CERN BIBLIOTHEQUE
C1

NOT TO BE
TAKEN AWAY

HH

UCRL-8417
Physics and Mathematics

UNIVERSITY OF CALIFORNIA
Radiation Laboratory
Berkeley, California

Contract No. W-7405-eng-48

CERN LIBRARIES, GENEVA
SCAN-9709037

SW9738

NOTES ON STATISTICS FOR PHYSICISTS

Jay Orear

August 13, 1958

9680

Printed for the U.S. Atomic Energy Commission

UNIVERSITY OF CALIFORNIA

Ernest O. Lawrence Radiation Laboratory
Berkeley, California

Contract No. W-7405-eng-48

ERRATA

Page		
6	last line	Change d to f.
8	1st eq.	$(x-a_0)$ should be $(x-a_0)^2$
14	1st eq.	Change $-\frac{1}{a_2^3}$ to $\frac{1}{a_2^3}$
14	3rd eq.	Square root sign should cover entire right hand side.
14	6th eq.	Change $\frac{\partial^2 w}{\partial a_1 a_2}$ to $\frac{\partial^2 w}{\partial a_1 \partial a_2}$
14	7th eq.	In matrix H change $\frac{H}{a_2^{*2}}$ to $\frac{N}{a_2^{*2}}$
15	Everywhere it appears change H_{ab}^{-1} to $(H^{-1})_{ab}$	
17	4th eq.	Remove subscript from da_1
23	3rd line	Change $a^* \neq a$ to $a^* \neq \bar{a}$
24	5th eq.	Change $\underset{\sim}{H}_{ij}^{-1}$ to $(\underset{\sim}{H}^{-1})_{ij}$
25	8th line	Change $a_5 \equiv a_2/a_3$ to $a_5 \equiv a_4/a_3$
30	1st line	Denominator on right hand side should be $2^{P/2}\,\Gamma\,(P/2)$

33 The 3 equations for $\mathcal{M}_0$ should go as follows:

$$\mathcal{M}_0 = \sum_{a=1}^{P} \sum_{b=1}^{M} \left[(Z_a - a_b^* F_{ab}) + (a_b^* - a_b) F_{ab} \right]^2$$

$$\mathcal{M}_0 = \sum_a \sum_b \left(\frac{y_a}{\sigma_a} - a_b \frac{f_b(x_a)}{\sigma_a} \right)^2 + 2(\underset{\sim}{Z} - \underset{\sim}{a}^* \cdot \underset{\sim}{\tilde{F}}) \underset{\sim}{F} (\underset{\sim}{\tilde{a}}^* - \underset{\sim}{\tilde{a}}) + (\underset{\sim}{a}^* - \underset{\sim}{a}) \underset{\sim}{\tilde{F}} \cdot \underset{\sim}{F} (\underset{\sim}{\tilde{a}}^* - \underset{\sim}{\tilde{a}})$$

$$\mathcal{M}_0 = \mathcal{M} + 2(\underset{\sim}{Z} \cdot \underset{\sim}{F} - \underset{\sim}{a}^* \cdot \underset{\sim}{F}\underset{\sim}{F})(\underset{\sim}{\tilde{a}}^* - \underset{\sim}{\tilde{a}}) + (\underset{\sim}{Z} \cdot \underset{\sim}{F} \cdot \underset{\sim}{H}^{-1} - \underset{\sim}{a}\underset{\sim}{H}\underset{\sim}{H}^{-1})\underset{\sim}{H}(\underset{\sim}{H}^{-1}\underset{\sim}{\tilde{F}}\underset{\sim}{\tilde{Z}} - \underset{\sim}{H}^{-1}\underset{\sim}{H}\underset{\sim}{\tilde{a}})$$

NOTES ON STATISTICS FOR PHYSICISTS

Contents

Preface 3
1. Direct Probability 4
2. Inverse Probability 4
3. Likelihood Ratios 5
4. Maximum-Likelihood Method 6
5. Gaussian Distributions 8
6. The Magic Formula: Maximum-Likelihood Error, One Parameter . 9
7. Maximum-Likelihood Errors, M-Parameters, Correlated Errors . 11
8. Propagation of Errors, the Error Matrix 15
9. Systematic Errors 15
10. Uniqueness of Maximum-Likelihood Solution 16
11. Confidence Intervals and Their Arbitrariness 17
12. Bartlett S-Function 18
13. Binomial Distribution 19
14. Poisson Distribution 21
15. Extended Maximum-Likelihood Method 23
16. Least-Squares Method 25
17. Goodness of Fit, the χ^2-Distribution 29
Appendix I: Prediction of Likelihood Ratios 32
Appendix II: Distribution of the Least-Squares Sum 34

NOTES ON STATISTICS FOR PHYSICISTS

Jay Orear*

Radiation Laboratory
University of California
Berkeley, California

August 13, 1958

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] Other useful references are Annis, Cheston, and Primakoff,[2] M. S. Bartlett,[3] and H. Cramer.[4] The general presentation will be to study the Gaussian distribution, binomial distribution, Poisson distribution, and least-squares method in that order as applications of the maximum-likelihood method.

---

*Permanent address: Department of Physics, Cornell University, Ithaca, New York.

[1] Frank Solmitz, Notes on the Least Squares and Maximum Likelihood Methods, Institute for Nuclear Studies Report, University of Chicago.

[2] M. Annis, W. Cheston, and H. Primakoff, On Statistical Estimation in Physics, Revs. Modern. Phys. 25, 818 (Oct. 1953).

[3] M. S. Bartlett, On the Statistical Estimation of Mean Lifetimes, Phil. Mag., 44, 249 (1953).

[4] H. Cramer, Mathematical Methods of Statistics (Princeton University Press, 1946).

NOTES ON STATISTICS FOR PHYSICISTS

Jay Orear

Radiation Laboratory
University of California
Berkeley, California

August 13, 1958

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-$\frac{1}{2}$ 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 a 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 a_0. Usually the physicist does an experiment and quotes a result $a = a^* \pm \Delta a$. The major portion of this report is devoted to the questions What do we mean by a^* and Δa? and What is the "best" way to calculate a^* and Δa? These are questions of extreme importance to all physicists.

