Fermi National Accelerator Laboratory

FERMILAB-Conf-94/358

Statistical Data Analysis

A.A. Hahn

Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

November 1994

Presented at the *Beam Instrumentation Workshop 94*, Vancouver, B.C., Canada, October 3-6, 1994

Operated by Universities Research Association Inc. under Contract No. DE-AC02-76CHO3000 with the United States Department of Energy

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Statistical Data Analysis

A.A. Hahn
Fermi National Accelerator Laboratory, Batavia, IL 60510 *

Abstract

The complexity of instrumentation sometimes requires data analysis to be done before the result is presented to the control room. This tutorial reviews some of the theoretical assumptions underlying the more popular forms of data analysis and presents simple examples to illuminate the advantages and hazards of different techniques.

INTRODUCTION

The function of this tutorial is to reintroduce the concepts of statistical data analysis to Instrumentation Engineers. I assume that everyone has already had an introduction sometime earlier in their course work. I hope to emphasize some practical considerations along with the theoretical underpinnings of the analysis. Another motivation is prompted by the availability of software packages which contain quite powerful statistical packages.

The primary reference is a graduate student text by Bevington (1). Before we get too deep into the subject, it is useful to remind ourselves why this is an important topic. All the examples given below are taken from work which has been done in the Accelerator Division Instrumentation Department at Fermilab.

Data Reduction and Precision Measurements

Several instruments produce from one to hundreds of kilobytes of raw data per measurement cycle. Some examples at Fermilab are the Synchrotron Light Detector which images the bunch by bunch transverse beam shape and is read out by a video camera, the Sample Bunch Display (SBD) which measures the individual bunch intensity and length in the Main Ring and Tevatron (2), the Booster Ion Profile Monitor which measures the transverse profile of the Booster beam (3), the Collision Point BPM system which measures the space-time collision point of the proton and antiproton beams at the two colliding points in the Tevatron, and the Flying Wire System which measures the bunch by bunch transverse beam profile in the Main Ring and Tevatron. What the Main Control Room wants from our systems is at most a few tens of words of summarized beam information. It is up to analysis to provide accurate and timely information.

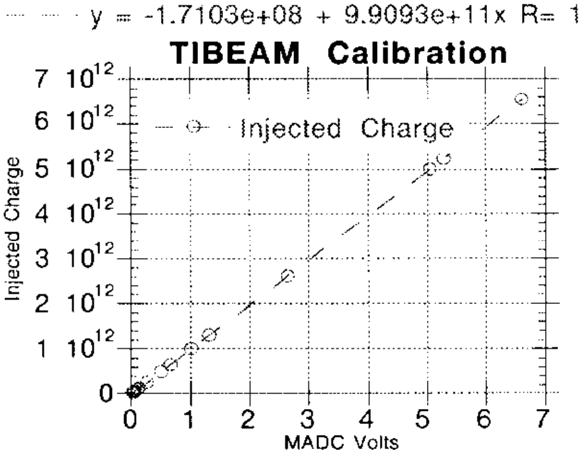
Several of the aforementioned systems are also being asked to provide automated measurements of the beam to the several percent level. It has been necessary to use sophisticated active noise subtraction techniques.

*Operated by the Universities Research Association under contract with the United States Department of Energy.

Calibration of Instruments

Figure 1 shows the calibration of the Tevatron DCCT Toroid (The plotted y value is the number which appears in the Main Control Room. The advantage of a fit over simply drawing a straight line is that given the data, anyone can get the same answer.

Figure 1. Calibration of Tevatron DCCT (T:IBEAM)



Parameterization of complex data by simple functions

Several times I have parameterized empirical curves by a relatively straightforward function. An example was the shape of the end field of a Tevatron dipole magnet. I needed a functional form which could be plugged into a formula to calculate the synchrotron photon yield. An error function shape fitted to the data was accurate enough for early prototyping.

SIMPLE STATISTICS-A REFRESHER

A fundamental requirement when a measurement is made is that a true value exists. This true value is often called the mean. If we make repeated measurements, we will find a distribution of the measurements about this mean. The width of the distribution tells us how precise our measurement is. A narrow distribution gives us confidence that we can determine the mean value well, while a wide distribution causes us worry. For an underlying probability distribution (either discrete, P_i, or continous, *P(x)*) we can define the mean μ and the width σ in the following manner:

$$\mu = \sum_{i=1}^{\infty} iP_i \xrightarrow{continuous} \int_{-\infty}^{\infty} xP(x)dx$$

$$\sigma = \sqrt{\sum_{i=1}^{\infty} (i-\mu)^2 P_i} \xrightarrow{continuous} \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 P(x)dx}$$

$$= \sqrt{(\sum_{i=1}^{\infty} i^2 P_i) - \mu^2} \xrightarrow{continuous} \sqrt{(\int_{-\infty}^{\infty} x^2 P(x)dx) - \mu^2}$$

I will sometimes refer to σ as the rms (root-mean-square) width. σ^2 is also known as the variance. The distributions are normalized, i.e.

$$1 = \sum_{i=1}^{\infty} P_i \xrightarrow{continuous} \int_{-\infty}^{\infty} P(x)dx \ .$$

Sample mean and variance

If we make n measurements of data, we can define a sample mean and width by:

$$\bar{x} = \left(\sum_{i=1}^{n} x_i / n \right)$$

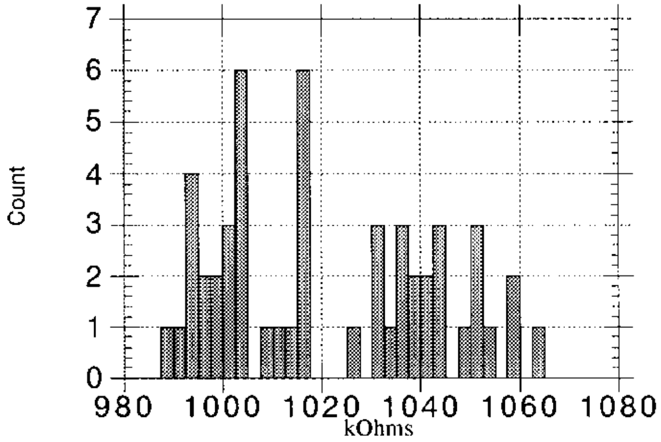
$$s = \sqrt{\overline{(x-\mu)^2}} = \sqrt{\overline{x^2} - \mu^2} \xrightarrow{\mu \to \bar{x}} \sqrt{(\overline{x^2} - \bar{x}^2)\left(\frac{n}{n-1}\right)}$$

The last step was a consequence of replacing the population mean by the sample mean. One degree of freedom in the data has been removed since the data set has already been used once in calculating the sample mean. In the limit of an infinite number of measurements, $\mu = \lim_{n\to\infty} \bar{x}$ and $\sigma = \lim_{n\to\infty} s$.

