NOTES ON STATISTICS FOR PHYSICISTS, REVISED

Jay Orear

Laboratory for Nuclear Studies
Cornell University
Ithaca, NY 14853

Table of Contents

REFERENCES

ORIGINAL PREFACE

These notes are based on a series of lectures given at the Radiation Laboratory in the summer of 1958. I wish to make clear my lack of familiarity with the mathematical literature and the corresponding lack of mathematical rigor in this presentation. The primary source for the basic material and approach presented here was Enrico Fermi. My first introduction to much of the material here was in a series of discussions with Enrico Fermi, Frank Solmitz, and George Backus at the University of Chicago in the autumn of 1953. I am grateful to Dr. Frank Solmitz for many helpful discussions and I have drawn heavily from his report "Notes on the Least Squares and Maximum Likelihood Methods." [1] The general presentation will be to study the Gausssian distribution, binomial distribution, Poisson distribution, and least-squares method in that order as applications of the maximum-likelihood method.

August 13, 1958

PREFACE TO REVISED EDITION

Lawrence Radiation Laboratory has granted permission to reproduce the original UCRL-8417. This revised version consists of the original version with corrections and clarifications including some new topics. Three completely new appendices have been added.

Jay Orear
July 1982

1. DIRECT PROBABILITY

Books have been written on the "definition" of probability. We shall merely note two properties: (*a*) statistical independence (events must be completely unrelated), and (*b*) the law of large numbers. This says that if p_1 is the probability of getting an event in Class 1 and we observe that N_1 out of N events are in Class 1, then we have

$$\lim_{N\to\infty}\left|\frac{N_1}{N}\right| = p_1.$$

A common example of direct probability in physics is that in which one has exact knowledge of a final-state wave function (or probability density). One such case is that in which we know in advance the angular distribution $f(x)$, where $x = \cos\theta$ of a certain scattering experiment, In this example one can predict with certainty that the number of particles that leave at an angle x_1 in an interval Δx_1 is $Nf(x_1)\Delta x_1$, where N, the total number of scattered particles, is a very large number. Note that the function $f(x)$ is normalized to unity:

$$\int_{-1}^{1} f(x)dx = 1.$$

As physicists, we call such a function a distribution function. Mathematicians call it a probability density function. Note that an element of probability, dp, is

$$dp = f(x)dx.$$

2. INVERSE PROBABILITY

The more common problem facing a physicist is that he wishes to determine the final-state wave function from experimental measurements. For example, consider the decay of a spin-½ particle, the muon, which does not conserve parity. Because of angular-momentum conservation, we have the a priori knowledge that

$$f(x) = \frac{1+\alpha x}{2}$$

However, the numerical value of α is some universal physical constant yet to be determined. We shall always use the subscript zero to denote the true physical value of the parameter under question. It is the job of the physicist to determine α_0. Usually the physicist does an experiment and quotes a result $\alpha = \alpha^* \pm \Delta\alpha$. The major portion of this report is devoted to the questions What do we mean by α^* and $\Delta\alpha$? and What is the "best" way to calculate α^* and $\Delta\alpha$? These are questions of extren importance to all physicists.

Crudely speaking, $\Delta\alpha$ is the standard deviation, [2] and what the physicist usually means is that the "probability" of finding

$$(\alpha^* - \Delta\alpha) < \alpha_0 < (\alpha^* + \Delta\alpha) \quad \text{is } 68.3\%$$

(the area under a Gaussian curve out to one standard deviation). The use of the word "probability" in the previous sentence would shock a mathematician. He would say the probability of having

$$(\alpha^* - \Delta\alpha) < \alpha_0 < (\alpha^* + \Delta\alpha) \quad \text{is either } 0 \text{ or } 1.$$

The kind of probability the physicist is talking about here we shall call inverse probability, in contrast to the direct probability used by the mathematician. Most physicists use the same word, probability, for the two completely different concepts: direct probability and inverse probability. In the remainder of this report we will conform to this sloppy physicist-usage of the word "probability."

3. LIKELIHOOD RATIOS

Suppose it is known that either Hypothesis A or Hypothesis B must be true. And it is also known that if A is true the experimental distribution of the variable x must be $f_A(x)$, and if B is true the distribution is $f_B(x)$. For example, if Hypothesis A is that the K meson has spin zero, and hypothesis B that it has spin 1, then it is "known" that $f_A(x) = 1$ and $f_B(x) = 2x$, where x is the kinetic energy of the decay π^- divided by its maximum value for the decay mode $K^+ \rightarrow \pi^- + 2\pi^+$.

If A is true, then the joint probability for getting a particular result of N events of values $x_1, x_2, ..., x_N$ is

$$dp_A = \prod_{i=1}^{N} f_A(x_i)dx_i.$$

The likelihood ratio is

$$\mathcal{R} = \prod_{i=1}^{N} \frac{f_A(x_i)}{f_B(x_i)}. \quad (1)$$

This is the probability, that the particular experimental result of N events turns out the way it did, assuming A is true, divided by the probability that the experiment turns out the way it did, assuming B is true. The foregoing lengthy sentence is a correct statement using direct probability. Physicists have a shorter way of saying it by using inverse probability. They say Eq. (1) is the betting odds of A against B. The formalism of inverse probability assigns inverse probabilities whose ratio is the likelihood ratio in the case in which there exist no prior probabilities favoring A or B. [3] All the remaining material in this report is based on this basic principle alone. The modifications applied when prior knowledge exists are discussed in Sec. 10.

An important job of a physicist planning new experiments is to estimate beforehand how many events he will need to "prove" a hypothesis. Suppose that for the $K^+ \rightarrow \pi^- + 2\pi^+$ one wishes to establish betting odds of 10^4 to 1 against spin 1. How many events will be needed for this? The problem and the general procedure involved are discussed in Appendix I: Prediction of Likelihood Ratios.

4. MAXIMUM-LIKELIHOOD METHOD

The preceding section was devoted to the case in which one had a discrete set of hypotheses among which to choose. It is more common in physics to have an infinite set of hypotheses; i.e., a parameter that is a continuous variable. For example, in the μ-e

decay distribution

$$f(\alpha;x) = \frac{1+\alpha x}{2},$$

the possible values for α_0 belong to a continuous rather than a discrete set. In this case, as before, we invoke the same basic principle which says the relative probability of any two different values of α is the ratio of the probabilities of getting our particular experimental results, x_i, assuming first one and then the other, value of α is true. This probability function of α is called the likelihood function, $\mathcal{L}(\alpha)$.

$$\mathcal{L}(\alpha) = \prod_{i=1}^{N} f(\alpha;x_i) \qquad (2)$$

The likelihood function, $\mathcal{L}(\alpha)$, is the joint probability density of getting a particular experimental result, x_1, ... , x_n, assuming $f(\alpha;x)$ is the true normalized distribution function:

$$\int f(\alpha;x)dx = 1.$$

The relative probabilities of α can be displayed as a plot of $\mathcal{L}(\alpha)$ vs. α. The most probable value of α is called the maximum-likelihood solution α^*. The rms (root-mean-square) spread of α about α^* is a conventional measure of the accuracy of the determination $\alpha = \alpha^*$. We shall call this $\Delta\alpha$.

$$\Delta\alpha = \left| \frac{\int (\alpha-\alpha^*)^2 \mathcal{L} d\alpha}{\int \mathcal{L} d\alpha} \right|^{\frac{1}{2}} \qquad (3)$$

In general, the likelihood function will be close to Gaussian (it can be shown to approach a Gaussian distribution as $N \to \infty$) and will look similar to Fig. 1b.

Fig. 1a represents what is called a case of poor statistics. In such a case, it is better to present a plot of $\mathcal{L}(\alpha)$ rather than merely quoting α^* and $\Delta\alpha$. Straightforward procedures for obtaining $\Delta\alpha$ are presented in Sections 6 and 7.

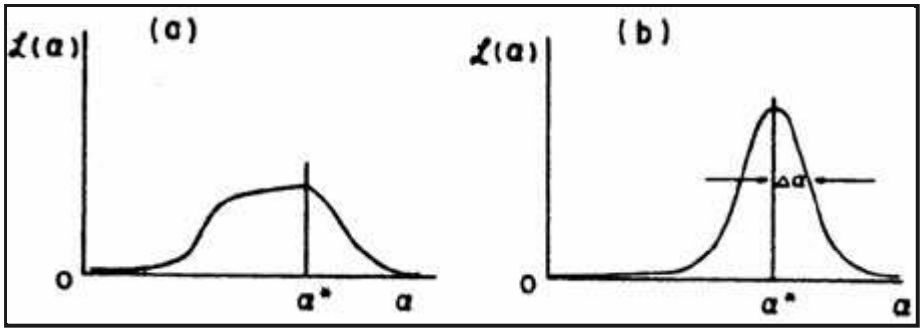

Figure 1. Two examples of likelihood functions $\mathcal{L}(\alpha)$.

A confirmation of this inverse probability approach is the Maximum-Likelihood Theorem, which is proved in Cramer [4] by use of direct probability. The theorem states that in the limit of large N, $\alpha^* \to \alpha_0$; and furthermore, there is no other method of estimation that is more accurate.

In the general case in which there are M parameters, α_1, ..., α_M, to be determined, the procedure for obtaining the maximum likelihood solution is to solve the M simultaneous equations,

$$\left.\frac{\partial w}{\partial \alpha_i}\right|_{\alpha_i=\alpha_i^*} = 0 \quad \text{where } w \equiv \ln \mathcal{L}(\alpha_1, \ldots, \alpha_M), \qquad (4)$$

5. GAUSSIAN DISTRIBUTIONS

As a first application of the maximum-likelihood method, we consider the example of the measurement of a physical parameter α_0, where x is the result of a particular type of measurement that is known to have a measuring error σ. Then if x i Gaussian-distributed, the distribution function is

$$f(\alpha_0; x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp[-(x-\alpha_0)^2/2\sigma^2].$$

For a set of N measurements x_i, each with its own measurement error σ_i the likelihood function is

$$\mathcal{L}(\alpha) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp[-(x_i-\alpha)^2/2\sigma_i^2];$$

then

$$w = -\frac{1}{2}\sum_{i=1}^{N} \frac{(x_i-\alpha)^2}{\sigma_i^2} + \text{ constant};$$

$$\frac{\partial w}{\partial \alpha} = \sum \frac{x_i-\alpha}{\sigma_i^2}, \qquad (5)$$

$$\sum \frac{x_i}{\sigma_i^2} - \sum \frac{\alpha^*}{\sigma_i^2} = 0;$$

The maximum-likelihood solution is

$$\alpha^* = \frac{\sum \frac{1}{\sigma_i^2} x_i}{\sum \frac{1}{\sigma_i^2}} \quad \text{The weighted mean.} \qquad (6)$$

Note that the measurements must be weighted according to the inverse squares of their errors. When all the measuring errors are the same we have

$$\alpha^* = \frac{\sum x_i}{N}.$$

Next we consider the accuracy of this determination.