Crudely speaking, Δa is the standard deviation,[5] and what the physicist usually means is that the "probability" of finding

$$(a^* - \Delta a) < a_0 < (a^* + \Delta a) \text{ is } 68.3\%$$

(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

$$(a^* - \Delta a) < a_0 < (a^* + \Delta a) \text{ is either 0 or 1.}$$

The kind of probability the physicist is talking about here is called 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 τ^+ 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.

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

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

---

[5] Some physicists use probable error rather than standard deviation. Also some physicists deliberately multiply their estimated standard deviations by a "safety" factor (such as π) before publishing their results. Such practices are confusing to other physicists who in the course of their work must combine, compare, interpret, or manipulate experimental results.

The likelihood ratio $\mathcal{R}$ is

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

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 experiments 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 a priori probabilities favoring A or B. All the remaining material in this report is based on this basic principle alone. The modifications applied when a priori 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 τ meson one wishes to establish betting odds of 10^4 to 1 against spin 1. How many events will be needed for this? This 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,

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

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, $\mathcal{L}(\alpha)$.

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

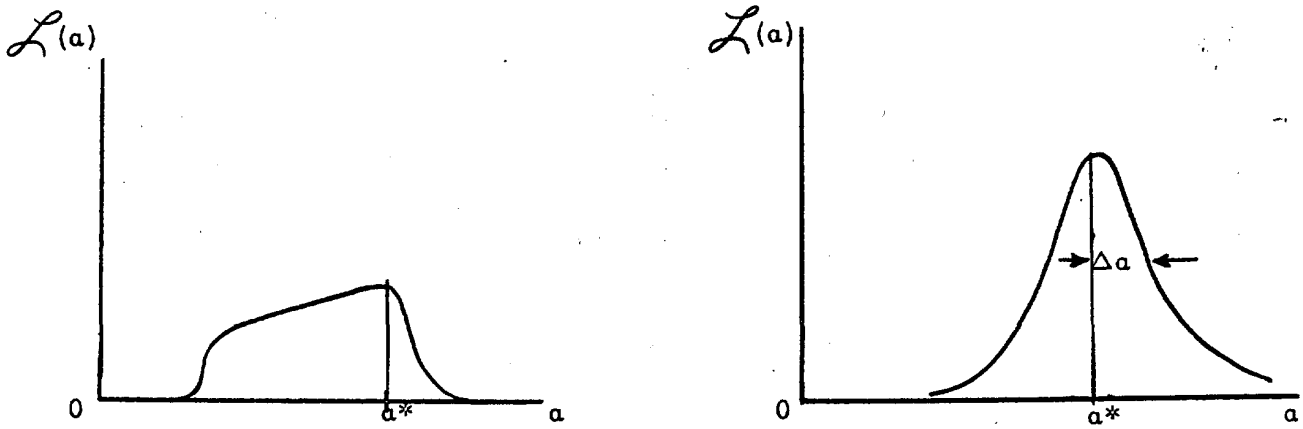
The likelihood function, $\mathcal{L}(\alpha)$, is the joint probability density of getting a particular experimental result, $x_1 \ldots, x_n$, assuming $f(\alpha;x)$ is the true normalized distribution function:

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

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

$$\Delta a = \left[\frac{\int (a-a^*)^2 \mathcal{L}\, da}{\int \mathcal{L}\, da} \right]^{\frac{1}{2}} \tag{3}$$

In general, the likelihood function will be close to Gaussian (it can be shown to approach a Gaussian distribution as $N \rightarrow \infty$) and will look similar to the right-hand figure below.



The left-hand figure represents what is called a case of poor statistics. In such a case, it is better to present a plot of $\mathcal{L}(a)$ rather than merely quoting a^* and Δa. Straightforward procedures for obtaining Δa are presented in Sections 6 and 7.

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, $a^* \rightarrow a_0$; and furthermore, there is no other method of estimation that is more accurate.

In the general case in which there are M parameters, $a_1 \ldots, a_M$, to be determined, the procedure for obtaining the maximum-likelihood solution is to solve the M simultaneous equations,

$$\boxed{\left.\frac{\partial w}{\partial a_i}\right|_{a_i = a_i^*} = 0} \qquad \text{where } w \equiv \ln \mathcal{L}(a_1, \ldots, a_M), \tag{4}$$

5. Gaussian Distributions

As a first application of the maximum-likelihood method, we consider the example of the measurement of a physical parameter a_0, where x is the result of a particular type of measurement that is known to have a measuring error σ. Then if x is Gaussian-distributed,[6] the distribution function is

$$f(a_0; x) = \frac{1}{\sqrt{2\pi}\,\sigma} \cdot \exp\left[-(x-a_0)^2/2\sigma^2\right] .$$

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

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

then

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

$$\frac{\partial w}{\partial a} = \Sigma \frac{x_i - a}{\sigma_i^2} , \tag{5}$$

$$\Sigma \frac{x_i}{\sigma_i^2} - \Sigma \frac{a^*}{\sigma_i^2} = 0 ;$$

$$a^* = \frac{\Sigma \frac{1}{\sigma_i^2} x_i}{\Sigma \frac{1}{\sigma_i^2}} \tag{6}$$

is the maximum-likelihood solution. 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

$$a^* = \frac{\Sigma x_i}{N} .$$

Next we consider the accuracy of this determination.

---

[6] A derivation of the Gaussian distribution and its relation to the binomial and Poisson distributions is given in Chapter II of Physical Statistics by R. B. Lindsay (Wiley, New York, 1941).