For example if we measure the resistance of a collection of 5% 1 MΩ resistors, we expect the mean to be 1 MΩ. (Is this really true?) We also expect to find a distribution of the individual resistors about the mean value. The exact width depends upon what the 5% specification really means. Figure 2 shows an actual distribution of 51 measurements of 1 MΩ resistors. This example illustrates the difference between experimental precision and measurement accuracy. The experimental precision of measuring was better than the resistor width distribution. This was confirmed by measuring the same resistor many times and noting the fluctuations were much smaller than 1%. But what of the absolute accuracy of the measurement? A summary of the data analysis is:

Mean	1021 kΩ
Std Deviation	22 kΩ
Std Deviation of Mean	3.1 kΩ.

Figure 2. Histogram of measurements of 1MΩ 5% resistors.



Statistical fluctuations due to the underlying processes (radioactive decay, industrial production methods, measurement techniques) are easily handled by traditional statistical methods. Systematic or calibration errors are much more difficult to deal with. For example, the resistance measurements gave reproducible readings at the 2 kΩ (0.2%) level. If we make ten measurements of the same resistor ,we can claim that our measurement is good (for that resistor) to 0.7 kΩ (we will show this in a following section). However the Ohmmeter may only be absolutely calibrated to the 1% level. What can we do to get a handle on systematic errors? Obviously we could buy a more accurate Ohmmeter or repeat the measurements with other brand Ohmmeters. (Why not use five of the same brand?). Changing measuring devices or techniques is equivalent to converting the systematic error into a statistical error.

Common Probability Distributions

Binomial

$$P(m,p,N) = \frac{N!p^m(1-p)^{N-m}}{m!(N-m)!}$$

The binomial distribution gives the probability for m successes out of N independent trials with the probability of a single success being p. It is known as a discrete distribution since the observables (m) are integers. The mean value and σ are $\mu = Np$ and $\sigma = \sqrt{Np(1-p)}$. An example of this distribution would be the number of times (m) one would expect to roll doubles on a pair of dice for N throws. The probability (p) per throw = 6/36 = 1/6.

Poisson

$$\mathrm{P}(\mathrm{m},\mu)=\frac{\mu^{\mathrm{m}}\mathrm{e}^{-\mu}}{\mathrm{m}!}$$

The Poisson distribution can be derived from the binomial distribution in the limit that p->0, and N-> infinity in a manner that their product (the mean) Np -> μ, a finite number. The observables "m" are integer and >0, although μ can be any positive real number. A unique and dearly beloved feature of this distribution is $\sigma=\sqrt{\mu}$. The Poisson distribution occurs in the counting statistics of radioactive decay.

Gaussian or Normal

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

Another limiting form of the binomial distribution when N >>1 is the Gaussian Distribution. The Gaussian distribution is a continuous distribution - x can vary continuously over the entire real axis. The Gaussian distribution is characterized by its mean μ and width σ. This distribution occurs everwhere!

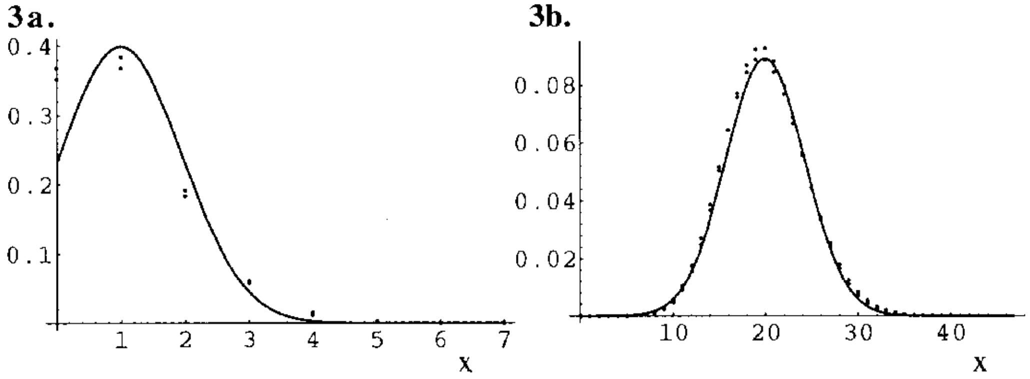
Uniform

$$\begin{aligned}P(x,\mu,w)&=\frac{1}{w}\ \text{if}\ |\mathrm{x}-\mu|\le\frac{w}{2}\\&=0\ \text{if}\ |\mathrm{x}-\mu|\le\frac{w}{2}\end{aligned}$$

The ability of calculators and computers to generate (pseudo)random numbers has made this distribution easily accessible to most people. This is a continuous distribution which has equal probability anywhere within the total window width "w" and identically zero probability outside. It can easily be shown to have a $\sigma=\sqrt{w^2/12}$. It serves nicely as a poor man's Gaussian if this value of σ is used.

The Binomial (12 trials, p=1/12), Poisson μ =1, and Gaussian (μ =1, σ = 1) distributions are plotted in fig. 3a. The Gaussian is plotted only for positive x values. Figure 3b shows the case for the binomial (N=240, p=1/12), Poisson (μ =20), and Gaussian (μ =1, σ = 1). The similarity of the distributions for these varied conditions shows why the Gaussian is used so often.

Figure 3(a,b). Plots of Binomial, Poisson , and Gaussian distributions for μ = 1 and μ = 20 respectively. The curve is the Gaussian Distribution. The plots illustrate the similarities of the three distributions.