6. MAXIMUM-LIKELIHOOD ERROR, ONE PARAMETER

It can be shown that for large N, $\mathcal{L}(\alpha)$ approaches a Gaussian distribution. To this approximation (actually the above example is always Gaussian in α), we have

$$\mathcal{L}(\alpha) \propto \exp[-(h/2)(\alpha-\alpha^*)^2],$$

where $1/\sqrt{h}$ is the rms spread of α about α^*,

$$\begin{aligned} w &= -\frac{h}{2}(\alpha-\alpha^*)^2 + \text{ constant}, \\ \frac{\partial w}{\partial \alpha} &= h(\alpha-\alpha^*), \\ \frac{\partial^2 w}{\partial \alpha^2} &= -h \end{aligned}$$

Since $\Delta\alpha$ as defined in Eq. (3) is $1/\sqrt{h}$, we have

$$\Delta\alpha = \left[-\frac{\partial^2 w}{\partial \alpha^2}\right]^{-\frac{1}{2}} \quad \text{Maximum-likelihood Error} \qquad (7)$$

It is also proven in Cramer [4] that no method of estimation can give an error smaller than that of Eq. 7 (or its alternate form Eq. 8). Eq. 7 is indeed very powerful and important. It should be at the fingertips of all physicists. Let us now apply this formula to determine the error associated with α^* in Eq. 6. We differentiate Eq. 5 with respect to α. The answer is

$$\frac{\partial^2 w}{\partial \alpha^2} = \sum \frac{-1}{\sigma_i^2}.$$

Using this in Eq. 7 gives

$$\Delta\alpha = \left[\sum \frac{1}{\sigma_i^2}\right]$$

This formula is commonly known as the law of combination of errors and refers to repeated measurements of the same quantity which are Gaussian-distributed with "errors" σ_i.

In many actual problems, neither α^* nor $\Delta\alpha$ may be found analytically. In such cases the curve $\mathcal{L}(\alpha)$ can be found numerically by trying several values of α and using Eq. (2) to get the corresponding values of $\mathcal{L}(\alpha)$. The complete function is then obtained by drawing a smooth curve through the points. If $\mathcal{L}(\alpha)$ is Gaussian-like, $\partial^2 w / \partial\alpha^2$ is the same everywhere. If not, it is best to use the average

$$\overline{\frac{\partial^2 w}{\partial \alpha^2}} = \frac{\int (\partial^2 w/\partial\alpha^2)\mathcal{L}\,d\alpha}{\int \mathcal{L}\,d\alpha}$$

A plausibility argument for using the above average goes as follows: If the tails of $\mathcal{L}(\alpha)$ drop off more slowly than Gaussian tails, $\overline{\frac{\partial^2 w}{\partial \alpha^2}}$ is smaller than

$$\left.\frac{\partial^2 w}{\partial \alpha^2}\right|_{\alpha^*}$$

Thus, use of the average second derivative gives the required larger error.

Note that use of Eq. 7 for $\Delta\alpha$ depends on having a particular experimental result before the error can be determined. However, it is often important in the design of experiments to be able to estimate in advance how many data will be needed in order to obtain a given accuracy. We shall now develop an alternate formula for the maximum-likelihood error, which depends only on

knowledge of $f(\alpha; x)$. Under these circumstances we wish to determine $\overline{\frac{\partial^2 w}{\partial \alpha^2}}$ averaged over many repeated experiments consisting of *N* events each. For one event we have

$$\overline{\frac{\partial^2 w}{\partial \alpha^2}} = \int \frac{\partial^2 \ln f}{\partial \alpha^2} f \, dx;$$

for *N* events

$$\overline{\frac{\partial^2 w}{\partial \alpha^2}} = N \int \frac{\partial^2 \ln f}{\partial \alpha^2} f \, dx$$

This can be put in the form of a first derivative as follows:

$$\frac{\partial^2 \ln f}{\partial \alpha^2} = \frac{\partial}{\partial \alpha}\left(\frac{1}{f}\frac{\partial f}{\partial \alpha}\right) = -\frac{1}{f^2}\left(\frac{\partial f}{\partial \alpha}\right)^2 + \frac{1}{f}\frac{\partial^2 f}{\partial \alpha^2}$$

$$\int \frac{\partial^2 \ln f}{\partial \alpha^2} f \, dx = -\int \frac{1}{f}\left(\frac{\partial f}{\partial \alpha}\right)^2 dx + \int \frac{\partial^2}{\partial \alpha^2} f \, dx.$$

The last integral vanishes if one integrates before the differentiation because

$$\int f \, dx = 1$$

Thus

$$\overline{\frac{\partial^2 w}{\partial \alpha^2}} = -N \int \frac{1}{f}\left(\frac{\partial f}{\partial \alpha}\right)^2 dx,$$

and Eq. (7) leads to

$$\Delta\alpha = \frac{1}{\sqrt{N}}\left[\int \frac{1}{f}\left(\frac{\partial f}{\partial \alpha}\right)^2 dx\right]^{-\frac{1}{2}} \quad \text{maximum-likelihood error} \qquad (8)$$

Example 1

Assume in the μ-e decay distribution function, $f(\alpha; x) = (1 + \alpha x) / 2$, that $\alpha_0 = -1/3$. How many μ-e decays are needed to establish a to a 1% accuracy (i.e., $\alpha / \Delta\alpha = 100$)?

$$\frac{\partial f}{\partial \alpha} = \frac{x}{2}$$

$$\int_{-1}^{1} \frac{1}{f}\left(\frac{\partial f}{\partial \alpha}\right)^2 dx = \int_{-1}^{1} \frac{x^2}{2(1+\alpha x)} dx = \frac{1}{2\alpha^3}\left[\ln\left(\frac{1+\alpha}{1-\alpha}\right) - 2\alpha\right]$$

$$\Delta\alpha = \frac{1}{\sqrt{N}}\sqrt{\frac{2\alpha^3}{\ln\frac{1+\alpha}{1-\alpha} - 2\alpha}}$$

Note that

$$\lim_{\alpha \to 0}[\Delta\alpha] = \sqrt{\frac{3}{N}}.$$

For

$$\alpha = -\frac{1}{3}, \quad \Delta\alpha = \sqrt{\frac{2.8}{N}}.$$

For this problem

$$\Delta\alpha = \frac{1}{300}, \quad N = 2.52 \times 10^5 \text{ events}.$$

7. MAXIMUM-LIKELIHOOD ERRORS, M-PARAMETERS CORRELATED ERRORS

When *M* parameters are to be determined from a single experiment containing *N* events, the error formulas of the preceding section are applicable only in the rare case in which the errors are uncorrelated.. Errors are uncorrelated only for $\overline{(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)} = 0$ for all cases with $i \neq j$. For the general case we Taylor-expand $w(\alpha)$ about (α^*):

$$w(\alpha) = w(\alpha^*) + \sum_{a=1}^{M} \left(\frac{\partial w}{\partial \alpha_a}\bigg|_{\alpha^*} \right) \beta_a - \frac{1}{2} \sum_a \sum_b H_{ab}\, \beta_a\, \beta_b + \cdots,$$

where

$$\beta_i \equiv \alpha_i - \alpha_i^*$$

and

$$H_{ij} \equiv -\frac{\partial^2 w}{\partial \alpha_i \partial \alpha_j}\bigg|_{\alpha^*}. \qquad (9)$$

The second term of the expansion vanishes because $\partial w / \partial \alpha = 0$ are the equations for α^*

$$\ln \mathcal{L}(\alpha) = w(\alpha^*) - \frac{1}{2} \sum_a \sum_b H_{ab}\, \beta_a\, \beta_b + \cdots.$$

Neglecting the higher-order terms, we have

$$\mathcal{L}(\alpha) = \mathrm{c}\ \exp\left(-\frac{1}{2} \sum_a \sum_b H_{ab}\, \beta_a\, \beta_b\right),$$

(an M-dimensional Gaussian surface). As before, our error formulas depend on the approximation that $\mathcal{L}(\alpha)$ is Gaussian-like in the region $\alpha_i \approx \alpha_i^*$. As mentioned in Section 4, if the statistics are so poor that this is a poor approximation, then one should merely present a plot of $\mathcal{L}(\alpha)$. (see Appendix IV).

According to Eq. (9), *H* is a symmetric matrix. Let *U* be the unitary matrix that diagonalizes *H*:

$$\underline{U}\cdot\underline{H}\cdot\underline{U}^{-1}=\begin{bmatrix} h_1 & & 0 \\ & h_2 & \\ 0 & \ddots & h_M \end{bmatrix}\equiv\underline{h}\quad\text{where }\underline{U}^T=\underline{U}^{-1}. \qquad (10)$$

Let $\underline{\beta}=(\beta_1,\beta_2,\ldots,\beta_M)$ and $\underline{\gamma}\equiv\underline{\beta}\cdot\underline{U}^{-1}$.. The element of probability in the β-space is

$$d^M p = \mathrm{c}\ \exp[-\frac{1}{2}(\underline{\gamma}\cdot\underline{U})\cdot H\cdot(\underline{\gamma}\cdot\underline{U})^T]d^M\beta.$$

Since $|\underline{U}| = 1$ is the Jacobian relating the volume elements $d^M\beta$ and $d^M\gamma$, we have

$$d^M p = \mathrm{c}\ \exp[-\frac{1}{2}\sum_a h_a\gamma_a^2]d^M\gamma.$$

Now that the general M-dimensional Gaussian surface has been put in the form of the product of independent one-dimensional Gaussians we have

$$\overline{\gamma_a\gamma_b}=\delta_{ab}h_a^{-1}.$$

Then

$$\begin{aligned}\overline{\beta_i\beta_j} &= \sum_a\sum_b \overline{\gamma_a\gamma_b}U_{ai}U_{bj}\\ &= \sum_a U_{ia}^{-1}h_a^{-1}U_{aj}\\ &= (\underline{U}^{-1}\cdot\underline{h}\cdot\underline{U})_{ij}^{-1}.\end{aligned}$$

According to Eq. (10), $\underline{H} = \underline{U}^{-1}\cdot\underline{h}\cdot\underline{U}$, so that the final result is

$$\overline{(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)}=(\underline{H}^{-1})_{ij}\quad\text{where } H_{ij}=-\frac{\partial^2 w}{\partial\alpha_i\partial\alpha_j}$$

Averaged over repeated experiments

$$\overline{H}_{ij}=N\int\frac{1}{f}\left(\frac{\partial f}{\partial\alpha_i}\right)\left(\frac{\partial f}{\partial\alpha_j}\right)\,dx$$

Maximum Likelihood Errors, M parameters (11)

(A rule for calculating the inverse matrix $\underline{H}^{-1}$ is

$$(H^{-1})_i=(-1)^{i+j}\times\frac{\text{ij th minor of }\underline{\mathrm{H}}}{\text{determinant of }\underline{\mathrm{H}}}\cdot)$$

If we use the alternate notation $\underline{V}$ for the error matrix $\underline{H}^{-1}$, then whenever $\underline{H}$ appears, it must be replaced with $\underline{V}^{-1}$; i.e., the likelihood function is

$$\mathcal{L}(\alpha)\propto\exp[-\frac{1}{2}\underline{\beta}\cdot\underline{V}^{-1}\cdot\underline{\beta}^T] \qquad (11a)$$

Example 2

Assume that the ranges of monoenergetic particles are Gaussian-distributed with mean range α_1 and straggling coefficient α_2 (the standard deviation). *N* particles having ranges x_1,..., x_N are observed. Find α_1*, α_2*, and their errors . Then

$$\begin{aligned}
\mathcal{L}(\alpha_1,\alpha_2) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\alpha_2} \exp[-(x_i-\alpha_1)^2/2\alpha_2^2] \\
w &= -\frac{1}{2}\sum_i \frac{(x_i-\alpha_1)^2}{\alpha_2^2} - N\ln\alpha_2 - N\ln(2\pi) \\
\frac{\partial w}{\partial \alpha_1} &= \sum_i \frac{(x_i-\alpha_1)}{\alpha_2^2}, \\
\frac{\partial w}{\partial \alpha_2} &= \frac{1}{\alpha_2^3}\sum_i (x_i-\alpha_1)^2 - \frac{N}{\alpha_2}.
\end{aligned}$$

The maximum-likelihood solution is obtained by setting the above two equations equal to zero.