6. The Magic Formula: Maximum-Likelihood Error, One Parameter

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

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

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

$$w = -\frac{h}{2}(a - a^*)^2 + \text{const},$$

$$\frac{\partial w}{\partial a} = -h(a - a^*),$$

$$\frac{\partial^2 w}{\partial a^2} = -h .$$

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

$$\boxed{\Delta a = \left[-\frac{\partial^2 w}{\partial a^2}\right]^{-\frac{1}{2}}} \qquad \text{Magic Formula I.} \qquad (7)$$

Now the error of the above determination, Eq. (6), can easily be found by differentiating Eq. (5) with respect to a. The answer is

$$\Delta a = \left[\Sigma \frac{1}{\sigma_i^2}\right]^{-\frac{1}{2}} .$$

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 a^* nor Δa may be found analytically. In such cases the curve $\mathcal{L}(a)$ can be found numerically by trying several values of a and using Eq. (2) to get the corresponding values of $\mathcal{L}(a)$. The complete function is then obtained by using a French curve. If $\mathcal{L}(a)$ is Gaussianlike, $\partial^2 w/\partial a^2$ is the same everywhere. If not, it is best to use the average

$$\overline{\frac{\partial^2 w}{\partial a^2}} = \frac{\int (\partial^2 w/\partial a^2)\,\mathcal{L}\,da}{\int \mathcal{L}\,da} .$$

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

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

Thus, use of the average second derivative gives the required larger error. This technique is discussed further in Section 12.

Note that Magic Formula I 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 Magic Formula II, which depends only on knowledge of $f(a;x)$. Under these circumstances we wish to determine $\overline{\partial^2 w/\partial a^2}$ averaged over many repeated experiments consisting of N events each. For one event we have

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

for N events,

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

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

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

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

The last integral vanishes if one integrates before the differentiation because

$$\int f dx = 1 .$$

Thus

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

and Eq. (7) leads to

$$\boxed{\Delta a = \frac{1}{\sqrt{N}} \left[\int \frac{1}{f} \left(\frac{\partial f}{\partial a}\right)^2 dx \right]^{-\frac{1}{2}}} \quad \text{Magic Formula II} \qquad (8)$$

(the case in which there is no experimental result).

Example

Assume in the μ-e decay distribution function, $f(a;x) = \frac{1+ax}{2}$, that $a_0 = -1/3$. How many μ-e decays are needed to establish a to a 1% accuracy (i.e., $a/\Delta a = 100$)?

$$\frac{\partial f}{\partial a} = \frac{x}{2} .$$

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

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

Note that

$$\lim_{a \to \infty} [\Delta a] = \sqrt{\frac{3}{N}} .$$

For

$$a = -\frac{1}{3} , \qquad \Delta a = \sqrt{\frac{2.8}{N}} .$$

For this problem

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

---

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 $\overline{(a_i - a_i^*)(a_j - a_j^*)} = 0$ for all cases with $i \neq j$. For the general case we Taylor-expand $w(a)$ about (a^*):

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

where

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

and

$$H_{ij} \equiv - \left.\frac{\partial^2 w}{\partial \alpha_i \partial \alpha_j}\right|_{\alpha^*} . \tag{9}$$

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

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

Neglecting the higher-order terms, we have

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

(an M-dimensional Gaussian surface). As before, our error formulas depend on the approximation that $\mathcal{L}(\alpha)$ is Gaussianlike 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 $\mathcal{L}(\alpha)$.

According to Eq. (9), $\underset{\sim}{H}$ is a symmetric matrix. Let $\underset{\sim}{U}$ be the unitary matrix that diagonalizes $\underset{\sim}{H}$:

$$\underset{\sim}{U} \cdot \underset{\sim}{H} \cdot \underset{\sim}{U}^{-1} = \begin{pmatrix} h_1 & & & 0 \\ & h_2 & & \\ & & \ddots & \\ 0 & & & h_M \end{pmatrix} \equiv \underset{\sim}{h} \qquad \text{where } \tilde{\underset{\sim}{U}} = \underset{\sim}{U}^{-1} . \tag{10}$$

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

$$d^M P = C \exp \left[-\frac{1}{2}(\underset{\sim}{\gamma} \cdot \underset{\sim}{U}) \cdot \underset{\sim}{H} (\widetilde{\underset{\sim}{\gamma} \cdot \underset{\sim}{U}})\right] d^M \beta .$$

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

$$d^M p = C \exp \left[-\left(\frac{1}{2}\right) \sum_a h_a \gamma_a^2\right] d^M \gamma .$$

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

$$\overline{\beta_i \beta_j} = \sum_a \sum_b \overline{\gamma_a \gamma_b}\, U_{ai}\, U_{bj}$$

$$= \sum_a U^{-1}_{ia}\, h_a^{-1}\, U_{aj}$$

$$= (\underset{\sim}{U}^{-1} \cdot \underset{\sim}{h} \cdot \underset{\sim}{U})^{-1}_{ij} .$$

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

$$\boxed{\begin{array}{l} \overline{(a_i - a_i^*)(a_j - a_j^*)} = (\underset{\sim}{H}^{-1})_{ij} \quad \text{where } H_{ij} = -\dfrac{\partial^2 w}{\partial a_i \partial a_j} \\ \\ \text{Averaged over repeated experiments} \\ \qquad \overline{H}_{ij} = N \displaystyle\int \frac{1}{f}\left(\frac{\partial f}{\partial a_i}\right)\left(\frac{\partial f}{\partial a_j}\right) dx \end{array}} \qquad \begin{array}{l} \text{Magic Formula III} \\ \\ (11) \end{array}$$

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

$$(H^{-1})_{ij} = \cancel{H^{-1}_{ij}} = (-1)^{i+j} \times \frac{\text{ijth minor of } \underset{\sim}{H}}{\text{determinant of } \underset{\sim}{H}} .$$

Example: Assume that the ranges of monoenergetic particles are Gaussian-distributed with mean range a_1 and straggling coefficient a_2 (the standard deviation). N particles having ranges $x_1, \ldots, x_N$ are observed. Find a_1^*, a_2^*, and their errors.