PROPAGATION OF ERRORS

Analytic approach

If a quantity "y", itself is a function of other variables which have errors, how can we determine σ_y? First one takes the derivative of the function with respect to the independent variables. Then the sum is squared and the cross terms dropped since they average to zero for independent measurements.

$$\sigma_R^2 = \left\langle \left(\sum_i \frac{\partial R}{\partial r_i} dr_i \right)^2 \right\rangle \Rightarrow \sum_i \left\langle \left(\frac{\partial R}{\partial r_i} d_{r_i} \right)^2 \right\rangle = \sum_i \left(\frac{\partial R}{\partial r_i} \sigma_{r_i} \right)^2$$

Some common functions are:

$$R = r_1 + r_2, \qquad \sigma_R = \sqrt{\sigma_{r_1}^2 + \sigma_{r_2}^2},$$

$$R = r_1 * r_2, \qquad \sigma_R / R = \sqrt{\left(\sigma_{r_1}/\mathrm{r}_1\right)^2 + \left(\sigma_{r_2}/\mathrm{r}_2\right)^2},$$

$$R = \frac{r_1}{r_2}, \qquad \sigma_R / R = \sqrt{\left(\sigma_{r_1}/\mathrm{r}_1\right)^2 + \left(\sigma_{r_2}/\mathrm{r}_2\right)^2},$$

$$R = r_1 r_2^2, \qquad \sigma_R / R = \sqrt{\left(\sigma_{r_1}/\mathrm{r}_1\right)^2 + \left(2\sigma_{r_2}/\mathrm{r}_2\right)^2}, \text{ and}$$

$$R = \cos(r), \quad {}^{\sigma_R}\!/\!{}_R = |\tan(r)|\sigma_r.$$

Example: Error in the mean from n measurements

We can calculate the error in the mean on n measurements by noting we can use the propagation equation for a sum of terms:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \text{ and } d\bar{x} = \sum_{i=1}^{n} \frac{dx_i}{n} dx_i, \text{ giving}$$

$$\sigma_{\bar{x}} = \frac{\sqrt{\sum_{i=1}^{n} \sigma_{x_i}^2}}{n}, \text{ if } \sigma_{x_i} \text{ are the same} = \sigma_x, \text{ then}$$

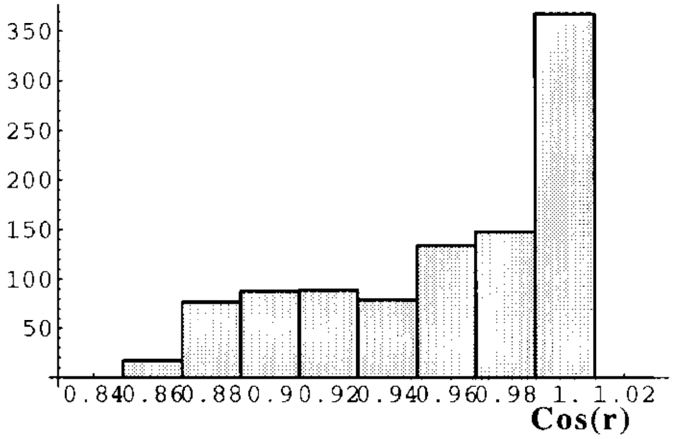
$$\sigma_{\bar{x}} = \frac{\sigma_x}{n}\sqrt{\sum_{i=1}^{n} 1} = \frac{\sigma_x}{n}\sqrt{n} = \frac{\sigma_x}{\sqrt{n}}$$

This is a very important result. It says that the uncertainty in the mean value decreases as the square root of the number of measurements. An important corollary is that the error does not decrease as fast as the number of measurements.

Monte Carlo approach to propagating errors

The analytic approach to error propagation assumes that all higher derivatives of R are negligible compared to the first derivative. This is obviously false whenever R is at a relative maximum or minimum since here the first derivative is zero. An example is the case R=cos(r) when r is 0. Notice that in the analytic description, σ_R= 0! At these points it is necessary to employ a Monte Carlo approach to error determination. Sometimes it is easier than calculating a complicated derivative, even if the analytic approach is in principle fine. The method is quite simple. One generates n random numbers. If a Gaussian generator is available, so much the better. However a uniform distribution with window w = 3.46σ works fine. So r = r_{mean}+(Ran-0.5)*w will do the trick (Ran is distributed from 0->1). Just calculate the function R using each of the generated r values. The resulting distribution represents the uncertainty in R. Note that it may be asymmetric due to the higher derivative terms which are always left out in the analytic approach. Figure 4 illustrates data generated for R=cos(r), with r = 0.0±0.1.

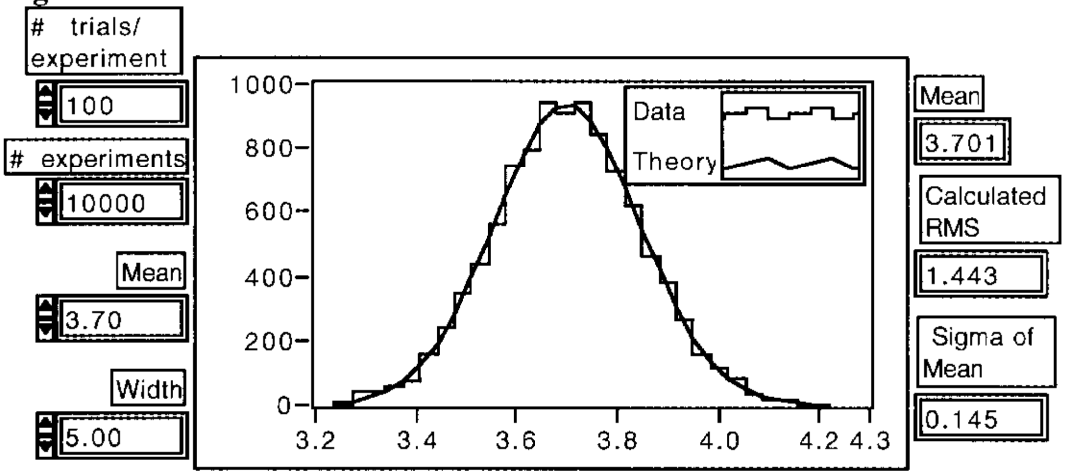
Figure 4. Histogram of Cos(r) with r randomly generated in the interval r = 0.0±0.1.



Central Limit theorem-or Why Daddy is everything a Gaussian?

Given an arbitrary distribution which has a mean and variance defined, the Central Limit Theorem tells us that the distribution of the AVERAGE of N measurements is Gaussian with a variance σ^2/N. The way this is typically interpreted is that most measurements we make are due to averages of processes which are ongoing at the microscopic level. For example, if we measure the current through a circuit, it is the sum of 10^{23} moving electrons. The individual electrons probably have a Maxwell-Boltzmann Distribution, but our measurements of the current will resemble a Gaussian distribution.