$$\begin{aligned}
\alpha_1^* &= \frac{1}{N}\sum_i x_i \\
\alpha_2^* &= \sqrt{\frac{\sum(x_i-\alpha_1)^2}{N}}
\end{aligned}$$

The reader may remember a standard-deviation formula in which *N* is replaced by (*N* - 1):

$$\overline{\alpha_2} = \sqrt{\frac{\sum(x_i-\alpha_1^*)^2}{N-1}}$$

This is because in this case the most probable value, α_2*, and the mean, $\bar{\alpha}_2$, do not occur at the same place. Mean values of such quantities are studied in Section 16. The matrix H is obtained by evaluating the following quantities at α_1* and α_2*:

$$\frac{\partial^2 w}{\partial \alpha_1^2} = -\frac{N}{\alpha_2^2},\ \frac{\partial^2 w}{\partial \alpha_2^2} = -\frac{3}{\alpha_2^4}\sum(x_i-\alpha_1)^2 + \frac{N}{\alpha_2^2} = -\frac{2N}{\alpha_2^{*2}} \text{ when } \alpha_1 = \alpha_1^*,$$

$$\frac{\partial^2 w}{\partial \alpha_1 \alpha_2} = -\frac{2}{\alpha_2^2}\sum(x_i-\alpha_1) = 0 \text{ when } \alpha_1 = \alpha_1^*,$$

$$\underline{H} = \begin{bmatrix} \frac{N}{\alpha_2^{*2}} & 0 \\ 0 & \frac{2N}{\alpha_2^{*2}} \end{bmatrix} \quad \text{and} \quad \underline{H}^{-1} = \begin{bmatrix} \frac{\alpha_2^{*2}}{N} & 0 \\ 0 & \frac{\alpha_2^{*2}}{2N} \end{bmatrix}$$

According to Eq. (11), the errors on α_1 and α_2 are the square roots of the diagonal elements of the error matrix, H^{-1}:

$$\Delta\alpha_1 = \frac{\alpha_2^*}{\sqrt{N}} \text{ and } \Delta\alpha_2 = \frac{\alpha_2^*}{\sqrt{2N}}$$ (this is sometimes called the error of the error).

We note that the error of the mean is 1/sqrt[*N*] σ where $\sigma = \alpha_2$ is the standard deviation. The error on the determination of σ is σ/sqrt[2*N*].

Correlated Errors

The matrix $V_{ij} \equiv \overline{(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)}$ is defined as the error matrix (also called the covariance matrix of α). In Eq. 11 we have shown that $\underline{V} = \underline{H}^{-1}$ where $H_{ij} = -\partial^2 w / (\partial\alpha_i \partial\alpha_j)$. The diagonal elements of $\underline{V}$ are the variances of the α's. If all the off-diagonal elements are zero, the errors in α are uncorrelated as in Example 2. In this case contours of constant *w* plotted in (α_1, α_2) space would be ellipses as shown in Fig. 2a. The errors in α_1 and α_2 would be the semi-major axes of the contour ellipse where *w* has dropped by ½ unit from its maximum-likelihood value. Only in the case of uncorrelated errors is the rms error $\Delta\alpha_j = (H_{jj})^{-½}$ and then there is no need to perform a matrix inversion.

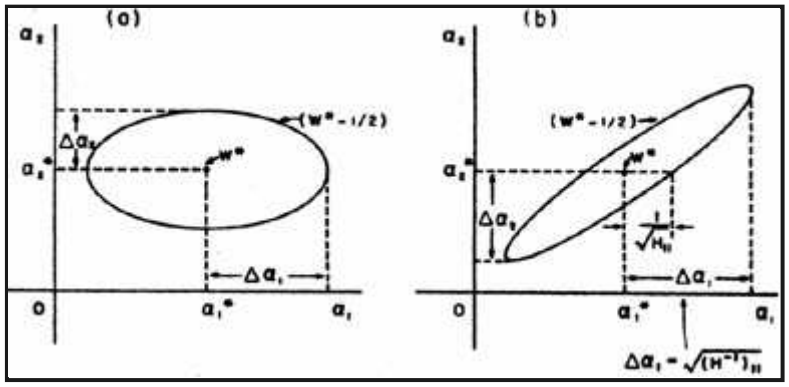

Figure 2. Contours of constant *w* as a function of α_1 and α_2. Maximum likelihood solution is at *w* = *w**. Errors in α_1 and α_2 are obtained from ellipse where *w* = (*w** - ½).
(a) Uncorrelated errors.
(b) Correlated errors. In either case $\Delta\alpha_1^2 = V_{11} = (H^{-1})_{11}$ and $\Delta\alpha_2^2 = V_{22} = (H^{-1})_{22}$. Note that it would be a serious mistake to use the ellipse "halfwidth" rather than the extremum for $\Delta\alpha$.

In the more common situation there will be one or more off-diagonal elements to $\underline{H}$ and the errors are correlated ($\underline{V}$ has off-diagonal elements). In this case (Fig. 2b) the contour ellipses are inclined to the α_1, α_2 axes. The rms spread of α_1 is still $\Delta\alpha_1 = \text{sqrt}[V_{11}]$, but it is the extreme limit of the ellipse projected on the α_1-axis. (The ellipse "halfwidth" axis is $(H_{11})^{-½}$ which is smaller.) In cases where Eq. 11 cannot be evaluated analytically, the α*'s can be found numerically and the errors in α can be found by Plotting the ellipsoid where *w* is 1/2 unit less than w^*. The extremums of this ellipsoid are the rms error in the α's. One should allow all the α_j to change freely and search for the maximum change in α_i which makes *w* = (w^* - ½). This maximum change in α_i, is the error in α_i and is sqrt[V_{11}].

8. PROPAGATION OF ERRORS: THE ERROR MATRIX

Consider the case in which a single physical quantity, *y*, is some function of the α's: $y = y(\alpha_1, ..., \alpha_M)$. The "best" value for *y* is then $y^* = y(\alpha_i^*)$. For example *y* could be the path radius of an electron circling in a uniform magnetic field where the measured quantities are $\alpha_1 = \tau$, the period of revolution, and $\alpha_2 = v$, the electron velocity. Our goal is to find the error in y given the errors in α. To first order in ($\alpha_i - \alpha_i^*$) we have

$$y - y^* = \sum \frac{\partial y}{\partial \alpha_a}(\alpha_a - \alpha_a^*),$$

$$\overline{(y - y^*)^2} = \sum_a \sum_b \frac{\partial y}{\partial \alpha_a}\frac{\partial y}{\partial \alpha_b}\overline{(\alpha_a - \alpha_a^*)(\alpha_b - \alpha_b^*)}, \qquad (12)$$

$$(\Delta y)_{\text{rms}} = \sqrt{\sum_a \sum_b \frac{\partial y}{\partial \alpha_a}\frac{\partial y}{\partial \alpha_b} V_{ab}}$$

A well-known special case of Eq. (12), which holds only when the variables are completely uncorrelated, is

$$(\Delta y)_{\rm rms} = \sqrt{\sum_a \left(\frac{\partial y}{\partial \alpha_a}\right)(\Delta\alpha_a)^2}.$$

In the example of orbit radius in terms of τ and v this becomes

$$\Delta R = \sqrt{\left(\frac{\partial R}{\partial \tau}\right)^2 (\Delta\tau)^2 + \left(\frac{\partial R}{\partial N}\right)^2 (\Delta v)^2} = \sqrt{\frac{v^2}{4\pi^2}(\Delta\tau)^2 + \frac{\tau^2}{4\pi^2}(\Delta v)^2}$$

in the case of uncorrelated errors. However, if $\overline{\Delta\tau\Delta v}$ is non-zero as one might expect, then Eq. (12) gives

$$\Delta R = \sqrt{\frac{v^2}{4\pi^2}(\Delta\tau)^2 + \frac{\tau^2}{4\pi^2}(\Delta v)^2 + 2\left(\frac{v}{2\pi}\right)\left(\frac{\tau}{2\pi}\right)\overline{\Delta\tau\Delta v}}$$

It is a common problem to be interested in M physical parameters, y_1, ..., y_M, which are known functions of the α_i. In fact the y_i can be thought of as a new set of α_i or a change of basis from α_i to y_i. If the error matrix of the α_i is known, then we have

$$\overline{(y_i - y_i^*)(y_j - y_j^*)} = \sum_a \sum_b \frac{\partial y_i}{\partial \alpha_a}\frac{\partial y_j}{\partial \alpha_b} H_{ab}^{-1}. \qquad (13)$$

In some such cases the ðy_i / ðα_a cannot be obtained directly, but the ðα_i / ðy_a are easily obtainable. Then

$$\frac{\partial y_i}{\partial \alpha_a} = (\underline{J}^{-1})_{ia}, \quad \text{where} \quad J_{ij} = \frac{\partial \alpha_i}{\partial y_j}.$$

Example 3

Suppose one wishes to use radius and acceleration to specify the circular orbit of an electron in a uniform magnetic field; i.e., $y_1 = r$ and $y_2 = a$. Suppose the original measured quantities are $\alpha_1 = \tau = (10 \pm 1)\mu$s and $\alpha_2 = v = (100 \pm 2)$ km/s. Also since the velocity measurement depended on the time measurement, there was a correlated error $\overline{\Delta\tau\Delta v} = 1.5 \times 10^{-3}$ m. Find r,**Δ**r, a, **Δ**a.

Since $r = v\tau$ / 2π = 0.159 m and $a = 2\pi v$ / $\tau = 6.28 \times 10^{10}$ m/s^2 we have $y_1 = \alpha_1\alpha_2$ / 2π and $y_2 = 2\pi\,\alpha_2$ / α_1. Then ðy_1 / ðα_1 = α_2 / 2π, ðy_1 / ðα_2 = α_1 / 2π, ðy_2 / ðα_1 = -2πα_2 / α_1^2, ðy_2 / ðα_2 = 2π / α_1 . The measurement errors specify the error matrix as

$$\underline{V} = \begin{bmatrix} 10^{12}\ \text{s}^2 & 1.5 \times 10^{-3}\ \text{m} \\ 1.5 \times 10^{-3}\ \text{m} & 4 \times 10^{6}\ \text{m2/s}^2 \end{bmatrix}$$

Eq. 13 gives

$$\begin{aligned}(\Delta y_1)^2 &= \left[\frac{\alpha_2}{2\pi}\right]^2 V_{11} + 2\left[\frac{\alpha_2}{2\pi}\right]\left[\frac{\alpha_1}{2\pi}\right]V_{12} + \left[\frac{\alpha_1}{2\pi}\right]^2 V_{22} \\ &= \frac{v^2}{4\pi^2}V_{11} + \frac{v\tau}{2\pi^2}V_{12} + \frac{\tau^2}{4\pi^2} = 3.39 \times 10^{-4}\ \text{m}^2\end{aligned}$$

Thus r = (0.159 ± 0.184) m

For y_2, Eq. 13 gives

$$(\Delta y_2)^2 = \left[-\frac{2\pi\alpha_2}{\alpha_1^2}\right]^2 V_{11} + 2\left[-\frac{2\pi\alpha_2}{\alpha_1^2}\right]\left[\frac{2\pi}{\alpha_1}\right] V_{12} + \left[\frac{2\pi}{\alpha_1}\right] V_{22} = 2.92 \times 10^{19}\ \frac{\mathrm{m}^2}{\mathrm{s}^4}$$

Thus $a = (6.28 \pm 0.54) \times 10^{10}$ m/s^2.