Then

$$\mathcal{L}(a_1, a_2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\, a_2} \exp\left[-(x_i - a_1)^2 / 2a_2^2\right]$$

$$w = -\tfrac{1}{2} \sum_i \frac{(x_i - a_1)^2}{a_2^2} - N \ln a_2 - N \ln(2\pi) ,$$

$$\frac{\partial w}{\partial a_1} = \sum_i \frac{(x_i - a_1)}{a_2^2} ,$$

$$\frac{\partial w}{\partial a_2} = \frac{1}{a_2^3} \sum_i (x_i - a_1)^2 - \frac{N}{a_2} .$$

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

$$\left\{ \begin{array}{l} a_1^* = \frac{1}{N} \sum_i x_i \\ \\ a_2^* = \sqrt{\frac{\Sigma (x_i - a_1^*)^2}{N}} \end{array} \right\}$$

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

$$\overline{a_2} = \sqrt{\frac{\Sigma (x_i - a_1^*)^2}{N-1}}$$

This is because in this case the most probable value, a_2^*, and the mean, a_2, do not occur at the same place. Mean values of such quantities are studied in Section 17. The matrix H is obtained by evaluating the following quantities at a_1^* and a_2^*:

$$\frac{\partial^2 w}{\partial a_1^2} = - \frac{N}{a_2^2} , \quad \frac{\partial^2 w}{\partial a_2^2} = - \frac{3}{a_2^4} \Sigma (x_i - a_1)^2 + \frac{N}{a_2^2} ,$$

$$\frac{\partial^2 w}{\partial a_1 \partial a_2} = - \frac{2}{a_2^2} \Sigma (x_i - a_1) ,$$

$$\underset{\sim}{H} = \begin{pmatrix} \frac{N}{a_2^{*2}} & 0 \\ 0 & \frac{2N}{a_2^{*2}} \end{pmatrix} \quad \text{and} \quad \underset{\sim}{H}^{-1} = \begin{pmatrix} \frac{a_2^{*2}}{N} & 0 \\ 0 & \frac{a_2^{*2}}{2N} \end{pmatrix}$$

According to Eq. (11), the errors on a_1 and a_2 are the square roots of the diagonal elements of the error matrix, $\underset{\sim}{H}^{-1}$:

$$\Delta a_1 = \frac{a_2^*}{\sqrt{N}} \quad \text{and} \quad \Delta a_2 = \frac{a_2^*}{\sqrt{2N}}$$ (this is sometimes called the error of the error).

---

8. Propagation of Errors: the Error Matrix

Consider the case in which a single physical quantity, y, is some function of the α's: $y = y(\alpha_1, \ldots, \alpha_M)$. The "best" value for y is then $y^* = y(\alpha_i^*)$. To first order in $(\alpha_i - \alpha_i^*)$ we have

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

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

$$(\Delta y)_{rms} = \sqrt{\sum_a \sum_b \frac{\partial y}{\partial \alpha_a} \frac{\partial y}{\partial \alpha_b} \cancel{H_{ab}^{-1}} (H^{-1})_{ab}} \tag{12}$$

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

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

It is a common problem to be interested in M physical parameters, $y_1, \ldots, y_M$, which are known functions of the α_i. If the error matrix $\underset{\sim}{H}^{-1}$ of the α_i is known, then we have

$$\overline{(y_i - y_i^*)(y_j - y_j^*)} = \sum_a \sum_b \frac{\partial y_i}{\partial \alpha_a} \frac{\partial y_j}{\partial \alpha_b} \cancel{H_{ab}^{-1}} (H^{-1})_{ab} \tag{13}$$

In some such cases the $\frac{\partial y_i}{\partial \alpha_a}$ cannot be obtained directly, but the $\frac{\partial \alpha_i}{\partial y_a}$ are easily obtainable. Then

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

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 a_{M+1} to account for the systematic effect.

In some cases a systematic effect cannot be estimated apart from the main experiment. The example given in Section 7 can be made into such a case. Let us assume that among the beam of monoenergetic particles there is an unknown background of particles uniformly distributed in range. In this case the distribution function would be

$$f(a_1, a_2, a_3; x) = \frac{1}{C}\left\{\frac{1}{\sqrt{2\pi}\, a_2} \exp[-(x-a_1)^2/2a_2^2] + a_3\right\},$$

where

$$C(a_1, a_2, a_3) = \int_{x_{min}}^{x_{max}} f\, dx.$$

The solution a_3^* is simply related to the percentage of background.

10. Uniqueness of Maximum-Likelihood Solution

Usually it is a matter of taste what physical quantity is chosen as a. 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 a. The maximum-likelihood solution for a is obtained from the equation $\partial w/\partial a = 0$. The maximum-likelihood solution for λ is obtained from $\partial w/\partial \lambda = 0$. But then we have

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

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 $\tau^* = 1/\lambda^*$.

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

$$\mathcal{L}(a) = G(a)\,\mathcal{H}(a);$$

$$w = \ln G + \ln \mathcal{H}.$$

$$\frac{\partial w}{\partial a} = \frac{\partial}{\partial a} \ln G + \frac{\partial}{\partial a} \ln \mathcal{H},$$

$$\frac{\partial}{\partial a} \ln \mathcal{H}(a^*) = - \frac{\partial}{\partial a} \ln G(a^*)$$

give the maximum-likelihood solution. In the absence of any a priori knowledge the term on the right-hand side is zero. In other words, the standard procedure in the absence of any a priori information is to use an a priori distribution in which all values of a are equally probable. Strictly speaking, it is impossible to know a "true" $G(a)$, because it in turn must depend on its own a priori probability. However, the above equation is useful when $G(a)$ is the combined likelihood function of all previous experiments and $\mathcal{H}(a)$ 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 a, $G(a)$, rather than a single value a_0. For example, one may wish to determine the momentum distribution of cosmic ray muons. Here one observes

$$\mathcal{L}(G) = \int G(a)\, \mathcal{H}(a;x)\, da,$$

where $\mathcal{H}(a;x)$ is known from the nature of the experiment and $G(a)$ is the function to be determined. This type of problem is discussed in Reference 2.

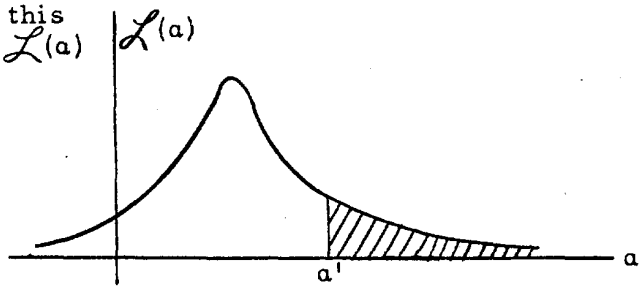
11. Confidence Intervals and Their Arbitrariness

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

$$P(a' < a < a'') = \int_{a'}^{a''} \mathcal{L}\, da \Big/ \int_{-\infty}^{\infty} \mathcal{L}\, da .$$

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

$$P(a > a') = \frac{\int_{a'}^{\infty} \mathcal{L}\, da}{\int_{-\infty}^{\infty} \mathcal{L}\, da}$$



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

$$P(\lambda > \lambda') = \frac{\int_{\lambda'}^{\infty} \mathcal{L}\, d\lambda}{\int_{-\infty}^{\infty} \mathcal{L}\, d\lambda} = \frac{\int_{a'}^{\infty} \mathcal{L} \frac{\partial \lambda}{\partial a}\, da}{\int_{-\infty}^{\infty} \mathcal{L}\, d\lambda}$$

$$\neq P(a > a') .$$

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 Δa. Only the maximum-likelihood solution and the relative probabilities are unaffected by the choice of a. 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.