Figure 5. Illustration of Central Limit Theorem.



trials/ experiment: 100

experiments: 10000

Mean: 3.70

Width: 5.00

Mean: 3.701

Calculated RMS: 1.443

Sigma of Mean: 0.145

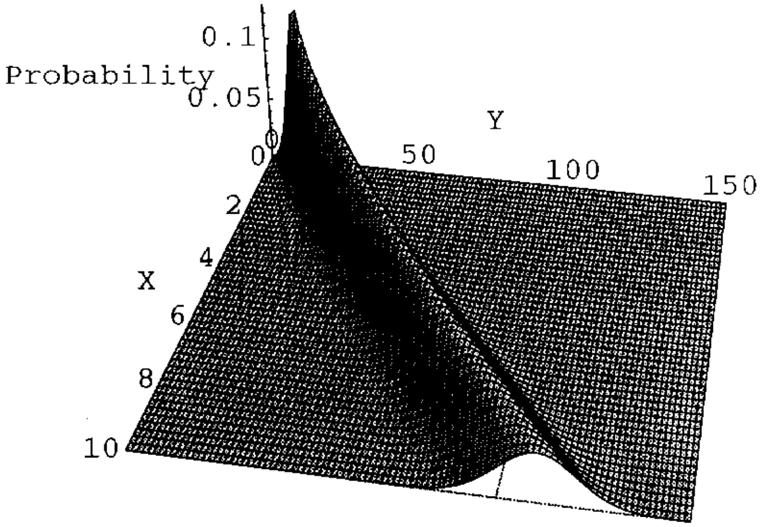
Figure 5 illustrates the Central Limit Theorem. The data on the left of the plot are inputs into the calculation and the data on the right are the outputs. An experiment consists of 100 random numbers (trials) generated uniformly in a window of *Width* =5.0 (rms=1.443), and centered at 3.70 (*Mean*). The sample mean was calculated from the data of each experiment. A total of 10000 experiments were run and the mean from each was histogrammed in the plot. The smooth curve is a Gaussian with $\mu = 3.7$, σ= rms/sqrt(#trials)=0.1443, and area equal to 10000. The curve is in excellent agreement with the histogram. The quantities on the right are calculated from the actual histogram.

ESTIMATION OF PARAMETERS FROM DATA.

Principle of Maximum Likelihood

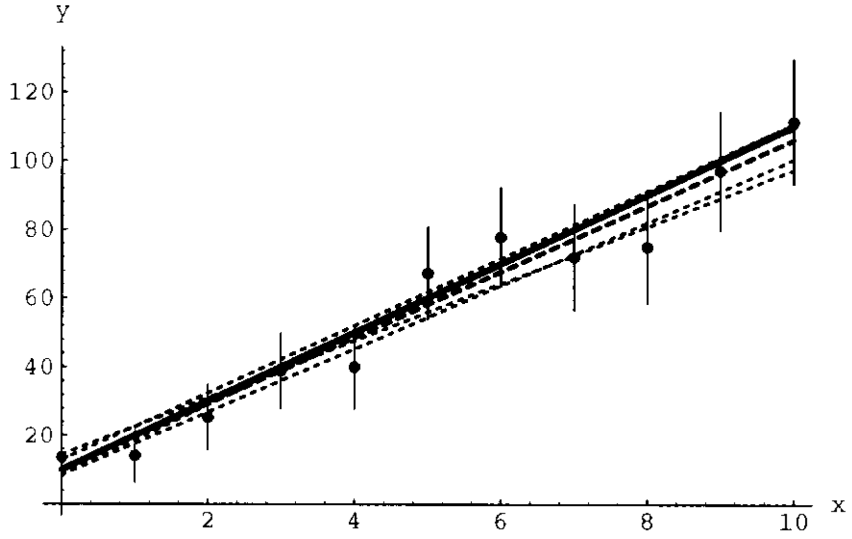
Suppose we have n independent data measurements y_i taken at n points x_i. These points x_i could all be the same point or they could just be the order in which the measurement was made. How can this data be used to learn something about the underlying functional relationship of y versus x? Clearly we have learned how to do this already in the case of an calculating a mean.

Figure 6. The Probablity of measuring y_i at x_i



Let's assume we know the form of the functional relationship, if not all the details. An slightly more complicated example than the mean would be $y(x_i)=a_1 + a_2 * x_i$, where a_1 and a_2 are unknown parameters that we would like to determine. At each independent point x_i, the measurement y_i is distributed about the "true" value $y(x_i)$ with a width σ_i. Figure 6 illustrates this concept. At any point x the measured y_i is distributed according to the Gaussian distribution. The true value *y(x)* is represented by the ridge *y(x)*=10+10*x* on the figure. σ(x) was made to vary as the squareroot of *y(x)* to illustrate the effect of a non-constant σ. Thus the Gaussian distribution is narrower and "peakier" at low values of *x*. However the area under each Gaussian along the *y* direction is equal to one. The probability for a particular measurement y_i is P_i which depends on y_i, $y(x_i)$, and σ_i. Figure 7 shows Monte Carlo data generated using the above assumptions. The data are plottted with the window error bars (3.46σ). The other curves will be explained in the next section.

Figure 7. Data generated along the line *y(x)* = 10 + 10 *x* . The data points are shown with window errors (3.46σ). The Curves are explained in the text.

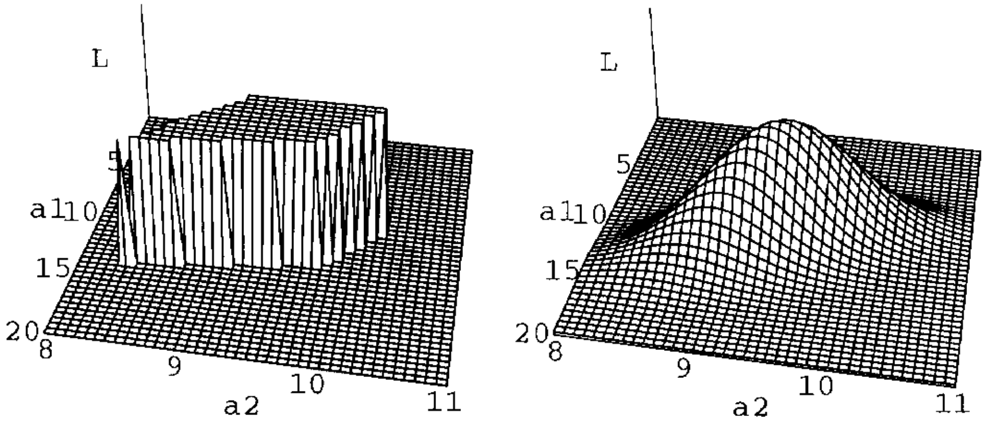


The total probability or Likelihood, L to have measured the particular data set of y_i values is $L = \prod_{i}^{n} P_i$, where P_i is the probability for each particular y_i . Given this data, we would like to estimate values for the parameters *a* of *y(x)*. Coming to our rescue is the Principle of Maximum Likelihood stating that the best estimate we can make for the parameters will be the ones which maximize the Likelihood (or total probability). I like to think of this as the Principle that Nature Plays Fair.