9. SYSTEMATIC ERRORS

"Systematic effects" is a general category which includes effects such as background, selection bias, scanning efficiency, energy resolution, angle resolution, variation of counter efficiency with beam position and energy, dead time, etc. The uncertainty in the estimation of such a systematic effect is called a "systematic error". Often such systematic effects and their errors are estimated by separate experiments designed for that specific purpose. In general, the maximum-likelihood method can be used in such an experiment to determine the systematic effect and its error. Then the systematic effect and its error are folded into the distribution function of the main experiment. Ideally, the two experiments can be treated as one joint experiment with an added parameter α_{M+1} to account for the systematic effect.

In some cases a systematic effect cannot be estimated apart from the main experiment. Example 2 can be made into such a case. Let us assume that among the beam of mono-energetic particles there is an unknown background of particles uniformly distributed in range. In this case the distribution function would be

$$f(\alpha_1, \alpha_2, \alpha_3; x) = \frac{1}{C}\left[\frac{1}{\sqrt{2\pi}\,\alpha_2}\exp[-(x-\alpha_1)^2/2\alpha_2^2] + \alpha_3\right]$$

where

$$C(\alpha_1, \alpha_2, \alpha_3) = \int_{x_{\min}}^{x_{\max}} f\, dx$$

The solution α_3* is simply related to the percentage of background. The systematic error is obtained using Eq. 11.

10. UNIQUENESS OF MAXIMUM-LIKELIHOOD SOLUTION

Usually it is a matter of taste what physical quantity is chosen as α. For example, in a lifetime experiment some workers would solve for the lifetime, τ*, while others would solve for λ*, where $\lambda = 1/\tau$. Some workers prefer to use momentum, and others energy, etc. Consider the case of two related physical parameters λ and α. The maximum-likelihood solution for α is obtained from the equation ðw / ðα = 0. The maximum-likelihood solution for λ is obtained from ðw / ðα = 0. But then we have

$$\frac{\partial w}{\partial \lambda} = \frac{\partial w}{\partial \alpha}\frac{\partial \alpha}{\partial \lambda} = 0, \quad \text{and} \quad \frac{\partial w}{\partial \alpha} = 0.$$

Thus the condition for the maximum-likelihood solution is unique and independent of the arbitrariness involved in choice of physical parameter. A lifetime result τ* would be related to the solution λ* by τ* = 1/λ*.

The basic shortcoming of the maximum-likelihood method is what to do about the prior probability of α. If the prior probability of α is $G(\alpha)$ and the likelihood function obtained for the experiment alone is $\mathcal{H}(\alpha)$, then the joint likelihood function is

$$\mathcal{L}(\alpha) = G(\alpha)\mathcal{H}(\alpha);$$
$$w = \ln G + \ln \mathcal{H}.$$
$$\frac{\partial w}{\partial \alpha} = \frac{\partial}{\partial \alpha}\ln G + \frac{\partial}{\partial \alpha}\ln \mathcal{H}.$$
$$\frac{\partial}{\partial \alpha}\ln \mathcal{H}(\alpha^*) = -\frac{\partial}{\partial \alpha}\ln G(\alpha^*)$$

give the maximum-likelihood solution. In the absence of any prior knowledge the term on the right-hand side is zero. In other words, the standard procedure in the absence of any prior information is to use a prior distribution in which all values of α are equally probable. Strictly speaking, it is impossible to know a "true" $G(\alpha)$, because it in turn must depend on its own prior probability. However, the above equation is useful when $G(\alpha)$ is the combined likelihood function of all previous experiments and $\mathcal{H}(\alpha)$ is the likelihood function of the experiment under consideration.

There is a class of problems in which one wishes to determine an unknown distribution in α, $G(\alpha)$, rather than a single value α. For example, one may wish to determine the momentum distribution of cosmic ray muons. Here one observes

$$\mathcal{L}(G) = \int G(\alpha)\mathcal{H}(\alpha; x)d\alpha$$

where $\mathcal{H}(\alpha; x)$ is known from the nature of the experiment and G(α) is the function to be determined. This type of problem is discussed in Reference 5.

11. CONFIDENCE INTERVALS AND THEIR ARBITRARINESS

So far we have worked only in terms of relative probabilities and rms values to give an idea of the accuracy of the determination $\alpha = \alpha^*$. One can also ask the question, What is the probability that α lies between two certain values such as α' and α''? This is called a confidence interval,

$$P(\alpha' < \alpha < \alpha'') = \int_{\alpha'}^{\alpha''} \mathcal{L}d\alpha \Big/ \int_{-\infty}^{\infty} \mathcal{L}d\alpha$$

Unfortunately such a probability depends on the arbitrary choice of what quantity is chosen for α. To show this consider the area under the tail of $\mathcal{L}(\alpha)$.

$$P(\alpha > \alpha') = \frac{\int_{\alpha'}^{\infty} \mathcal{L}d\alpha}{\int_{-\infty}^{\infty} \mathcal{L}d\alpha}$$

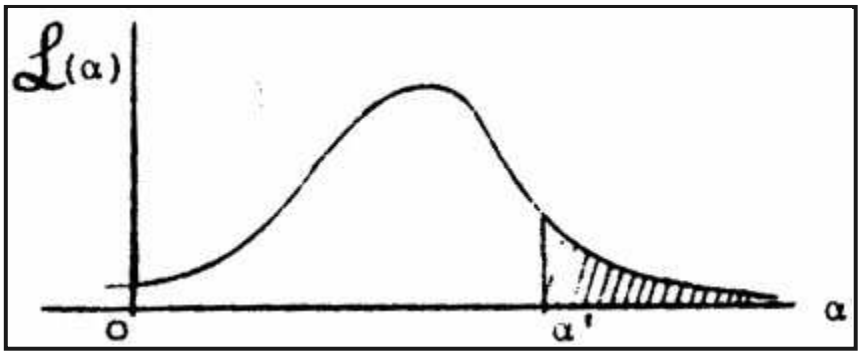

Figure 3. Shaded area is $P(\alpha > \alpha')$. (Sometimes called the confidence limit of α'.)

If $\lambda = \lambda(\alpha)$ had been chosen as the physical parameter instead, the same confidence interval is

$$P(\lambda > \lambda') = \frac{\int_{\lambda'}^{\infty} \mathcal{L} d\lambda}{\int_{-\infty}^{\infty} \mathcal{L} d\lambda} = \frac{\int_{\alpha'}^{\infty} \mathcal{L} \frac{\partial \lambda}{\partial \alpha} d\alpha}{\int_{-\infty}^{\infty} \mathcal{L} d\lambda} \neq P(\alpha > \alpha').$$

Thus, in general, the numerical value of a confidence interval depends on the choice of the physical parameter. This is also true to some extent in evaluating $\mathbf{\Delta}\alpha$. Only the maximum likelihood solution and the relative probabilities are unaffected by the choice of α. For Gaussian distributions, confidence intervals can be evaluated by using tables of the probability integral. Tables of cumulative binomial distributions and cumulative Poisson distributions are also available. Appendix V contains a plot of the cumulative Gaussian distribution.

12. BINOMIAL DISTRIBUTION

Here we are concerned with the case in which an event must be one of two classes, such as up or down, forward or back, positive or negative, etc. Let p be the probability for an event of Class 1. Then $(1 - p)$ is the probability for Class 2, and the joint probability for observing N_1 events in Class 1 out of N total events is

$$P(N_1, N) = \frac{N!}{N_1!\,(N - N_1)!}\, p^{N_1} (1 - p)^{N - N_1}. \quad \text{The binomial distribution} \qquad (14)$$

Note that $\sum_{j=1}^{N} p(j, N) = [p + (1 - p)]^N = 1$. The factorials correct for the fact that we are not interested in the order in which the events occurred. For a given experimental result of N_1 out of N events in Class 1, the likelihood function $\mathcal{L}(p)$ is then

$$\begin{aligned} \mathcal{L}(p) &= \frac{N!}{N_1!\,(N - N_1)!}\, p^{N_1} (1 - p)^{N - N_1} \\ w &= N_1 \ln p + (N - N_1) \ln(1 - p) + \text{const} \qquad (15) \\ \frac{\partial w}{\partial p} &= \frac{N_1}{p} - \frac{N - N_1}{1 - p}, \\ \frac{\partial^2 w}{\partial p^2} &= \frac{N_1}{p^2} - \frac{N - N_1}{(1 - p)^2}. \qquad (16) \end{aligned}$$

From Eq. (15) we have

$$p^* = \frac{N_1}{N} \qquad (17)$$

From (16) and (17):

$$\begin{aligned} \overline{(p - p^*)^2} &= \frac{1}{\frac{N_1}{p^{*2}} - \frac{N - N_1}{(1 - p^*)^2}}, \qquad (18) \\ \Delta p &= \frac{\sqrt{p^*(1 - p^*)}}{N} \end{aligned}$$

The results, Eqs. (17) and (18), also happen to be the same as those using direct probability. Then

$$N_1 = pN$$

and

$$\overline{(N_1 - N_1)^2} = Np(1-p).$$

Example 4

In Example 1 on the μ-e decay angular distribution we found that

$$\Delta\alpha \approx \sqrt{\frac{3}{N}}$$

is the error on the asymmetry parameter α. Suppose that the individual cosine, x_i, of each event is not known. In this problem all we know is the number up vs. the number down. What then is $\Delta\alpha$? Let p be the probability of a decay in the up hemisphere; then we have

$$p = \int_0^1 \frac{1+\alpha x}{2} dx = \frac{1+\frac{\alpha}{2}}{2}.$$

By Eq. (18),

$$\begin{aligned} \Delta\alpha &= 4\sqrt{\frac{p^*(1-p^*)}{N}}, \\ \Delta\alpha &= \sqrt{\frac{4}{N}(1-\frac{\alpha^{*2}}{4})} \end{aligned}$$

For small α this is $\Delta\alpha$ = sqrt[4 / N] as compared to sqrt[3 / N] when the full information is used.

13. POISSON DISTRIBUTION

A common type of problem which falls into this category is the determination of a cross section or a mean free path. For a mean free path λ, the probability of getting an event in an interval dx is dx / λ. Let $P(0, x)$ be the probability of getting no events in a length x. Then we have

$$\begin{aligned} dP(0,x) &= -P(0,x) \times \frac{dx}{\lambda}, \\ \ln P(0,x) &= -\frac{x}{\lambda} + \text{ const}, \\ P(0,x) &= e^{-x/\lambda} \ (\text{at } x=0,\ P(0,x)=1). \end{aligned} \qquad (19)$$

Let $P(N, x)$ be the probability of finding N events in a length x. An element of this probability is the joint probability of N events at dx_1, ..., dx_N times the probability of no events in the remaining length:

$$d^N P(N,x) = \prod_{i=1}^{N} \left[\frac{dx_i}{\lambda}\right] e^{-x/\lambda} \qquad (20)$$

The entire probability is obtained by integrating over the N-dimensional space. Note that the integral

$$\prod_{i=1}^{N} \int_0^x \frac{x_i}{\lambda} = \left[\frac{x}{\lambda}\right]^N$$

does the job except that the particular probability element in Eq. (20) is swept through *N*! times. Dividing by *N*! gives

$$P(N,x) = \frac{\left[\frac{x}{\lambda}\right]^N}{N!} e^{-x/\lambda}, \quad \begin{matrix}\text{the Poisson}\\ \text{distribution}\end{matrix} \qquad (21)$$

As a check, note

$$\sum_{j=1}^{\infty} P(j,x) = e^{-x/\lambda}\left(\sum \frac{(x/\lambda)^j}{j!}\right) = e^{-x/\lambda}\left(e^{x/\lambda}\right) = 1.$$

$$\overline{N} = \sum_{N=1}^{\infty} N \frac{(x/\lambda)^N}{N!} e^{-x/\lambda} = \frac{x}{\lambda}\left[\sum_{N=1}^{\infty} \frac{\left[\frac{x}{\lambda}\right]^{N-1}}{N-1}\right] e^{-x/\lambda} = \frac{x}{\lambda}.$$

Likewise it can be shown that $\overline{(N-\overline{N})^2} = \bar{N}$. Equation (21) is often expressed in terms of $\bar{N}$:

$$P(N,\overline{N}) = \frac{\overline{N}^N}{N!} e^{-\overline{N}}, \quad \begin{matrix}\text{the Poisson}\\ \text{distribution}\end{matrix} \qquad (22)$$

This form is useful in analyzing counting experiments. Then the "true" counting rate is $\bar{N}$.