12. Bartlett S. Function

M. S. Bartlett discusses a type of confidence interval that avoids some of the above objections.[3] He defines a function $S(a)$ which always has a mean of zero and standard deviation of one, independent of the choice of a:

$$S(a) \equiv \frac{1}{C} \frac{\partial w}{\partial a} \qquad \text{where } C^2 \equiv - \int_{a_{min}}^{a_{max}} \frac{\partial^2 w}{\partial a^2} \mathcal{L}(a) da .$$

For an $\mathcal{L}(a)$ which is a Gaussian curve with standard deviation Δa, $S(a)$ would then be

$$S(a) = - \frac{a - a^*}{\Delta a} .$$

Bartlett proposes that, since S is closer than a to being Gaussian distributed, the 68.3% confidence interval (one standard deviation) in a can be obtained by solving for the two values of a which give $S(a') = +1$ and $S(a'') = -1$. Similarly the 2-standard-deviation confidence interval is obtained by solving for $S(a) = \pm 2$. Bartlett's paper also contains a further refinement of a skewness correction.[3] We now demonstrate that we have $\overline{S} = 0$ and $\overline{S^2} = 1$.

$$\overline{S} = \frac{1}{C}\int\left(\frac{\partial w}{\partial a}\right)\mathcal{L}\,da = \frac{1}{C}\int\frac{\partial\mathcal{L}}{\partial a}\,da = [\mathcal{L}(a_{max}) - \mathcal{L}(a_{min})] = 0\,,$$

$$\overline{S^2} = \frac{1}{C^2}\int\frac{1}{\mathcal{L}^2}\left(\frac{\partial\mathcal{L}}{\partial a}\right)^2\mathcal{L}\,da = \frac{\int\frac{1}{\mathcal{L}}\left(\frac{\partial\mathcal{L}}{\partial a}\right)^2 da}{-\int\frac{\partial}{\partial a}\left(\frac{1}{\mathcal{L}}\frac{\partial\mathcal{L}}{\partial a}\right)\mathcal{L}\,da} =$$

$$\frac{\int\frac{1}{\mathcal{L}}\left(\frac{\partial\mathcal{L}}{\partial a}\right)^2 da}{-\int\frac{\partial^2\mathcal{L}}{\partial a^2}\,da + \int\frac{1}{\mathcal{L}}\left(\frac{\partial\mathcal{L}}{\partial a}\right)^2 da}\,,$$

$\therefore\ \overline{S^2} = 1$, because the term $-\int\frac{\partial^2\mathcal{L}}{\partial a^2}\,da = \left.\frac{\partial\mathcal{L}}{\partial a}\right|_{a_{min}} - \left.\frac{\partial\mathcal{L}}{\partial a}\right|_{a_{max}} = 0.$

13. 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

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

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

$$\mathcal{L}(p) = \frac{N!}{N_1!\,(N-N_1)!}\; p^{N_1}(1-p)^{N-N_1}$$

$$w = N_1 \ln p + (N-N_1)\ln(1-p) + \text{const}$$

$$\frac{\partial w}{\partial p} = \frac{N_1}{p} - \frac{N-N_1}{1-p} \quad , \tag{15}$$

$$\frac{\partial^2 w}{\partial p^2} = -\frac{N_1}{p^2} - \frac{N-N_1}{(1-p)^2} \quad , \tag{16}$$

From Eq. (15): we have

$$\boxed{p^* = \frac{N_1}{N}} \tag{17}$$

From (16) and (17):

$$\overline{(p-p^*)^2} = \frac{1}{\dfrac{N_1}{p^{*2}} - \dfrac{N-N_1}{(1-p^*)^2}} \quad ,$$

$$\boxed{\Delta p = \sqrt{\frac{p^*(1-p^*)}{N}}} \tag{18}$$

The results, Eqs. (17) and (18), also happen to be the same as those using direct probability.[5] Then

$$\overline{N}_1 = pN$$

and

$$\overline{(N_1-\overline{N}_1)^2} = Np(1-p).$$

Example: In the previous example (see Section 6) 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

$$p = \int_0^1 \frac{1 + ax}{2}\, dx = \frac{1 + \frac{a}{2}}{2} \quad ,$$

$dp = \frac{1}{4}da$, and in the limit of small errors, $\Delta a = 4\Delta p$.

By Eq. (18),

$$\Delta a = 4\sqrt{\frac{p^*(1-p^*)}{N}} \quad ,$$

$$\Delta a = \sqrt{\frac{4}{N}\left(1 - \frac{a^{*2}}{4}\right)} \quad .$$

---

14. 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

$$dP(0, x) = - P(0, x) \times \frac{dx}{\lambda} \quad ,$$

$$\ln P(0, x) = - \frac{x}{\lambda} + \text{const},$$

$$P(0, x) = e^{-x/\lambda} \quad (\text{at } x = 0,\ P(0, x) = 1). \tag{19}$$

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 $dx_1, \ldots dx_N$ times the probability of no events in the remaining length:

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

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

$$\prod_{i=1}^{N} \int_0^x \frac{dx_i}{\lambda} = \left(\frac{x}{\lambda}\right)^N$$

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

$$\boxed{P(N, x) = \frac{\left(\frac{x}{\lambda}\right)^N}{N!} \quad e^{-x/\lambda} ,} \qquad \text{the Poisson distribution.} \qquad (21)$$

As a check, note

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

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

Likewise it can be shown that $(N-\overline{N})^2 = N$.