Uniform and Gaussian examples

If the underlying statistical distribution is known, one can calculate the probabilities for different values of the unknown parameters. Figure 8 plots the Uniform and Gaussian Likelihood distribution (using the same equivalent rms (Sqrt(*y*(*x*)))) for the data shown in fig.7 by varying the parameters a_1 and a_2 and recalculating L. Notice that the Uniform Likelihood Distribution is flat - all values for a_1 and a_2 in the plateau are equally likely. All values outside are excluded (L=0). Without invoking a new principle we cannot go further with this distribution. The Gaussian Likelihood is peaked about a particular value of *(a1,a2)* = (10.1,9.6). These values represent the most likely values for *(a1,a2)*. It will turn out that the Gaussian Distribution will have some very convenient mathematical features which will save us from having to do all this work. In fig. 7, the dotted lines are the *(a1,a2)* values from the corners of the Uniform Likelihood distribution. The heavy solid line is the actual theoretical curve *(a1,a2)*=(10,10). The heavy dashed line is determined from the peak of the Gaussian Likelihood Distribution

Figure 8. The Likelihood distributions for the data. The left plot is from a uniform probability function. The right plot from a Gaussian probability function. Both use the same data and rms width.



Development of Least Squares Fits

If the individual *y(x)* distributions are Gaussian, some wonderful things happen. The Likelihood function is:

$$L = \prod_i^n P(y_i, y(x), \sigma(x_i)) = \prod_i^n \frac{1}{\sqrt{2\pi}\sigma(x_i)} e^{-\frac{1}{2}\left(\frac{(y_i - y(x_i))}{\sigma(x_i)}\right)^2}$$

$$= \left(\prod_i^n \left(\frac{1}{\sqrt{2\pi}\sigma(x_i)}\right)\right) e^{-\frac{1}{2}\sum_i^n \left(\frac{(y_i - y(x_i))}{\sigma(x_i)}\right)^2} .$$

Notice that once we have our data set, and have determined the errors at each point, the only dependence in L is in the parameters *a*, which appear only in the function *y(x,a)*. Therefore maximizing the Likelihood is equivalent to minimizing sum of the squares in the exponent. This exponent is known as chisquare,

$$\chi^2 = \sum_i^n \left(\frac{\left(y_i - y(x_i)\right)}{\sigma(x_i)} \right)^2$$

Linear fit to parameters

We can gain some insight by expanding χ^2 in a second order Taylor's series about about its minimum,

$$\chi^2(a) = \chi^2\left(a^0\right) + \sum_j \frac{\partial \chi^2\left(a^0\right)}{\partial a_j} da_j + \frac{1}{2} \sum_{j,k} \frac{\partial^2 \chi^2\left(a^0\right)}{\partial a_j \partial a_k} da_j da_k .$$

First derivative terms

The first term in the expansion is just the value of χ^2 at the minimum. The second term is

$$\frac{\partial \chi^2\left(a^0\right)}{\partial a_j} = \frac{\partial \sum_i \left(\frac{y(x_i,a) - y_i}{\sigma_i} \right)^2}{\partial a_j} = 2 \sum_i \left(\frac{y(x_i,a) - y_i}{\sigma_i^2} \right) \frac{\partial y(x_i,a)}{a_j} .$$

The nomenclature is that "*j*" refers to the jth parameter, and "*i*" to the ith data point.

Since we want to minimize χ^2 ,we set the each of the first derivative terms to zero. If *y(x,a)* can be written as a linear function of the parameters a_j,

$$y(x_i,a) = \sum_j f_j(x_i) a_j ,$$

the mathematics simplifies even further and we have what is known as a Linear Least Squares Fit. With this assumption the equations simplify to:

$$0 = \sum_k \alpha_{jk} a_k - \beta_j \ \text{, where}$$

$$\alpha_{jk} = \sum_i \frac{f_j(x_i) f_k(x_i)}{\sigma_i^2}, \quad \text{and } \beta_j = \sum_i \frac{y_i f_j(x_i)}{\sigma_i^2} .$$

Notice that α_{jk} , which is an element of the curvature matrix (see next section), depends only on the functional form of *y(x)* and the error values but NOT the actual y_i data values. All information about the data are contained in the β_j . These equations simpify in form into a matrix equation $\{\alpha\}(a)=(\beta)$. This equation can be inverted to solve for the a parameters giving

$$(a)=\{\alpha\}^{-1}(\beta)=\{\varepsilon\}(\beta) \text{ or}$$
$$a_j=\varepsilon_{jk}\beta_k \ .$$

ε_{jk} is an element of what is known as the error matrix.

Second derivative terms

With $y(x_i,a)$ a linear function of the a, it is easy to show that the coefficient of the second derivative of χ^2 is α_{jk}. Therefore around the minimum,

$$\chi^2(a)= \chi^2\left(a^0\right)+\sum_{j,k}\alpha_{jk}da_jda_k\,.$$

This is exact for the linear function case. α_{jk}.represents the curvature of χ^2. We will see that a steep curvature of χ^2 implies small uncertainties in the determination of the parameters a.