We now consider the case in which, in a certain experiment, *N* events were observed. The problem is to determine the maximum-likelihood solution for $\alpha \equiv \bar{N}$ and its error:

$$\begin{aligned} \mathcal{L}(\alpha) &= \frac{\alpha^N}{N!} e^{-\alpha} \\ w &= N \ln \alpha - \alpha - \ln N!, \\ \frac{\partial w}{\partial \alpha} &= \frac{N}{\alpha} - 1. \\ \frac{\partial^2 w}{\partial \alpha^2} &= -\frac{N}{\alpha^2}. \end{aligned}$$

Thus we have

$$\alpha^* = N$$

and by Eq. (7),

$$\Delta\alpha = \frac{\alpha}{\sqrt{N}}$$

In a cross-section determination, we have $\alpha = \rho x \sigma$, where ρ is the number of target nuclei per cm^3 and *x* is the total path length. Then

$$\sigma^* = \frac{N}{\rho x} \quad \text{and} \quad \frac{\Delta\sigma}{\sigma^*} = \frac{1}{\sqrt{N}}$$

In conclusion we note that $\alpha \neq \bar{\alpha}$:

$$\bar{\alpha} = \frac{\int_0^\infty \alpha \mathcal{L}(\alpha) d\alpha}{\int \mathcal{L}(\alpha) d\alpha} = \frac{\int_0^\infty \alpha^{N+1} e^{-\alpha} d\alpha}{\int_0^\infty \alpha^N e^{-\alpha} d\alpha} = \frac{(N+1)!}{N!} = N + 1.$$

14. GENERALIZED MAXIMUM-LIKELIHOOD METHOD

So far we have always worked with the standard maximum-likelihood formalism, whereby the distribution functions are always normalized to unity. Fermi has pointed out that the normalization requirement is not necessary so long as the basic principle is observed: namely, that if one correctly writes down the probability of getting his experimental result, then this likelihood function gives the relative probabilities of the parameters in question. The only requirement is that the probability of getting a particular result be correctly written. We shall now consider the general case in which the probability of getting an event in dx is $F(x)dx$, and

$$\int_{x_{\min}}^{x_{\max}} F dx \equiv \overline{N}(\alpha)$$

is the average number of events one would get if the same experiment were repeated many times. According to Eq. (19), the probability of getting no events in a small finite interval Δx is

$$\exp(-\int_x^{x+\Delta x} F dx).$$

The probability of getting no events in the entire interval $x_{\min} < x < x_{\max}$ is the product of such exponentials or

$$\exp(-\int_{x_{\min}}^{x_{\max}} F dx) = e^{-\overline{N}}$$

The element of probability for a particular experimental result of N events at $x = x_1, \ldots, x_N$ is then

$$d^N p = e^{-\overline{N}} \prod_{i=1}^N F(x_i) dx_i.$$

Thus we have

$$\mathcal{L}(\alpha) = e^{-\overline{N}(\alpha)} \prod_{i=1}^N F(\alpha; x_i)$$

and

$$w(\alpha) = \sum_{i=1}^N \ln F(\alpha; x_i) - \int_{x_{\min}}^{x_{\max}} F(\alpha; x) dx.$$

The solutions $\alpha_i = \alpha_i^*$ are still given by the M simultaneous equations:

$$\frac{\partial w}{\partial \alpha_i} = 0.$$

The errors are still given by

$$\overline{(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)} = (\underline{H}^{-1})_{ij}.$$

where

$$H_{ij} = -\frac{\partial^2 w}{\partial \alpha_i \partial \alpha_j}$$

The only change is that *N* no longer appears explicitly in the formula

$$-\frac{\overline{\partial^2 w}}{\partial \alpha_i \partial \alpha_j} = \int \frac{1}{F}\left[\frac{\partial F}{\partial \alpha_i}\right]\left[\frac{\partial F}{\partial \alpha_i}\right] dx.$$

A derivation similar to that used for Eq. (8) shows that *N* is already taken care of in the integration over *F*(*x*).

In a private communication, George Backus has proven, using direct probability, that the Maximum-Likelihood Theorem also holds for this generalized maximum-likelihood method and that in the limit of large *N* there is no method of estimation that is more accurate. Also see Sect. 9.8 of Ref. 6.

In the absence of the generalized maximum-likelihood method our procedure would have been to normalize $F(\alpha; x)$ to unity by using

$$f(\alpha; x) = \frac{F(\alpha; x)}{\int F dx}.$$

For example, consider the sample containing just two radioactive species, of lifetimes α_1 and α_2. Let α_3 and α_4 be the two initial decay rates. Then we have

$$F(\alpha_i; x) = \alpha_3 \, \mathrm{e}^{-x/\alpha_1} + \alpha_4 \, \mathrm{e}^{-x/\alpha_2},$$

where *x* is the time. The standard method would then be to use

$$f(\alpha; x) = \frac{\mathrm{e}^{-x/\alpha_1} + \alpha_5 \, \mathrm{e}^{-x/\alpha_2}}{\alpha_1 + \alpha_5 \alpha_2},$$

which is normalized to one. Note that the four original parameters have been reduced to three by using $\alpha_5 \equiv \alpha_4 / \alpha_3$. Then α_3 and α_4 would be found by using the auxiliary equation

$$\int_0^\infty F dx = N,$$

the total number of counts. In this standard procedure the equation

$$\overline{N}(\alpha_i) = N,$$

must always hold. However, in the generalized maximum-likelihood method these two quantities are not necessarily equal. Thus the generalized maximum-likelihood method will give a different solution for the α_i, which should, in principle, be better.

Another example is that the best value for a cross section σ is not obtained by the usual procedure of setting $\rho\,\sigma L = N$ (the number of events in a path length L). The fact that one has additional prior information such as the shape of the angular distribution enables one to do a somewhat better job of calculating the cross section.

15. THE LEAST-SQUARES METHOD

Until now we have been discussing the situation in which the experimental result is N events giving precise values $x_1, \ldots, x_N$ where the x_i may or may not, as the case may be, be all different.

From now on we shall confine our attention to the case of p measurements (not p events) at the points $x_1, \ldots, x_p$. The experimental results are $(y_1 \pm \sigma_1), \ldots, (y_p \pm \sigma_p)$. One such type of experiment is where each measurement consists of N_i events. Then $y_i = N_i$ and is Poisson-distributed with $\sigma_i = \mathrm{sqrt}[N_i]$. In this case the likelihood function is

$$\mathcal{L} = \prod_{i=1}^{p} \frac{[\overline{y}(x_i)]^{N_i}}{N_i!}\, e^{-\overline{y}(x_i)}$$

and

$$w = \sum_{i=1}^{p} N_i \ln \overline{y}(x_i) - \sum_{i=1}^{p} \overline{y}(x_i) + \text{ const.}$$

We use the notation $\bar{y}(\alpha_i; x)$ for the curve that is to be fitted to the experimental points. The best-fit curve corresponds to $\alpha_i = \alpha_i^*$. In this case of Poisson-distributed points, the solutions are obtained from the M simultaneous equations

$$\sum_{a=1}^{p} \frac{\partial \overline{y}(x_a)}{\partial \alpha_i} = \sum_{a=1}^{p} \frac{N_a}{\overline{y}(x_a)} \frac{\partial \bar{(y)}(x_a)}{\partial \alpha_i}$$

If all the $N_i >> 1$, then it is a good approximation to assume each y_i is Gaussian-distributed with standard deviation σ_i. (It is better to use $\bar{N}_i$ rather than N_i for σ_i^2 where $\bar{N}_i$ can be obtained by integrating $\bar{y}(x)$ over the ith interval.) Then one can use the famous least squares method.

The remainder of this section is devoted to the case in which y_i are Gaussian-distributed with standard deviations σ_i. See Fig. 4. We shall now see that the least-squares method is mathematically equivalent to the maximum likelihood method. In this Gaussian case the likelihood function is

$$\mathcal{L} = \prod_{a=1}^{p} \frac{1}{\sqrt{2\pi}\,\sigma_a} \exp\{-[y_a - \overline{y}(x_a)]^2/2\sigma_a^2\}$$
$$w(\alpha) = -\frac{1}{2} S(\alpha) - \sum_{a=1}^{p} \ln \sqrt{2\pi}\,\sigma_a. \qquad (23)$$

where

$$S(\alpha) \equiv \sum_{a=1}^{p} \frac{[y_a - \overline{y}(x_a)]^2}{\sigma_a^2} \qquad (24)$$

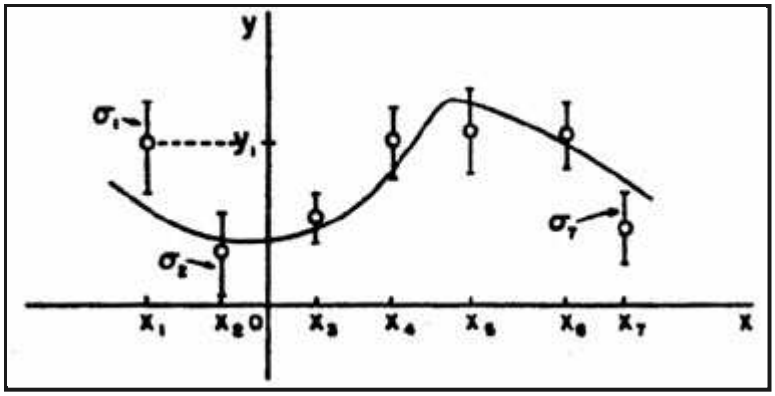

Figure 4. $\bar{y}(x)$ is a function of known shape to be fitted to the 7 experimental points.

The solutions $\alpha_i = \alpha_i^*$ are given by minimizing $S(\alpha)$ (maximizing w):

$$\frac{\partial S(\alpha)}{\partial \alpha_i} = 0. \qquad (25)$$

This minimum value of S is called S^*, the least squares sum. The values of α_i which minimize are called the least-squares solutions. Thus the maximum-likelihood and least-squares solutions are identical. According to Eq. (11), the least-squares errors are

$$\overline{(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)} = (\underline{H}^{-1})_{ij}, \quad \text{where} \quad H_{ij} = \frac{1}{2}\frac{\partial^2 S}{\partial \alpha_i \partial \alpha_j}.$$

Let us consider the special case in which $\bar{y}(\alpha_i; x)$ is linear in the α_i:

$$\overline{y}(\alpha_i; x) = \sum_{a=1}^{M} \alpha_a f_a(x).$$

(Do not confuse this $f(x)$ with the $f(x)$ on page 2.)