Equation (21) is often expressed in terms of $\overline{N}$:

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

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

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 $a \equiv \overline{N}$ and its error:

$$\mathcal{L}(a) = \frac{a^N}{N!} \quad e^{-a} ,$$

$$w = N \ln a - a - \ln N! ,$$

$$\frac{\partial w}{\partial a} = \frac{N}{a} - 1 .$$

$$\frac{\partial^2 w}{\partial a^2} = - \frac{N}{a^2} .$$

Thus we have $\qquad a^* = N$

and by Eq. (7), $\qquad \Delta a = \frac{a}{\sqrt{N}}$.

In a cross-section determination, we have $a = \rho x \sigma$, 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 $a^* \neq \bar{a}$:

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

15. Extended 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

$$\int_{x_{min}}^{x_{max}} F\,dx \equiv \bar{N}(a)$$

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

$$\Delta x \text{ is } \exp\left(-\int_x^{x+\Delta x} F\,dx\right) .$$

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

$$\exp\left(-\int_{x_{min}}^{x_{max}} F\,dx\right) = e^{-\bar{N}} .$$

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

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

and

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

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

$$\frac{\partial w}{\partial a_i} = 0 \; .$$

The errors are still given by

$$\overline{(a_i - a_i^*)(a_j - a_j^*)} = \cancel{H_{ij}^{-1}}\ (\underset{\sim}{H}^{-1})_{ij}$$

where

$$H_{ij} = -\frac{\partial^2 w}{\partial a_i \partial a_j} \; .$$

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

$$-\overline{\frac{\partial^2 w}{\partial a_i \partial a_j}} = \int \frac{1}{F} \left(\frac{\partial F}{\partial a_i}\right)\left(\frac{\partial F}{\partial a_i}\right) dx \; .$$

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 the extended maximum-likelihood method and that in the limit of large N there is no method of estimation that is more accurate.

In the absence of the extended maximum-likelihood method our procedure would have been to normalize F(a;x) to unity by using

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

For example, consider a sample containing just two radioactive species, of lifetimes a_1 and a_2. Let a_3 and a_4 be the two initial decay rates. Then we have

$$F(a_i;x) = a_3\, e^{-x/a_1} + a_4\, e^{-x/a_2} ,$$

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

$$f(a;x) = \frac{e^{-x/a_1} + a_5\, e^{-x/a_2}}{a_1 + a_5\, a_2} ,$$

which is normalized to one. Note that the four original parameters have been reduced to three by using $a_5 \equiv a_4/a_3$. Then a_3 and a_4 would be found by using the auxilary equation

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

the total number of counts In this standard procedure the equation

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

must always hold. However, in the extended maximum-likelihood method these two quantities are not necessarily equal. Thus the extended maximum-likelihood method will give a different solution for the a_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 a priori information such as the shape of the angular distribution enables one to do a somewhat better job of calculating the cross section. In a private communication Frank Crawford has pointed out that the two methods give exactly the same answers in the special case in which $F(a_i;x)$ is homogeneous in the a_i.

16. 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. The case in which the x_i have known measurement errors is discussed in Reference 1.

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 $\sigma_i = \sqrt{N_i}$. In this case the likelihood function is

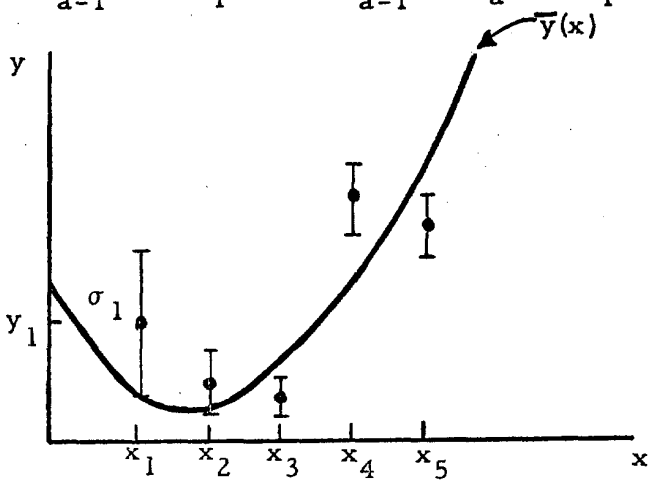
$$\mathcal{L} = \prod_{i=1}^{p} \frac{[\bar{y}(x_i)]^{N_i}}{N_i\,!} \; e^{-\bar{y}(x_i)}$$

and

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

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

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



The remainder of this section is devoted to the case in which the y_i are Gaussian-distributed with standard deviations σ_i. Here the famous least-squares method is applicable. 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

$$\mathcal{L} = \prod_{a=1}^{p} \frac{1}{\sqrt{2\pi}\,\sigma_a} \exp\left[-(y_a - \bar{y}(x_a))^2/2\sigma_a^2\right] \tag{23}$$

$$w = -\tfrac{1}{2}\mathcal{M} - \sum_{a=1}^{p} \ln \sqrt{2\pi}\,\sigma_a$$

where

$$\boxed{\mathcal{M} \equiv \sum_{a=1}^{p} \frac{[y_a - \bar{y}(x_a)]^2}{\sigma_a^2}} \tag{24}$$

The solutions $a_i = a_i^*$ are given by minimizing $\mathcal{M}$ (maximizing w):

$$\frac{\partial \mathcal{M}}{\partial a_i} = 0 . \tag{25}$$

This minimum value of $\mathcal{M}$ is called $\mathcal{M}^*$, the least-squares sum. The values of a_i which minimize $\mathcal{M}$ 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

$$\overline{(a_i - a_i^*)(a_j - a_j^*)} = (\underset{\sim}{H}^{-1})_{ij}, \quad \text{where } H_{ij} = \tfrac{1}{2} \frac{\partial^2 \mathcal{M}}{\partial a_i \partial a_j} .$$

Finally we consider the special case in which $\overline{y}(a_i; x)$ is linear in the a_i:

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

Then

$$\frac{\partial \mathcal{M}}{\partial a_i} = -2 \sum_{a=1}^{p} \left[\frac{y_a - \sum_{b=1}^{M} a_b f_b(x_a)}{\sigma_a^2} \right] f_i(x_a) , \tag{26}$$

and

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

Define

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

Then

$$\frac{\partial \mathcal{M}}{\partial a_i} = -2 \left[U_i - \sum_{b=1}^{M} a_b H_{bi} \right] .$$

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

$$0 = \underline{u} - \underline{a}^* \cdot \underset{\sim}{H} , \tag{29}$$

$$\underline{a}^* = \underline{u} \cdot \underset{\sim}{H}^{-1} ;$$

$$\boxed{\begin{aligned} a_i^* &= \sum_{a=1}^{M} \sum_{b=1}^{p} \frac{y_b f_a(x_b)}{\sigma_b^2} (\underset{\sim}{H}^{-1})_{ai} , \\ \overline{(a_i - a_i^*)(a_j - a_j^*)} &= \underset{\sim}{H}^{-1}{}_{ij} \quad \text{where } H_{ij} \equiv \sum_{a=1}^{p} \frac{f_i(x_a) f_j(x_a)}{\sigma_a^2} . \end{aligned}} \quad (30)$$

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 a_i) may jointly be fitted. Then the $\mathcal{M}$ value is the sum over all the points on all the curves.

Example: The curve 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.

$$\bar{y}(x) = a_1 + a_2 x + a_3 x^2 ,$$

$$f_i = 1, \; f_2 = x, \; f_3 = x^2 ,$$

$$H_{11} = \sum_{a=1}^{p} \frac{1}{\sigma_a^2} , \; H_{22} = \sum_a \frac{x_a^2}{\sigma_a^2} , \; H_{33} = \sum_a \frac{x_a^4}{\sigma_a^2} ,$$

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

$$\underset{\sim}{H} = \begin{pmatrix} 2.5 & 0 & 0.26 \\ 0 & 0.26 & 0 \\ 0.26 & 0 & 0.068 \end{pmatrix} \quad \underset{\sim}{H}^{-1} = \begin{pmatrix} 0.664 & 0 & -2.54 \\ 0 & 3.847 & 0 \\ -2.54 & 0 & 24.418 \end{pmatrix} ,$$

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

$$a_1^* = 3.685, \quad \Delta a_1 = 0.815, \quad \overline{\Delta a_1 \Delta a_2} = 0,$$
$$a_2^* = 3.27, \quad \Delta a_2 = 1.96, \quad \overline{\Delta a_1 \Delta a_3} = -2.54,$$
$$a_3^* = 7.808, \quad \Delta a_3 = 4.94.$$

$\overline{y}(x) = (3.685 \pm 0.815) + (3.27 \pm 1.96)x + (7.808 \pm 4.94)x^2$ is the best-fit curve.

---

17. Goodness of Fit, the χ^2 Distribution

The numerical value of the likelihood function at $\mathcal{L}(a^*)$ can, in principle, be used as a check on whether one is using the correct type of function for $f(a;x)$. 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 $\mathcal{L}(a^*)$ assuming a particular true $f(a_0, x)$. Then the probability of getting an $\mathcal{L}(a^*)$ 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 $\mathcal{L}(a^*)$, one would seriously question either the experiment or the function $f(a;x)$ that was used.

In practice, the determination of the distribution of $\mathcal{L}(a^*)$ 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 $\mathcal{M}(a^*)$ rather than of $\mathcal{L}(a^*)$. We shall first consider the distribution of $\mathcal{M}(a_0)$. According to Eqs. (23) and (24) the probability element is

$$d^p P \propto \exp[-\mathcal{M}/2]\, d^p y_i .$$

Note that $\mathcal{M} = \rho^2$, 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

$$d^p y_i \propto \rho^{p-1}\, d\rho \propto \mathcal{M}^{(p-1)/2} \mathcal{M}^{\frac{1}{2}}\, d\mathcal{M}.$$

Thus

$$dP(\mathcal{M}) \propto \mathcal{M}^{(p/2)-1}\, e^{(-\mathcal{M}/2)}\, d\mathcal{M}.$$

The normalization is obtained by integrating from $\mathcal{M}=0$ to $\mathcal{M} = \infty$.

$$dP(\mathcal{M}_{a_0}) = \frac{1}{2^{p/2}\,\Gamma(p/2)}\,\mathcal{M}_0^{(p/2)-1}\,e^{-\mathcal{M}_0/2}\,d\mathcal{M}_0, \tag{29}$$

where $\mathcal{M}_0 \equiv \mathcal{M}(a_0)$. $\Gamma(x) = \int_0^\infty e^{-t}\,t^{x-1}\,dt \qquad \Gamma(1) = 1$

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

$$\int_{\mathcal{M}_0}^{\infty} dP(\mathcal{M})$$

for several degrees of freedom are in the Handbook of Chemistry and Physics and other common mathematical tables.

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

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

Usually a_0 is not known. In such a case one is interested in the distribution of

$$\mathcal{M}^* \equiv \mathcal{M}(a^*).$$

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 have

$$\boxed{\begin{array}{l} dP(\mathcal{M}^*) = \chi^2 \text{ distribution for (p-M) degrees of freedom,} \\ \overline{\mathcal{M}^*} = (p-M) \text{ and } \Delta M^* = \sqrt{2(p-M)} \end{array}} \tag{31}$$

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

<u>Example 1:</u> Determine the χ^2 probability of the solution to the problem at the end of Section 16.

$$\mathcal{M}^* = \left(\frac{5 - \bar{y}(-.6)}{2}\right)^2 + \left(\frac{3 - \bar{y}(-.2)}{1}\right)^2 + \left(\frac{5 - \bar{y}(.2)}{1}\right)^2 + \left(\frac{8 - \bar{y}(.6)}{2}\right)^2,$$

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

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

$$\bar{y} = a_1 + a_2 x + a_3 x^2 .$$

Example 2: Two different laboratories have measures the lifetime of the θ_1 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 $a^* = 1.008 \times 10^{-10}$ sec. Thus

$$\mathcal{M}^* = \left(\frac{1.00 - 1.008}{0.01}\right)^2 + \left(\frac{1.04 - 1.008}{0.02}\right)^2 = 3.2 \qquad \overline{\mathcal{M}^*} = 2 - 1 = 1$$