Errors for the parameters a

Now that we have a mathematical prescription to find the parameters, it is reasonable to ask what are the rms errors of the parameters themselves? The uncertainty can be propagated back to the source of the uncertainty - the errors in y_i. Since $a_j=\varepsilon_{jk}\beta_k$,

$$da_j=\sum_i\frac{\partial a_j}{\partial y_i}dy_i=\sum_i\frac{\partial\left(\sum_k\varepsilon_{jk}\beta_k\right)}{\partial y_i}dy_i=\sum_k\varepsilon_{jk}\sum_i\frac{\partial\beta_k}{\partial y_i}dy_i\,,$$

$$da_j=\sum_k\varepsilon_{jk}\sum_i\frac{\partial\left(\sum_l\frac{y_lf_k(x_l)}{\sigma_l^2}\right)}{\partial y_i}dy_i=\sum_k\varepsilon_{jk}\sum_i\frac{f_k(x_i)}{\sigma_i^2}dy_i\,.\text{ Therefore,}$$

$$\sigma_{jl}^2 = \langle da_j da_l \rangle = \left\langle \left(\sum_k \varepsilon_{jk} \sum_i \frac{f_k(x_i)}{\sigma_i^2} dy_i \right)\left(\sum_m \varepsilon_{lm} \sum_n \frac{f_m(x_n)}{\sigma_n^2} dy_n \right) \right\rangle$$

$$= \sum_{k,m} \varepsilon_{jk}\varepsilon_{lm} \sum_i \frac{f_k(x_i) f_m(x_i)}{\sigma_i^4} \underset{i}{\sigma_i^2} = \sum_{k,m}\varepsilon_{jk}\varepsilon_{lm}\alpha_{km} = \sum_k \varepsilon_{jk}\delta_{lk} = \varepsilon_{jl}$$

This explains why ε_{jk} is called the error matrix.

Explicit example

This section calculates the fit and errors for the sample generated data which has been used throughout this paper. With the function $y_{actual}(x) = 10.0 + 10.0x$, the data are

$$x_i = [0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10]$$
$$y_i = [13.6,\ 14.1,\ 25.3,\ 38.9,\ 40.0,\ 67.3,\ 77.8,\ 72.0, 75.0,\ 97.1,\ 115.3]$$
$$\sigma_{y_i} = [3.2,\ 4.5,\ 5.5,\ 6.3,\ 7.1,\ 7.8,\ 8.9,\ 9.5,\ 10.0,\ 10.5]$$

With these values the matrices are:

$$\{\alpha\} = \begin{Bmatrix} \sum_{i=1}^{11} \frac{1}{\sigma_{y_i}{}^2} & \sum_{i=1}^{11} \frac{x_i}{\sigma_{y_i}{}^2} \\ \sum_{i=1}^{11} \frac{x_i}{\sigma_{y_i}{}^2} & \sum_{i=1}^{11} \frac{x_i^2}{\sigma_{y_i}{}^2} \end{Bmatrix} = \begin{Bmatrix} 0.302 & 0.798 \\ 0.798 & 4.702 \end{Bmatrix} \text{ and,}$$

$$(\beta) = \begin{pmatrix} \sum_{i=1}^{11} \frac{y_i}{\sigma_{y_i}{}^2} \\ \sum_{i=1}^{11} \frac{y_i x_i}{\sigma_{y_i}{}^2} \end{pmatrix} = \begin{pmatrix} 10.67 \\ 53.00 \end{pmatrix} \qquad \{\varepsilon\} = \{\alpha\}^{-1} = \begin{Bmatrix} 6.00 & -1.02 \\ -1.02 & 0.38 \end{Bmatrix}.$$

Solving for the parameters and their errors gives

$$(a) = \{\varepsilon\}(\beta) = \begin{pmatrix} 10.1 \\ 9.56 \end{pmatrix} = \begin{pmatrix} \text{constant} \\ \text{slope} \end{pmatrix} \qquad \sigma_a = \begin{pmatrix} \sqrt{\varepsilon_{11}} \\ \sqrt{\varepsilon_{22}} \end{pmatrix} = \begin{pmatrix} 2.4 \\ 0.62 \end{pmatrix}, \text{ with}$$

$\chi^2 = 10.4\,/\,9$ degrees of freedom at the minimum.

Calibration errors from results

Now that we have determined our parameters it would be useful to use the results. Suppose *y(x,a)* represent a calibration of a system. How do we state the error of the calibration? It is incorrect (in principle) to claim the uncertainty of the calibration as being given by the data fluctuations from the curve. Why? Because if (a big if!) the data are really described by our chosen function, we should be able to do better since the fit is using all the points. A good example would be the error in the mean. If we make 100 measurements of a value, the error in the mean should be a factor of 10 lower than the rms width of the actual data distribution. If we know the rms width, an easy check is calculating the χ^2 value. It should be (as we will see) approximately equal to the number of data points for a good fit. Likewise in the case we have just calculated, our knowledge of the equation of the line is much better than the fluctuations of the data around it. But how well do we know it? The uncertainty can be calculated by propagating the errors. However we will propagate the error only to the parameters instead of going all the way back to the data. The error in $y(x)=a_1+a_2x$ is

$$dy(x,a) = \sum_j \frac{\partial y(x,a)}{\partial a_j} da_j = \sum_j f_j(x) da_j .$$

$$\sigma^2_{y(x,a)} = \left\langle \sum_j f_j(x) da_j \sum_k f_k(x) da_k \right\rangle = \sum_{j,k} f_j(x) f_k(x) \langle da_j da_k \rangle$$

$$= \sum_{j,k} f_j(x) f_k(x) \varepsilon_{jk}$$

In our particular example, the last equation becomes

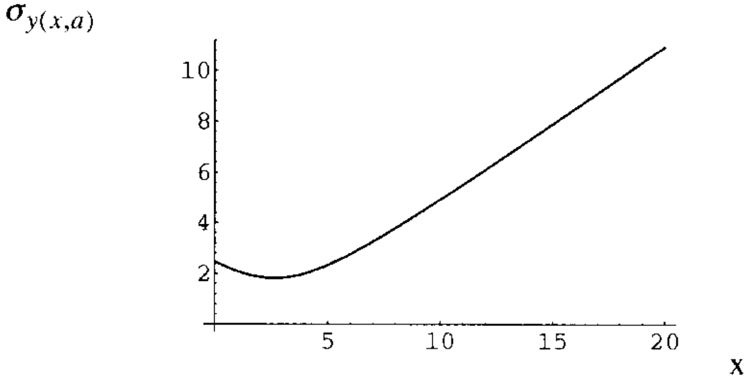
$$\begin{aligned}\sigma_{y(x,a)} &= \sqrt{f_1(x)f_1(x)\varepsilon_{11} + 2f_1(x)f_2(x)\varepsilon_{12} + f_2(x)f_2(x)\varepsilon_{22}} \\ &= \sqrt{\varepsilon_{11} + 2x\varepsilon_{12} + x^2\varepsilon_{22}} \\ &= \sqrt{6.00 - 2.04x + .38x^2}\end{aligned}$$

.