Then

$$\frac{\partial S}{\partial \alpha_i} = -2 \sum_{a=1}^{p} \left[\frac{y_a - \sum_{b=1}^{M} \alpha_b f_b(x_a)}{\sigma_a^2} \right] f_i(x_a). \qquad (26)$$

Differentiating with respect to α_j gives

$$H_{ij} = \sum_{a=1}^{p} \frac{f_i(x_a) f_j(x_a)}{\sigma_a^2} \qquad (27)$$

Define

$$U_i \equiv \sum_{a=1}^{p} \frac{y_a f_i(x_a)}{\sigma_a^2} \qquad (28)$$

Then

$$\frac{\partial S}{\partial \alpha_i} = -2\left[U_i - \sum_{b=1}^{M} \alpha_b H_{bi}\right].$$

In matrix notation the M simultaneous equations giving the least-squares solution are

$$\begin{aligned} 0 &= \underline{u} - \underline{\alpha}^* \cdot \underline{H}, \\ \underline{\alpha}^* &= \underline{u} \cdot \underline{H}^{-1} \end{aligned} \qquad (29)$$

is the solution for the α*'s. The errors in α are obtained using Eq. 11. To summarize:

$$\text{If } \overline{y}(\alpha; x) = \sum_{a=1}^{M} \alpha_a f_a(x)$$

$$\alpha_i^* = \sum_{a=1}^{M}\sum_{b=1}^{p} \frac{y_b f_a(x_b)}{\alpha_b^2}(\underline{H}^{-1})_{ai}, \qquad (30)$$

$$\overline{(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)} = \underline{H}_{ij}^{-1} \quad \text{where} \quad H_{ij} \equiv \sum_{a=1}^{p} \frac{f_i(x_a) f_j(x_a)}{\sigma_a^2}$$

Equation (30) is the complete procedure for calculating the least squares solutions and their errors. Note that even though this procedure is called curve-fitting it is never necessary to plot any curves. Quite often the complete experiment may be a combination of several experiments in which several different curves (all functions of the α_i) may be jointly fitted. Then the S-value is the sum over all the points on all the curves. Note that since $w(\alpha^*)$ decreases by ½ unit when one of the α_j has the value ($\alpha_i^* \pm \Delta\alpha_j$), the *S*-value must increase by one unit. That is,

$$S(\alpha_1^*, \ldots, \alpha_j \pm \Delta\alpha_j, \ldots, \alpha_M) = S^* + 1.$$

Example 5 Linear regression with equal errors

$\overline{y}(x)$ is known to be of the form $\overline{y}(x) = \alpha_1 + \alpha_2 x$. There are *p* experimental measurements ($y_j \pm \sigma$).Using Eq. (30) we have

$$f_1 = 1, \; f_2 = x,$$

$$\underline{H} = \begin{bmatrix} \frac{p}{\sigma^2} & \frac{\sum x_a}{\sigma^2} \\ \sum \frac{x_a}{\sigma^2} & \frac{\sum x_a^2}{\sigma^2} \end{bmatrix}$$

$$\underline{H}^{-1} = \frac{\sigma^2}{p\sum x_a^2 - (\sum x_a)^2} \begin{bmatrix} \sum x_a^2 & -\sum x_a \\ -\sum x_a & p \end{bmatrix}$$

$$\alpha_1^* = \frac{\sum y_a \sum x_a^2 - \sum x_a \sum (x_a y_a)}{p\sum x_a^2 - (\sum x_a)^2}$$

$$\alpha_2^* = \frac{p\sum (x_a y_a) - \sum x_a \sum y_a}{p\sum x_a^2 - (\sum x_a)^2}$$

These are the linear regression formulas which are programmed into many pocket calculators. They should not be used in those cases where the σ_i are not all the same. If the σ_i are all equal, the errors

$$(\Delta\alpha_1)^2 = (H^{-1})_{11}$$

or

$$\Delta\alpha_1 = \sigma\sqrt{\frac{\sum x_a^2}{p\sum x_a^2 - (\sum x_a)^2}}$$

$$\Delta\alpha_2 = \sqrt{(H^{-1})_{22}} = \sigma\sqrt{\frac{p}{p\sum x_a^2 - (\sum x_a)^2}}$$

Example 6 Quadratic regression with unequal errors

The curve to be fitted is known to be a parabola. There are four experimental points at x = - 0.6, - 0.2, 0.2, and 0.6. The experimental results are 5 ± 2, 3 ± 1, 5 ± 1, and 8 ± 2. Find the best-fit curve.

$$\overline{y}(x) = \alpha_1 + \alpha_2 x + \alpha_3 x^2$$

$$f_1 = 1, \quad f_2 = x, \quad f_3 = x^2,$$

$$H_{11} = \sum_a \frac{1}{\sigma_a^2}, \quad H_{22} = \sum_a \frac{x_a^2}{\sigma_a^2}, \quad H_{33} = \sum_a \frac{x_a^4}{\sigma_a^2},$$

$$H_{12} = \sum \frac{x_a}{\sigma_a^2}, \quad H_{13} = \sum \frac{x_a^2}{\sigma_a^2} = H_{22}, \quad H_{23} = \sum \frac{x_a^3}{\sigma_a^2}$$

$$\underline{H} = \begin{bmatrix} 2.5 & 0 & 0.26 \\ 0 & 0.26 & 0 \\ 0.26 & 0 & 0.068 \end{bmatrix} \quad \underline{H}^{-1} = \begin{bmatrix} 0.664 & 0 & -2.54 \\ 0 & 3.847 & 0 \\ -2.54 & 0 & 24.418 \end{bmatrix} \equiv \underline{V}$$

$$\underline{u} = (11.25 \quad 0.85 \quad 1.49)$$

$$\alpha_1^* = 3.685, \quad \Delta\alpha_1 = 0.815, \quad \overline{\Delta\alpha_1\Delta\alpha_2} = 0$$
$$\alpha_2^* = 3.27, \quad \Delta\alpha_2 = 1.96, \quad \overline{\Delta\alpha_1\Delta\alpha_3} = -2.54$$
$$\alpha_3^* = 7.808, \quad \Delta\alpha_3 = 4.94.$$

$\bar{y}(x)$ = (3.685 ± 0.815) + (3.27 ± 1.96)x + (7.808 ± 4.94)x^2 is the best fit curve. This is shown with the experimental points in Fig. 5.

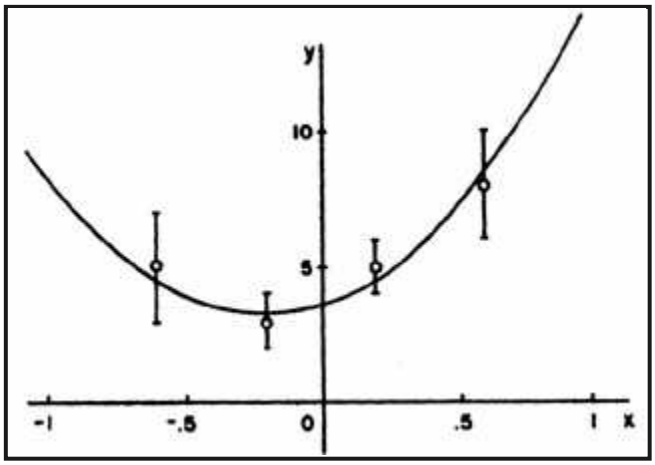

Figure 5. This parabola is the least squares fit to the 4 experimental points in Example 6.

Example 7

In example 6 what is the best estimate of y at $x = 1$? What is the error of this estimate?

Solution: Putting $x = 1$ into the above equation gives

$$y = 3.685 + 3.27 + 7.808 = 14.763.$$

Δy is obtained using Eq. 12.

$$\begin{aligned} \Delta y &= \sqrt{f_1^2 V_{11} + f_2^2 V_{22} + f_3^2 V_{33} + 2 f_1 f_2 V_{12} + 2 f_1 f_3 V_{13} + 2 f_2 f_3 V_{23}} \\ &= \sqrt{0.664 + x^2(3.847) + x^4(24.418) + 0 + 2x^2(-2.54) + 0} \end{aligned}$$

Setting $x = 1$ gives

$$\Delta y = 5.137$$

So at $x = 1$, $y = 14.763 \pm 5.137$.

Least Squares When the y_i are Not Independent

Let

$$V_{ij} = \overline{(y_i - \overline{y}_i)(y_j - \overline{y}_j)}$$

be the error matrix-of the y measurements. Now we shall treat the more general case where the off diagonal elements need not be zero; i.e., the quantities y_i are not independent. We see immediately from Eq. 11a that the log likelihood function is

$$w = -\frac{1}{2}(\underline{y - \overline{y}}) \cdot \underline{V}^{-1} \cdot \underline{(y - \overline{y})}^T + \text{const.}$$

The maximum likelihood solution is found by minimizing

$$S = (\underline{y - \overline{y}}) \cdot \underline{V}^{-1} \cdot \underline{(y - \overline{y})}^T$$

where $V_{ij} = \overline{(y_i - \overline{y}_i)(y_j - \overline{y}_j)}$ Generalized least squares sum

16. GOODNESS OF FIT, THE χ^2 DISTRIBUTION

The numerical value of the likelihood function at $\mathcal{L}(\alpha^*)$ can, in principle, be used as a check on whether one is using the correct type of function for $f(\alpha; x)$. If one is using the wrong f, the likelihood function will be lower in height and of greater width. In principle, one can calculate, using direct probability, the distribution of $\mathcal{L}(\alpha^*)$ assuming a particular true $f(\alpha_0, x)$. Then the probability of getting an $\mathcal{L}(\alpha^*)$ smaller than the value observed would be a useful indication of whether the wrong type of function for f had been used. If for a particular experiment one got the answer that there was one chance in 10^4 of getting such a low value of $\mathcal{L}(\alpha^*)$, one would seriously question either the experiment or the function $f(\alpha;x)$ that was used.

In practice, the determination of the distribution of $\mathcal{L}(\alpha^*)$ is usually an impossibly difficult numerical integration in N-dimensional space. However, in the special case of the least-square problem, the integration limits turn out to be the radius vector in p-dimensional space. In this case we use the distribution of $S(\alpha^*)$ rather than of $\mathcal{L}(\alpha^*)$. We shall first consider the

distribution of $S(\alpha_0)$. According to Eqs. (23) and (24) the probability element is

$$d^p P \propto \exp[-S/2] d^p y_i.$$

Note that $S = \rho^2$, where ρ is the magnitude of the radius vector in p-dimensional space. The volume of a p-dimensional sphere is $U \propto \rho_p$. The volume element in this space is then

$$d^p y_i \propto \rho^{p-1} d\rho \propto S^{(p-1)/2} S^{-\frac{1}{2}} dS.$$

Thus

$$dP(S) \propto S^{(p/2)-1}\, e^{(-S/2)} dS.$$

The normalization is obtained by integrating from $S = 0$ to $S = \infty$.

$$dP(S_0) = \frac{1}{2^{p/2}\Gamma(p/2)} S_0^{(p/2)-1}\, e^{-S_0/2} dS_0 \qquad (30a)$$

where $S \equiv S(\alpha_0)$.

This distribution is the well-known χ^2 distribution with p degrees of freedom. χ^2 tables of

$$\int_{S_0}^{\infty} dP(S)$$

for several degrees of freedom are commonly available - see Appendix V for plots of the above integral.

From the definition of S (Eq. (24)) it is obvious that $\overline{S}_0 = p$. One can show, using Eq. (29) that $\overline{(S_0 - \overline{S}_0)^2} = 2p$. Hence, one should be suspicious if his experimental result gives an S-value much greater than

$$(p + \sqrt{2p}).$$

Usually α is not known. In such a case one is interested in the distribution of

$$S^* \equiv S(\alpha^*).$$

Fortunately, this distribution is also quite simple. It is merely the χ^2 distribution of (p - M) degrees of freedom, where p is the number of experimental points, and M is the number of parameters solved for. Thus we haved

$$\begin{aligned} dP(S^*) &= \chi^2 \text{ distribution for } (p - M) \text{ degrees of freedom} \\ \overline{S^*} &= (p - M) \text{ and } \Delta S^* = \sqrt{(S^* - \overline{S^*})} = \sqrt{2(p - M)} \end{aligned} \qquad (31)$$

Since the derivation of Eq. (31) is somewhat lengthy, it is given in Appendix II.