According to the χ^2 table for one degree of freedom, the probability of getting $\mathcal{M}^* > 3.2$ 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

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

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 τ 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

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

$$N = 30.$$

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 tau is spin zero. The fluctuation (rms spread) of log $\mathcal{L}$ for a given N is

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

Appendix II: Distribution of the Least-Squares Sum

We shall define

$$Z_i \equiv \frac{y_i}{\sigma_i} \quad \text{and} \quad F_{ij} \equiv \frac{f_j(x_i)}{\sigma_i} \quad .$$

Note that $\underset{\sim}{H} = \underset{\sim}{\tilde{F}} \cdot \underset{\sim}{F}$ by Eq. (27),

$$\underset{\sim}{Z} \cdot \underset{\sim}{F} = \underline{a}^* \cdot \underset{\sim}{H} \text{ by Eq. (28) and (29).} \tag{33}$$

Then $\underline{a}^* = \underset{\sim}{Z} \cdot \underset{\sim}{F} \cdot \underset{\sim}{H}^{-1}$. (34)

$$\mathcal{M}_0 = \sum_{a=1}^{\ell} \sum_{b=1}^{M} [(Z_a - \tilde{a}_b{}^* \tilde{F}_{ab}) + (\tilde{a}^*{}_b - \tilde{a}_b)\tilde{F}_{ab}]^2 ,$$

where the unstarred a is used for a_0.

$$\mathcal{M}_0 = \sum_a^p \sum_b^M \left(\frac{y_a}{\sigma_a} - \frac{a^*{}_b f_b(x_a)}{\sigma_a}\right)^2 + 2(\underline{Z} - \underline{a}^* \cdot \underset{\sim}{\tilde{F}})\underset{\sim}{F}(\underline{\tilde{a}}^* - \underline{\tilde{a}}) + (\underline{a}^* - \underline{a})\underset{\sim}{\tilde{F}} \cdot \underset{\sim}{F}(\underline{\tilde{a}}^* - \underline{\tilde{a}}) ,$$

$$\mathcal{M}_0 = \mathcal{M}^* + 2(\underline{Z} \cdot \underset{\sim}{F} - \underline{a}^* \cdot \underset{\sim}{\tilde{F}}\underset{\sim}{F})(\underline{\tilde{a}}^* - \underline{\tilde{a}}) + (\underline{Z} \cdot \underset{\sim}{F} \cdot \underset{\sim}{H}^{-1} - \underline{a}\underset{\sim}{H}\underset{\sim}{H}^{-1})\underset{\sim}{H}(\underset{\sim}{H}^{-1}\underset{\sim}{\tilde{F}}\underset{\sim}{\tilde{Z}} - \underset{\sim}{H}^{-1}\underset{\sim}{H}\underline{\tilde{a}})$$

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

$$\mathcal{M}^* = \mathcal{M}_0 - (\underline{Z} \cdot \underset{\sim}{F} - \underline{a}\underset{\sim}{\tilde{F}}\underset{\sim}{F})\underset{\sim}{H}^{-1}\underset{\sim}{H}\underset{\sim}{H}^{-1}(\underset{\sim}{\tilde{F}}\underset{\sim}{Z} - \underset{\sim}{\tilde{F}}\underset{\sim}{F}\underline{a}) ,$$

$$\mathcal{M}^* = (\underline{Z} - \underline{\bar{Z}})(\underset{\sim}{1} - \underset{\sim}{S})\widetilde{(\underset{\sim}{Z} - \underset{\sim}{\bar{Z}})} \quad \text{where } \underline{a} \cdot \underset{\sim}{\tilde{F}} = \underset{\sim}{\bar{Z}} \text{ and}$$

$$\underset{\sim}{S} \equiv \underset{\sim}{F}\underset{\sim}{H}^{-1}\underset{\sim}{\tilde{F}} . \tag{35}$$

Note that

$$\underset{\sim}{S}^2 = (\underset{\sim}{F}\underset{\sim}{H}^{-1}\underset{\sim}{\tilde{F}})(\underset{\sim}{F}\underset{\sim}{H}^{-1}\underset{\sim}{\tilde{F}}) = \underset{\sim}{F}\underset{\sim}{H}^{-1}\underset{\sim}{\tilde{F}} = \underset{\sim}{S}.$$

If s_i is an eigenvalue of $\underset{\sim}{S}$, it must equal s_i^2, an eigenvalue of S^2. Thus $s_i = 0$ or 1. The trace of S is

$$\mathrm{Tr}S = \sum_{a,b,c} F_{ab}H_{bc}^{-1}\tilde{F}_{ca} = \sum_{b,c} H_{cb}H_{bc}^{-1} = \mathrm{Tr}\,\underset{\sim}{I} = M .$$

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 S are one, and (p-M) are zero. Let $\underset{\sim}{U}$ be the unitary

matrix which diagonalizes $\underset{\sim}{S}$ (and also $(\underset{\sim}{1}-\underset{\sim}{S})$). According to Eq. (35),

$$\mathcal{M}^* = \underline{\eta} \cdot \underset{\sim}{U}(\underset{\sim}{1}-\underset{\sim}{S})\,\underset{\sim}{U}^{-1} \cdot \tilde{\underline{\eta}}, \quad \text{where } \underline{\eta} \equiv (\underline{Z}-\bar{\underline{Z}}) \cdot \underset{\sim}{U},$$

$$= \sum_{a=1}^{p} \mathcal{m}_a \eta_a^2 \quad \text{where } \mathcal{m}_a \text{ are the eigenvalues of } (\underset{\sim}{1}-\underset{\sim}{S}).$$

$$= \sum_{a=1}^{p-M} \eta_a^2 \quad \text{since the M nonzero eigenvalues of } \underset{\sim}{S}$$

cancel out M of the eigenvalues of $\underset{\sim}{1}$.

Thus

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

where $\mathcal{M}^*$ is the square of the radius vector in (p-M)-dimensional space. By definition (see Section 17) this is the χ^2 distribution with (p-M) degrees of freedom.

Information Division
br

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission to the extent that such employee or contractor prepares, handles or distributes, or provides access to, any information pursuant to his employment or contract with the Commission.

26.7.84