Notice that the parameters are correlated. We cannot ignore the cross terms. The reason for this is that they have been determined from the the same set of data.

$\sigma_{y(x,a)}$ is plotted as a function of x in fig. 9. Notice how the error is smaller at all x values than our data estimations. Also note how the error grows very fast for x values outside our data set (x>10). This tells us to beware extrapolations!

Figure 9. The error in the fit $y(x)= a_1+ a_2x$ as a function of x . The parameters a are the fit values.



Non-Linear Least Square Fits

Log and other endruns

Some functions with non-linear parameters lend themselves to a linear fit by a remapping of the independent and dependent variables. The most common are

$$y = a_1 e^{a_2 x} \rightarrow \log(y) = \log(a_1) + a_2 x \text{ and}$$

$$\text{y} = a_1 x^{a_2} \rightarrow \log(y) = \log(a_1) + a_2 \log(x).$$

The new variables are log(y) and x, and log(y) and log(x) respectively. This was an important feature, especially when most data analysis took place on graph paper (remember semi-log, logarithmic, and probability paper ?). Even now the technique is useful due to the simplicity and availability of software on computers and calculators. However caution should be applied since the error bars are certainly not equal (generally what the software assumes), even if they were for the original y values. Typically the smaller values of y are overweighted if one assumes equal errors for log(y). This will cause the fit to overemphasize the small y values. In addition the error bars are no longer symmetric about the "true" value. For example if the error is ±10 and y=10, y_{data} might range from 0 to 20. On a log scale, the error bar should range from -infinity to log(20).

Linearization

If $y(x)$ is not a linear function of the parameters, what can be done? Due to the success of the Linear Least Squares Fit mathematics, the answer is obvious- linearize the function! Expanding $y(x,a)$ about the parameters a gives:

$$y(x,a) = y\left(x,a^0\right) + \sum_j \frac{\partial y\left(x,a^0\right)}{\partial a_j} da_j, \text{ and}$$

$$\chi^2 = \sum_i \frac{\left(y(x_i,a) - y_i\right)^2}{\sigma_i^2} \to \sum_i \frac{\left(y\left(x_i,a^0\right) - y_i + \sum_j \frac{\partial y\left(x,a^0\right)}{\partial a_j} da_j\right)^2}{\sigma_i^2}.$$

The entire formalism for the linear least squares fit can be taken over with the substitution of $(y(x_i) - y_i)$ for $(-y_i)$ and realizing that the parameters are now da_j. The a^0 's are considered as constants, and hopefully close to their true values. The implemented algorithms usually require the user to make an initial guess at the a^0 . In addition one usually supplies the fit with the analytic partial derivatives, again computed at the a^0 starting points. Some algorithms will compute the derivatives numerically for you, but this triples the number of calculations of a complicated function (which is already being evaluated at every x value). The fit calculates the da's which then are added to the a^0's and the process starts over again, this time using the new a^0's to calculate $y(x)$ and its derivatives again. The process can be numerically intensive if there are a lot of data points. Generally the user uses some requirement to stop the iteration such as Chisquare not changing. The definition of errors in the parameters is the same ε_{jj} as in the linear case.

As has been mentioned, many packages exist in spreadsheet, statistical, and lab data acquisition software which will fit non-linear functions. The art of the process is starting the fit off at the right point. It is one thing to fit a curve by hand-specifying the starting parameters, and quite another to automate the sequence under all conditions the aaccelerator provides. If the signal to noise of the data is good, it will usually be easy to automate the process. Most of our work has been done with functions which are close to being Gaussian in shape and on a reasonable background:

$$y(x,a) = a_1 + a_2 x + a_3 e^{-\frac{1}{2}\left(\frac{x-a_4}{a_5}\right)^2}.$$

We typicallly estimate the background a_1 at the ends of the data array, set a_2 = 0, a_3 = (maximum value - a_1), a_4 = x value of the maximum value, and finally find the full width half maximum (FWHM) peak heights on either side of the maximum and set a_5 =(FWHM/2.35). This method is reasonably insensitive the poor signal to noise ratios.

Chisquare Phenomenology

Chisquare Contours

If the contours of χ^2 are rotated ellipses with respect to the parameter axes (easiest to visualize in a two parameter fit), the error matrix is not diagonal (normal case). If the ellipse's major and minor axes are aligned with the parameter axes, then the errors in the parameters have no correlations (and the error matrix is diagonal). A simple example (case 1) can be had from our simple Linear Least Squares Fit to the line *y(x)*= 10+10*x*. (assume constant errors for this example). The contour of χ^2 is shown in fig.10a. It is rotated. If we use a function (case 2) y(x)=a1+a2*(x-5), the χ^2 contour is shown fig.10b. What we have done is replaced the y intercept of the first function with a constant which is now equivalent to the average y value (recall that x = 5 is the center x value of our data). Since the average y value is independent of the slope,we have "decoupled" the two parameters. If one recalls the formula for the off-diagonal element of {α}, the two parameter version has the value

$$\alpha_{12} = \sum_{i=1}^{n} \left(\frac{\left(\frac{\partial y(x_i)}{\partial a_1} \frac{\partial y(x_i)}{\partial a_2} \right)}{\sigma_i^2} \right).$$

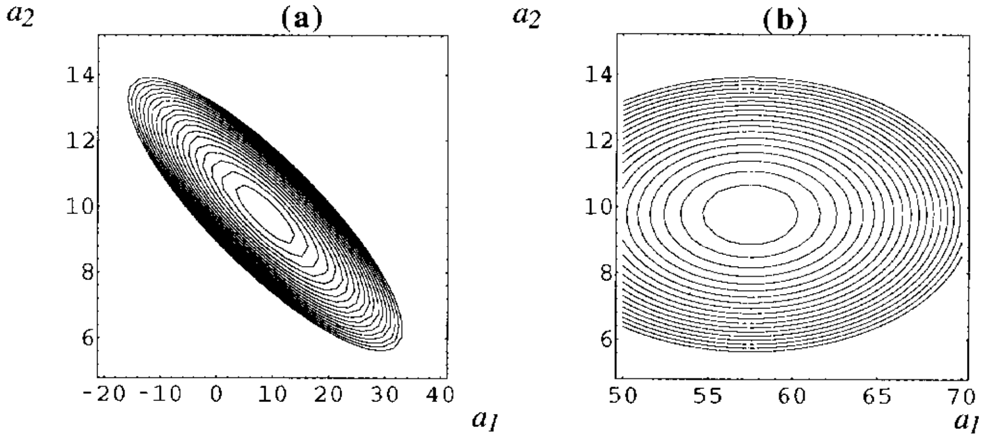
For case 1 and 2 respectively, these values are $\alpha_{12} = 0.55$ and $\alpha_{12} = 0$. The values of x = (0, 1, 2, 3,..10), are the values used in generating the data. This example suggests how to "orthogonalize" the function, even in the case where σ_i is not constant. Set the off-diagonal elements of {α}=0 and solve for the constant. The actual goodness of the fit is exactly the same in both cases. The difference is that it is easier to understand the errors of the parameters in case 2. Practically speaking, this technique is almost never used. However I plan to use it to make a point in the next paragraph.