Example 8

Determine the χ^2 probability of the solution to Example 6.

$$S^* = \left[\frac{5 - \overline{y}(-0.6)}{2}\right]^2 + \left[\frac{3 - \overline{y}(-0.2)}{1}\right]^2 + \left[\frac{5 - \overline{y}(0.2)}{1}\right]^2 + \left[\frac{8 - \overline{y}(0.6)}{2}\right]^2$$

$$S^* = 0.674 \text{ compared to } \overline{S^*} = 4 - 3 = 1.$$

According to the χ^2 table for one degree of freedom the probability of getting $S^* > 0.674$ is 0.41. Thus the experimental data are quite consistent with the assumed theoretical shape of

$$\overline{y} = \alpha_1 + \alpha_2 x + \alpha_3 x^2.$$

Example 9 Combining Experiments

Two different laboratories have measured the lifetime of the K_1^0 to be $(1.00 \pm 0.01) \times 10^{-10}$ sec and $(1.04 \pm 0.02) \times 10^{10}$ sec respectively. Are these results really inconsistent?

According to Eq. (6) the weighted mean is $\alpha^* = 1.008 \times 10^{-10}$ sec. (This is also the least squares solution for τ_{KO}.

Thus

$$S^* = \left[\frac{1.00 - 1.008}{0.01}\right]^2 + \left[\frac{1.04 - 1.008}{0.02}\right]^2 = 3.2 \quad \overline{S} = 2 - 1 = 1$$

According to the χ^2 table for one degree of freedom, the probability of getting $S^* > 3.2$ is 0.074. Therefore, according to statistics, two measurements of the same quantity should be at least this far apart 7.4% of the time.

APPENDIX I: PREDICTION OF LIKELIHOOD RATIOS

An important job for a physicist who plans new experiments is to estimate beforehand just how many events will be needed to "prove" a certain hypothesis. The usual procedure is to calculate the average logarithm of the likelihood ratio. The average logarithm is better behaved mathematically than the average of the ratio itself. We have

$$\overline{\log \mathcal{R}} = N \int \log \frac{f_A}{f_B} f_A(x) dx, \text{ assuming A is true} \qquad (32)$$

or

$$\overline{\log \mathcal{R}} = N \int \log \frac{f_A}{f_B} f_B(x) dx, \text{ assuming B is true}$$

Consider the example (given in Section 3) of the K^+ meson. We believe spin zero is true, and we wish to establish betting odds of 10^4 to 1 against spin 1. How many events will be needed for this? In this case Eq. (32) gives

$$\log 10^4 = 4 = \int_0^1 \log(\frac{1}{2x}) dx = -N \int_0^1 \log(2x) dx,$$

$$N = 30$$

Thus about 30 events would be needed on the average. However, if one is lucky, one might not need so many events. Consider the extreme case of just one event with $x = 0$: $\mathcal{R}$ would then be infinite and this one single event would be complete proof in itself that the K^+ is spin zero. The fluctuation (rms spread) of log $\mathcal{L}$ for a given N is

$$\overline{(\log \mathcal{R} - \overline{\log \mathcal{R}})^2} = N\left[\int (\log \frac{f_A}{f_B})^2 f_A dx - (\int \log \frac{f_A}{f_B} f_A dx)^2\right].$$

APPENDIX II: DISTRIBUTION OF THE LEAST-SQUARES SUM

We shall define the vector $Z_i \equiv y_i / \sigma_i$ and the matrix $F_{ij} \equiv f_j(x_i) / \sigma_i$.

Note that $\underline{H} = \underline{F}^T \cdot \underline{F}$ by Eq. (27),

$$\underline{Z} \cdot \underline{F} = \underline{\alpha}^* \cdot \underline{H} \text{ by Eq. (28) and (29).} \qquad (33)$$

Then

$$\underline{\alpha}^* = \underline{Z} \cdot \underline{F} \cdot \underline{H}^{-1}. \qquad (34)$$

$$S_0 = \sum_{a=1}^{p} \sum_{b=1}^{M} [(Z_a - \alpha_b^* F_{ab}) + (\alpha_b^* - a_b) F_{ab}]^2$$

where the unstarred α is used for α_0.

$$S_0 = \sum_a^p \sum_b^M \left[\frac{y_a}{a} - \frac{\alpha_b^* f_b(x_a)}{\sigma_a}\right]^2 + 2(\underline{Z} - \underline{\alpha}^* \cdot \underline{F}^T)\underline{F}(\underline{\alpha^* - \alpha})^T + (\underline{\alpha}^* - \underline{\alpha})\underline{F}^T \cdot \underline{F}(\underline{\alpha^* - \alpha})^T,$$

$$S_0 = S^* + 2(\underline{Z} \cdot \underline{F} - \alpha^* \cdot \underline{F}^T \underline{F})(\underline{\alpha^* - \alpha})^T + (\underline{Z} \cdot \underline{F} \cdot \underline{H}^{-1} - \underline{\alpha H H^{-1}})\underline{H}(\underline{H}^{-1}\underline{F}^T \underline{Z}^T - \underline{H}^{-1}\underline{H\alpha}^T)$$

using Eq. (34). The second term on the right is zero because of Eq. (33).

$$S^* = S_0 - (\underline{Z} \cdot \underline{F} - \underline{\alpha F^T F})\underline{H}^{-1}\underline{H H^{-1}}(\underline{F}^T \underline{Z}^T - \underline{F}^T \underline{F\alpha}^T),$$

$$S^* = (\underline{Z} - \overline{\underline{Z}})(\underline{1} - \underline{Q})(\underline{Z} - \overline{\underline{Z}})^T \text{ where } \underline{\alpha} \cdot \underline{F}^T = \overline{\underline{Z}} \text{ and}$$

$$\underline{Q} \equiv \underline{F H}^{-1} \underline{F}^T. \qquad (34)$$

Note that

$$\underline{Q}^2 = (\underline{F H}^{-1} \underline{F}^T)(\underline{F H}^{-1} \underline{F}^T) = \underline{F H}^{-1} \underline{F}^T = \underline{Q}$$

If q_i is an eigenvalue of $\underline{Q}$, it must be equal q_i^2, an eigenvalue of $\underline{Q}^2$. Thus $q_i = 0$ or 1. The trace of $\underline{Q}$ is

$$\text{Tr } \underline{Q} = \sum_{a,b,c} F_{ab} H_{bc}^{-1} F_{ca}^T = \sum_{b,c} H_{cb} H_{bc}^{-1} = \text{Tr } \underline{I} = M.$$

Since the trace of a matrix is invariant under a unitary transformation, the trace always equals the sum of the eigenvalues of the matrix. Therefore M of the eigenvalues of $\underline{Q}$ are one, and (p - M) are zero. Let U be the unitary matrix which diagonalizes $\underline{Q}$ (and also ($\underline{1}$ - $\underline{Q}$)). According to Eq. (35),

$$S^* = \underline{\eta} \cdot \underline{U}(\underline{1} - \underline{Q})\underline{U}^{-1} \cdot \underline{\eta}, \text{ where } \underline{\eta} \equiv (\underline{Z} - \overline{\underline{Z}}) \cdot \underline{U}^{-1},$$

$$S^* = \sum_{a=1}^{p} m_a \eta_a^2 \text{ where } m_a \text{ are the eigenvalues of } (\underline{1} - \underline{Q}).$$

$$S^* = \sum_{a=1}^{p-M} \eta_a^2 \text{ since the M nonzero eigenvalues of } \underline{Q} \text{ cancel out M of the eigenvalues of } \underline{1}.$$

Thus

$$dP(S^*) \propto e^{-S^*/2}\, d^{(p-M)}\, \eta_a,$$

where S^* is the square of the radius vector in $(p - M)$-dimensional space. By definition (see Section 16) this is the χ^2 distribution with $(p - M)$ degrees of freedom.

APPENDIX III. LEAST SQUARES WITH ERRORS IN BOTH VARIABLES

Experiments in physics designed to determine parameters in the functional relationship between quantities x and y involve a series of measurements of x and the corresponding y. In many cases not only are there measurement errors δy_i for each y_j, but also measurement errors δx_j for each x_j. Most physicists treat the problem as if all the $\delta x_j = 0$ using the standard least squares method. Such a procedure loses accuracy in the determination of the unknown parameters contained in the function $y = f(x)$ and it gives estimates of errors which are smaller than the true errors.

The standard least squares method of Section 15 should be used only when all the $\delta x_j << \delta y_i$. Otherwise one must replace the weighting factors $1 / \sigma_i^2$ in Eq. (24) with $(\delta_j)^{-2}$ where

$$\delta_j^2 \equiv \left[\frac{\partial f}{\partial x}\right]_j^2 [\delta x_j]^2 + [\delta y_j]^2 \quad (36)$$

Eq. (24) then becomes

$$S = \sum_{j=1}^{n} \left[\frac{y_i - f(x_j)}{\delta_j}\right]^2 \quad (37)$$

A proof is given in Ref. 7.

We see that the standard least squares computer programs may still be used. In the case where $y = \alpha_1 + \alpha_2 x$ one may use what are called linear regression programs, and where y is a polynomial in x one may use multiple polynomial regression programs. The usual procedure is to guess starting values for $\partial f / \partial x$ and then solve for the parameters α_j* using Eq. (30) with σ_j replaced by σ_j. Then new $[\partial f / \partial x]_j$ can be evaluated and the procedure repeated. Usually only two iterations are necessary. The effective variance method is exact in the limit that $\partial f / \partial x$ is constant over the region δx_j. This means it is always exact for linear regressions.

Next Contents Previous

APPENDIX IV. NUMERICAL METHODS FOR MAXIMUM LIKELIHOOD AND LEAST SQUARES SOLUTIONS

In many cases the likelihood function is not analytical or else, if analytical, the procedure for finding the α_j* and the errors is too cumbersome and time consuming compared to numerical methods using modern computers.

For reasons of clarity we shall first discuss an inefficient, cumbersome method called the grid method. After such an

introduction we shall be equipped to go on to a more efficient and practical method called the method of steepest descent.

The grid method

If there are M parameters $\alpha_1, \ldots, \alpha_M$ to be determined one could in principle map out a fine grid in M-dimensional space evaluating $w(\alpha)$ (or $S(\alpha)$) at each point. The maximum value obtained for w is the maximum likelihood solution w^*. One could then map out contour surfaces of $w = (w^* - ½), (w^* - 1)$, etc. This is illustrated for $M = 2$ in Fig. 6.

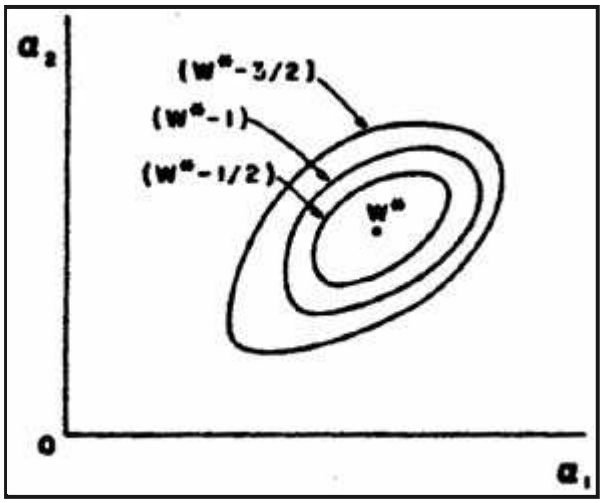

Figure 6. Contours of fixed w enclosing the max. likelihood solution w^*.

In the case of good statistics the contours would be small ellipsoids. Fig. 7 illustrates a case of poor statistics.

