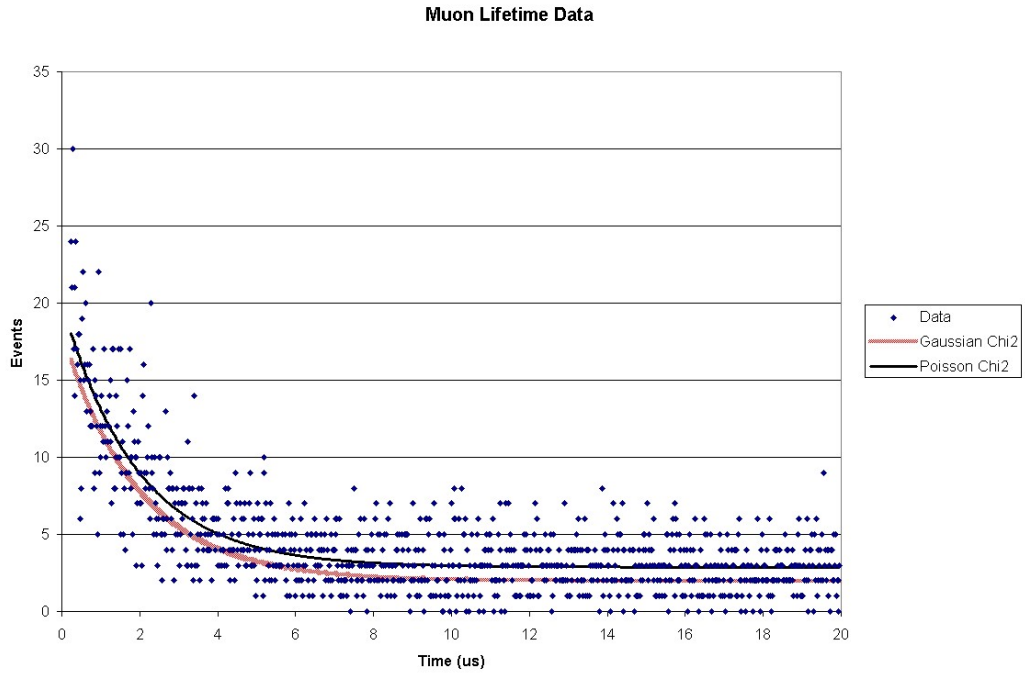

Figure 5: Histogram of time coincidence data from a muon lifetime experiment. A Gaussian Chi-Square fit and a Poisson likelihood fit are shown.

Propagation of Errors

Another major topic is the propagation of statistical uncertainties. Once you have found a particular parameter, you may wish to perform a mathematical transformation on it and its associated error. For example, suppose you wish to fit an exponential distribution $y(x) = ae^{-x/\tau}$ to your radioactive decay data (x_i, y_i), but have only a linear regression package. You could transform your data by taking the natural logarithm so that your fitting function becomes $y'(x) = \ln y(x) = -x/\tau + \ln A = a'x + b'$. But how do the uncertainties on y_i transform to uncertainties on y_i' ?

We apply the principles of calculus by assuming that the errors on a quantity are nearly infinitesimal ($\sigma_x \to dx$). Then if we transform our data with a function $f(x)$, the transformation of an infinitesimal length becomes:

$$df = \left|\frac{df}{dx}\right| dx \quad \Rightarrow \quad \sigma_f = \left|\frac{df}{dx}\right| \sigma_x$$

In other words, we are assuming that the transforming function is approximately linear over a region σ_x about *x*, and we are using the slope of the curve to determine how to transform the interval σ_x.

If our transformation involves several variables, and these variables are all independent and uncorrelated, then the error propagation formula becomes:

$$\sigma_f^2 = \left|\frac{\partial f}{\partial x}\right|^2 \sigma_x^2 + \left|\frac{\partial f}{\partial y}\right|^2 \sigma_y^2 + \left|\frac{\partial f}{\partial z}\right|^2 \sigma_z^2 + \cdots$$

We can apply this formula to several special cases. Suppose we are adding *N* measurements:

$$s = x_1 + x_2 + \cdots + x_N$$
$$\sigma_s^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 + \cdots + \sigma_{x_N}^2$$

In other words, we say that the errors "add in quadrature." If we were instead finding the average of *N* measurements where each measurement has the same uncertainty σ, it is easy to show that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \qquad \text{error on mean}$$

Now suppose we wish to find how the error propagates for measurements multiplied together. For example:

$$f = \frac{x}{y} = xy^{-1}$$

$$\frac{\partial f}{\partial x} = \frac{1}{y} \qquad \frac{\partial f}{\partial y} = -\frac{x}{y^2}$$

$$\sigma_f^2 = \frac{1}{y^2}\sigma_x^2 + \frac{x^2}{y^4}\sigma_y^2$$

But note that if we divide by the function f itself, we find that:

$$\left(\frac{\sigma_f}{f}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$$

In other words, it's the *relative* errors that add in quadrature for multiplication, not the absolute errors. If the errors on both x and y are 10%, then the error on f will be 14%.