If we look at the contours in fig. 10, the first contour drawn is the one which is one (1) higher than the minimum value of χ^2. This means that the probability for our data to have been generated by the values of the pararameters which lie on this contour is

$$L \propto e^{-\frac{1}{2}\left(\chi^2_{\min}+1\right)} = L_{\max} e^{-\frac{1}{2}} = 0.61 L_{\max}.$$

In case 2 with the "orthogonalized" function, the error matrix diagonal elements are just the inverse of the curvature matrix diagonals (since {α} is diagonal). Thus it is true that the one sigma error in the parameter is given by the increment which increase χ^2 by one unit since $\chi^2(a)= \chi^2\left(a^0\right)+\alpha_{jj}\varepsilon_{jj}=\chi^2\left(a^0\right)+1$. This can shown to be true even in case 1, as long as the other parameters are re-optimised. As a matter of fact, the technique of finding the 1 sigma contour is the best way to define the parameter errors since it works even in the case when χ^2 is not well described by the second order expansion - the case in some nonlinear fits. It should be noted that the second order expansion of χ^2 is exact for all functions which are linear in their parameters. Also note again that {α}, the curvature, is completely set by the estimate of the data error (once the function has been set and the data measured).The actual data points y_i have nothing to do with {α}. This means that the parameter errors are very sensitive to the estimation of errors. As a side note, the parameters are also sensitive to the x_i values that the y_i are measured at. One cannot expect a fit to work without sampling the data where the function is significant.

Figure 10. Contour Plots of χ^2 for the Linear Fit. The horizontal axis is a_1 , and the vertical axis is a_2. (a) is using the function $y(x) = a_1 + a_2 x$ while (b) uses the function $y(x) = a_1 + a_2 (x -5)$.



After all this work it would be nice to get error bars from our fits. Unfortunately most software packages leave out this step. This is fine if the error bars are the same magnitude for all data points, because at least the formulae for the the parameters are correct. Even still most packages don't calculate either the curvature or error matrix so they leave it to you anyway. Calculating the curvature matrix is not such a big deal if you really need it, and most packages do have matrix inversion routines which will get you the error matrix. If the errors are really varying dramatically across the data set, these packages may not even calculate the parameters very well.

Of course the major difficulty is estimating the error bars in the first place. I suspect most of us do this empirically by checking the repeatability of our measurements. If one is fitting curves automatically under widely varying conditions, this option isn't generally available. A trick which does work if one can assume equal errors is to measure the point to point fluctuations in the background region of the fit with a simple rms calculation. If it works, use it!

Chisquare and goodness of Fit:

Most of us have heard about the goodness of the fit or χ^2. If we look back at the definitions of χ^2 we see that if the theoretical function *y(x)* is close to reality, and if we have estimated the errors at each point correctly, $(y(x)\text{-}y_i)$ should sometimes be less than σ and sometimes greater, since a Gaussian distribution has been assumed. Therefore over a data set we should expect that each term in the sum should contribute one unit to χ^2, and χ^2 will approximately equal the total number of data points N. Usually we divide χ^2 by the number of degrees of freedom ν = (N-n) to give what is known as the "reduced" chisquare or $\chi^2_{\nu}. = \chi^2/(\text{N-n})$. χ^2_{ν} by the preceding arguments should now be approximately 1 if the fit is good. Probability tables exist for χ^2_{ν}. which can be used to check the goodness of the fit.

What if χ^2_{ν}. is too ridiculously big or small? The first question should deal with the σ estimations. Are they too small (large χ^2_{ν}) or too big (small χ^2_{ν}). If this looks fine then one wonders about the function and/or the data themselves. Unfortunately there is no easy way to disentangle the two except by looking. Is the data asymmetric about the mean?, ...Is the background handled correctly?, ...Are there obviously bad data points? This is the black magic of the process.

Fitting versus simple statistical analysis

Suppose one is trying to extract the mean and width of a gaussian-looking peak, which is riding on a noise floor, i.e.

$$y(x) = Ae^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} + B.$$

When is a simple moment analysis appropriate to get the centroid and width of the peak?

I have found that when the signal to noise is 10/1 or better, it is possible to forego fancy fitting and do a simple moments analysis. By this is meant

$$\mu = \frac{\sum_{i=1}^{n}(y_i - B)x_i}{\sum_{i=1}^{n}(y_i - B)} \text{ and } \sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i - B)(x_i - \mu)^2}{\sum_{i=1}^{n}(y_i - B)}} \; .$$

These are simply the weighted μ and σ of the x values, using the y_i 's as weights Before one can make this calculation it is necessary to strip away the noise floor. The resulting (y_i-B) (in the background region) will fluctuate both positively and negatively around zero, and on the average cancel. However if the fluctuations are large compared to A, the peak amplitude, the results (especially σ) will be very erratic. One must always be on guard under these conditions.

CONCLUSION

Statistical Analysis of data is within the reach of all Instrumentation Engineers. This tutorial has tried to stress the assumptions which are implicit in most analyses. One should realise that we have only scratched the tip of the iceberg.

1(1) P.R.Bevington, Data Reduction and Error Analysis for the Physical Sciences,McGraw-Hill Book Co.(1969).
(2) E.L.Barsotti Jr., "A Longitudinal Bunch Monitoring System Using LabVIEW and High-speed Oscilloscopes", Contributed Poster this Workshop
(3) J.R.Zagel, D.Chen, and J.Crisp, " Fermilab Booster Ion Profile Monitor System Using LabVIEW", Contributed Poster this Workshop