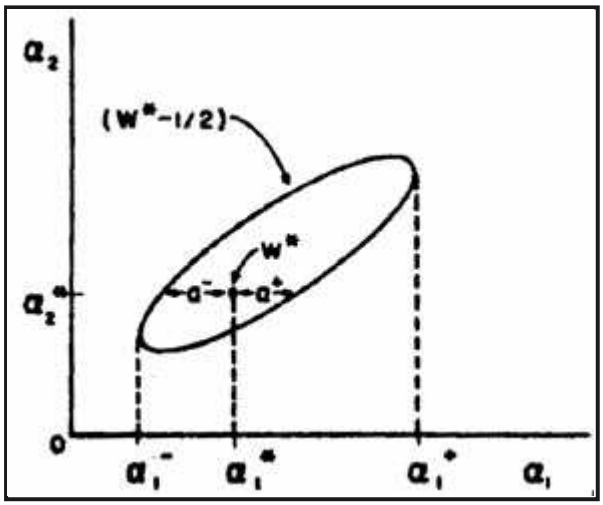

Figure 7. A poor statistics case of Fig. 6.

Here it is better to present the $(w^* - ½)$ contour surface (or the $(S^* + 1)$ surface) than to try to quote errors on α. If one is to quote errors it should be in the form $\alpha_1^- < \alpha_1 < \alpha_1^+$ where α_1^- and α_1^+ are the extreme excursions the surface makes in α_1 (see Fig. 7). It could be a serious mistake to quote a^- or a^+ as the errors in α_1.

In the case of good statistics the second derivatives $\partial^2 w / \partial\alpha_a \partial\alpha_b = - H_{ab}$ could be found numerically in the region near w^*. The errors in the α's are then found by inverting the H-matrix to obtain the error matrix for α; i.e., $\overline{(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)} = (H^{-1})_{ij}$. The second derivatives can be found numerically by using

$$\frac{\partial^2 w}{\partial\alpha_i \partial\alpha_j} = [w(\alpha_i + \Delta\alpha_i, \alpha_j + \Delta\alpha_j) + w(\alpha_i, \alpha_j) - w(\alpha_i + \Delta\alpha_i, \alpha_j) - w(\alpha_i, \alpha_j + \Delta\alpha_j)]/\Delta\alpha_i \Delta\alpha_j.$$

In the case of least squares use H_{ij} = ½ ðS / ðα_i ðα_j .

So far we have for the sake of simplicity talked in terms of evaluating $w(\alpha)$ over a fine grid in M-dimensional space. In most cases this would be much too time consuming. A rather extensive methodology has been developed for finding maxima or minima numerically. In this appendix we shall outline just one such approach called the method of steepest descent. We shall show how to find the least squares minimum of $S(\alpha)$. (This is the same as finding a maximum in $w(\alpha)$).

Method of Steepest Descent

At first thought one might be tempted to vary α_1 (keeping the other α's fixed) until a minimum is found. Then vary α_2 (keeping the others fixed) until a mew minimum is found, and so on. This is illustrated in Fig. 8 where M = 2 and the errors are strongly correlated. But in Fig. 8 many trials are needed. This stepwise procedure does converge, but in the case of Fig. 8, much too slowly. In the method of steepest descent one moves against the gradient in α-space:

$$\nabla_\alpha S = \frac{\partial S}{\partial\alpha_1}\hat{\alpha}_1 + \frac{\partial S}{\partial\alpha_2}\hat{\alpha}^2 + \ldots$$

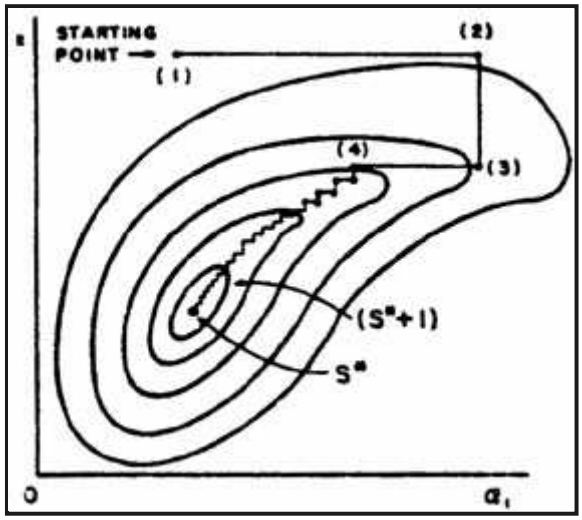

Figure 8. Contours of constant S vs. α_1 and α_2. Stepwise search for the minimum.

So we change all the α's simultaneously in the ratio ðS / ðα_1 : ðS / ðα_2 : ðS / ðα_3 : ... . In order to find the minimum along this line in α-space one should use an efficient step size. An effective method is to assume $S(s)$ varies quadratically from the minimum position s* where s is the distance along this line. Then the step size to the minimum is

$$s_0 = s_1 + \frac{\Delta s}{2}\frac{3S_1 - 4S_2 + S_3}{S_1 - 2S_2 + S_3}$$

where S_1, S_2, and S_3 are equally spaced evaluations of $S(s)$ along s with step size Δs starting from s_1; i.e., $s_2 = s_1 + \Delta s$, $s_3 = s_1 + 2\Delta s$. One or two iterations using the above formula will reach the minimum along s shown as point (2) in Fig. 9. The next repetition of the above procedure takes us to point (3) in Fig. 9. It is clear by comparing Fig. 9 with Fig. 8 that the method of steepest descent requires much fewer computer evaluations of $S(\alpha)$ than does the one variable at a time method.

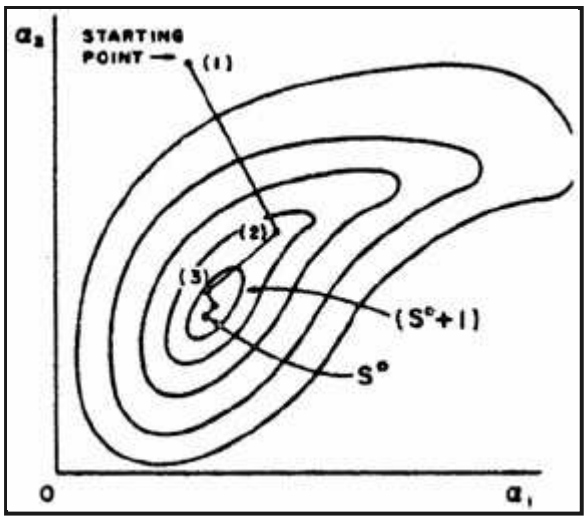

Figure 9. Same as Fig. 8, but using the method of steepest descent.

Least Squares with Constraints

In some problems the possible values of the α_j are restricted by subsidiary constraint relations. For example, consider an elastic scattering event in a bubble chamber where the measurements y_j are track coordinates and the α_i are track directions and momenta. However, the combinations of α_i that are physically possible are restricted by energy-momentum conservation. The most common way of handling this situation is to use the 4 constraint equations to eliminate 4 of the α's in $S(\alpha)$. Then S is minimized with respect to the remaining α's. In this example there would be (9 - 4) = 5 independent α's: two for orientation of the scattering plane, one for direction of incoming track in this plane, one for momentum of incoming track, and one for scattering angle. There could also be constraint relations among the measurable quantities y_i. In either case, if the method of substitution is too cumbersome, one can use the method of Lagrange multipliers.

In some cases the constraining relations are inequalities rather than equations. For example, suppose it is known that α_1 must be a positive quantity. Then one could define a new set of α's where $(\alpha_1')^2 = \alpha_1$, $\alpha_2' = \alpha_2$, etc. Now if $S(\alpha')$ is minimized n non-physical values of a will be used in the search for the minimum.

Appendix V. Cumulative Gaussian and Chi-Squared Distributions

The χ^2 confidence limit is the probability of Chi-squared exceeding the observed value; i.e.,

$$\mathrm{CL} = \int_{\chi^2}^{\infty} P_p(\chi^2)\, d\chi^2$$

where P_p for p degrees of freedom is given by Eq. (30a).

Gaussian Confidence Limits

Let $\chi^2 = [x / \sigma]^2$. Then for $n_D = 1$,

$$dP_1 = \frac{1}{\sqrt{2}\,\Gamma(\frac{1}{2})}\left[\frac{x}{\sigma}\right]^{-1} \exp\left[-\frac{x^2}{2\sigma^2}\right] d\left[\frac{x}{\sigma}\right]^2 = 2\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)\right] dx$$

Thus CL for n_D is twice the area under a single Gaussian tail. For example the $n_D = 1$ curve for $\chi^2 = 4$ has a value of $CL = 0.046$. This means that the probability of getting $|x| \geq 2\sigma$ is 4.6% for a Gaussian distribution.

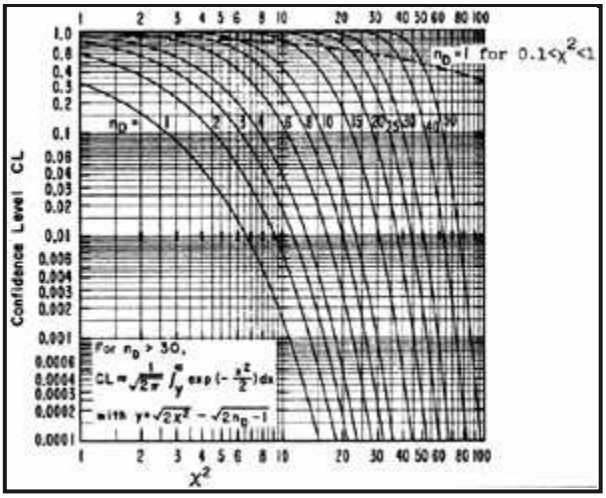

Figure 10. χ^2 Confidence Level vs. χ^2 for n_D Degrees of Freedom (9).

REFERENCES

1. Frank Solmitz, Ann. Rev. Nucl. Sci. 14, 375 (1964)
2. In 1958 it was common to use probable error rather than standard deviation. Also some physicists would deliberately multiply their estimated standard deviations by a "safety" factor (such as π). Such practices are confusing to other physicists who in the course of their work must combine, compare, interpret, or manipulate experimental results. By 1980 most of these misleading practices had been discontinued.
3. An equivalent statement is that in the inverse probability approach (also called Baysean approach) one is implicitly assuming that the prior probabilities are equal.
4. H. Cramer, Mathematical Methods of Statistics, Princeton University Press, 1946.
5. M. Annis, W. Cheston, and H. Primakoff, Rev. Mod. Phys. 25, 818 (1953).
6. A. G. Frodesen, O. Skjeggestad, and H. Tofte, Probability and Statistics in Particle Physics. (Columbia University Press, 1979) ISBN 82-00-01906-3. The title is misleading, this is an excellent book for physicists in all fields who wish to pursue the subject more deeply than is done in these notes.
7. J. Orear, "Least Squares When Both Variables have Uncertainties", Amer. Jour. Phys., Oct. 1982.
8. Some statistics books written specifically for physicists are: H. D. Young, "Statistical Treatment of Experimental Data," (McGraw-Hill Book Co., New York, 1962). P. R. Bevington, "Data Reduction and Error Analysis for the Physical Sciences," (McGraw-Hill Book Co., New York 1969). W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet, "Statistical Methods in Experimental Physics," (North-Holland Publishing Co., Amsterdam-London, 1971). S. Brandt, "Statistical and Computational Methods in Data Analysis," second edition (Elsevier North-Holland Inc., New York, 1976.) S. L. Meyer, "Data Analysis for Scientists and Engineers" (John Wiley and Sons, New York, 1975).
9. Reprinted from Rev. Mod. Phys. 52, No. 2, Part 11, April 1980 (page 536).