ANALYSIS OF EXPERIMENTS IN PARTICLE PHYSICS

By Frank T. Solmitz

Lawrence Radiation Laboratory, University of California, Berkeley, California

CONTENTS

Introduction

The field of high energy physics has grown at a tremendous rate in the last few years. Not only has the number of experiments been increasing, but so also have the amount and complexity of the data obtained in individual experiments. The analysis of such vast quantities of data has been made feasible by the increasing availability of high-speed digital computers. This article reviews some of the methods currently in use, as well as the principles on which they are based.

In the author's experience many experimentalists working in high energy physics have had very little, if any, formal training in the theory of probability and statistics. They may, however, be quite adept and experienced in the manipulation of distributions. (In fact they often rediscover theorems well known to the statisticians.) It therefore seems essential to start with some discussion, necessarily extremely sketchy, of the underlying ideas.

In essence, science consists in learning from experience: our observations lead us to make statements about nature. The process is known as induction. It is utterly and completely different from deduction, the principal tool of mathematics. Observations may lead one to believe—more or less strongly—

certain laws of nature, but such laws cannot be proved in the sense that a mathematical theorem can be proved on the basis of a set of axioms. Induction is basically the application of common sense. Our belief in a theory is generally strengthened if our observations are a probable (or certain) consequence of the theory; it is weakened if the observations are improbable on the basis of the theory. Our observations thus allow us to arrange in order our relative degrees of belief in various theories. This is, of course, generally not done on the basis of the observations of a single experiment; a large amount of prior knowledge also comes into play. The formal statement of the principle involved was given by Bayes in the 18th century. The fact that the principle was given a mathematical form did not reduce induction to deduction. An attempt to do so would be absurd. Equally absurd would be attempts to eliminate induction, since the pursuit of science is impossible without it.

There is, unfortunately, a great deal of controversy among the experts in the field of probability and statistics about both the basic principles and the methods to be used. One might crudely divide the different schools into two groups, the "Bayesians" and the "anti-Bayesians."

The Bayesian builds induction into the structure of his theory. This may be done by introducing the concept of "degree of belief" and giving rules governing it [see, e.g., Jeffreys (1)]. Alternatively, one can formulate theories of "inductive behavior," which give rules of action on the basis of all the available information; such theories of behavior can be applied to many fields such as (for instance) economics or agriculture, but generally not to physics. Anti-Bayesians give rules by which a scientist can combine and distill his data to the point where the inference is "obvious." They would have the scientist lead his readers to the brink of induction, but then let each reader take the plunge in his own way. This may seem to be the coward's approach, but it does have a great practical advantage: It allows the scientist to limit the scope of his discussion to his own experiment and things which he feels have an immediate bearing on it; it leaves each reader free to evaluate the results in the light of his own knowledge and experience.

1. Review of Properties of Some Distributions

1.1 *General remarks and definitions.*—Our basic notions of probability distributions are tied to the idea of relative frequency. When we say that a tossed coin has a 50 percent probability of coming up "heads," we imagine that in a large series of tosses this coin would come up heads about half the time. In the modern form of the calculus of probability, one constructs an abstract definition which endows the probability P with the kind of properties one associates with relative frequency:

(*a*) P is a positive set function $0 \leq P \leq 1$. In the case of the coin toss, the set has two "elements," heads and tails.

(*b*) P is additive: $P(A \text{ or } B) = P(A) + P(B)$: the probability that in the throw of a die either the "one" or the "two" will come up is $\frac{1}{3}$, if the probability that the "one" will come up is $\frac{1}{6}$ and that the "two" will come up is $\frac{1}{6}$.

(*c*) The probability for the whole set is unity, i.e., P is normalized.

A variable that takes on a distinct numerical value for each element of the set (e.g., the number of dots on each of the six sides of a die) is called a random variable. In what follows we shall speak only of such random variables and their probability distributions. A random variable may range over a discrete spectrum or a continuous spectrum, or a spectrum which is partly discrete and partly continuous. To treat the distributions of random variables with rigor and generality takes a rather formidable mathematical apparatus. It seems, however, that very little if anything of practical importance is lost by sticking to the rather crude and simple approach customary among physicists. For a single continuous random variable x, we say that $P(x)dx$ is the probability that the variable lies in the interval dx at x. If the variable ranges over the interval from a to b, the normalization condition is clearly $\int_a^b P(x)dx = 1$. We shall call $P(x)$ the probability density.

Suppose that we have two random variables x and y, with x lying between a and b and having a probability density $P(x)$, and y between a' and b' with probability density $Q(y)$; we shall assume that they obey a joint distribution, characterized by a joint probability density, say $R(x, y)$, such that $R(x, y)dx\,dy$ is the probability that the variables lie in the area element $dxdy$ at the point (x, y). We then require that R have the properties

$$\int_{a'}^{b'} R(x, y)dy = P(x), \qquad \int_a^b R(x, y)dx = Q(y)$$

The normalization of R follows from the normalization of P (or Q):

$$\int_a^b \int_{a'}^{b'} R(x, y)dxdy = \int_a^b P(x)dx = 1$$

We define the conditional probability density of y, given x, $S(y|x)$, by the relation $S(y|x)P(x) = R(x, y)$. Similarly we define $T(x|y)$, the conditional probability density of x, given y, by $T(x|y)Q(y) = R(x, y)$. From these definitions we immediately obtain the relationship known as Bayes' theorem:

$$T(x \mid y) = \frac{S(y \mid x)P(x)}{Q(y)} = \frac{S(y \mid x)P(x)}{\int S(y \mid x)P(x)dx} \qquad 1.$$

We shall see below how "Bayesians" apply this relationship to the problem of induction.

Two random variables, x and y, are said to be statistically independent if their joint probability density $R(x, y)$ can be written as a product,

$$R(x, y) = P(x)Q(y)$$

This "product rule" clearly cannot be proved, since it serves to define the concept of statistical independence. The definition is, however, in accord with our primitive concept. Two independent dice would come up snake eyes $\frac{1}{36}$ of the time if each one has a probability of $\frac{1}{6}$ of coming up on the "one"; by "independent" we mean here that the two dice do not exert any influence on each other. One of the important things in the analysis of any experiment is to try to establish which variables can be considered independent in this sense.

All the definitions made so far for two random variables can be extended in an obvious way to any number of random variables; they also apply to discrete variables if we replace integrals by sums and "probability density" by "probability." One often encounters distributions in which some of the variables are discrete and some are continuous; this also does not cause any complications.

A function, $y(x)$, of a random variable x is also a random variable. Two such variables are said to be functionally dependent (as opposed to statistically dependent). It should be noted that the joint probability density is not really a well-defined concept for two functionally dependent variables, since the density would have to vanish everywhere in the $x-y$ plane except on the line $y(x)$. If we are given the probability density $P(x)$ for x, we obtain the probability density $Q(y)$ for y by requiring that $Q(y)dy=P(x)dx$ or $Q(y)=P(x)|dx/dy|$. (This has to be modified slightly if x is not a single-valued function of y.) We generalize the above in the usual way to the transformation of r variables: If $y_1, y_2, \cdots, y_r$ are functions of $x_1, x_2, \cdots, x_r$, then

$$Q(y_1, y_2, \cdots, y_r) = P(x_1, x_2, \cdots, x_r) \left| J\left(\frac{x_1, x_2, \cdots, x_r}{y_1, y_2, \cdots, y_r}\right) \right|$$

Here P and Q are the joint probability densities for the x_i and y_i, respectively, and the Jacobian J gives the transformation of the volume element in the r-dimensional space.

1.2 *Mean, moments.*—We define the mean or expectation value of a random variable x with probability density $P(x)$ as $\langle x\rangle=\int xP(x)dx$. Similarly, the mean of some function $f(x)$ is defined by $\langle f(x)\rangle=\int f(x)P(x)dx$. The quantity $\langle x^n\rangle$ is called the nth moment of the distribution; the nth moment about the mean, $\langle(x-\langle x\rangle)^n\rangle$, is called the nth central moment. The variance of x, V_x, is the second central moment:

$$V_x = \langle(x - \langle x\rangle)^2\rangle = \langle x^2\rangle - \langle x\rangle^2$$

The standard deviation σ_x is defined as $\sigma_x=(V_x)^{1/2}$. In a distribution in r variables $x_1, x_2, \cdots x_r$, we define the mean of x_k by

$$\langle x_k\rangle = \int\int\cdots\int dx_1dx_2\cdots dx_rx_kP(x_1, x_2, \cdots, x_r)$$

The generalization of the variance is the moment matrix M:

$$\begin{aligned} M_{kl} &= \langle(x_k - \langle x_k\rangle)(x_l - \langle x_l\rangle)\rangle \\ &= \int\int\cdots\int dx_1\cdots dx_r(x_k - \langle x_k\rangle)(x_l - \langle x_l\rangle)P(x_1, \cdots, x_r) \\ &= \langle x_kx_l\rangle - \langle x_k\rangle\langle x_l\rangle \end{aligned}$$

Here M is symmetric and nonnegative; its r diagonal terms are the variances, its off-diagonal terms the covariances. The correlation coefficient ρ_{kl} for the variables x_k, x_l is defined by $\rho_{kl} = M_{kl}/(M_{kk}M_{ll})^{1/2} = M_{kl}/\sigma_k\sigma_l$. From the definition of M and the well-known Schwarz inequality, one deduces that $|\rho_{kl}| \leq 1$. The equality holds if and only if x_k is a linear function of x_l; in that case we are dealing with a distribution in only $(r-1)$ random variables; the moment matrix is then singular, since two of its rows are proportional. More generally, if there are s linear relations among the x_k, the rank of M is $(r-s)$.

Consider S linear functions $y_1, y_2, \cdots, y_s$ of the N random variables $x_1, x_2, \cdots, x_N$,

$$y_i = \sum_{j=1}^{N} T_{ij}x_j + a_i, \qquad \text{for } 1 \leq i \leq S$$

or, in matrix notation,

$$\boldsymbol{y} = \boldsymbol{T}\cdot\boldsymbol{x} + \boldsymbol{a} \qquad 2.$$

Now the moments of the y's are clearly linear functions of the moments of the x's; for the mean we have

$$\langle \boldsymbol{y} \rangle = \boldsymbol{T}\cdot\langle \boldsymbol{x} \rangle + \boldsymbol{a} \qquad 3.$$

The moment matrix of the y's, say $\boldsymbol{M}_y$, is given by

$$\begin{aligned} \boldsymbol{M}_y &= \langle(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)\cdot(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)^{\dagger}\rangle \\ &= \boldsymbol{T}\cdot(\boldsymbol{x} - \langle \boldsymbol{x} \rangle)\cdot(\boldsymbol{x} - \langle \boldsymbol{x} \rangle)^{\dagger}\cdot\boldsymbol{T}^{\dagger} \end{aligned}$$

or

$$\boldsymbol{M}_y = \boldsymbol{T}\cdot\boldsymbol{M}_x\cdot\boldsymbol{T}^{\dagger} \qquad 4.$$

(here $\boldsymbol{M}_x$ is the moment matrix of the x's, and the dagger indicates transposition). If the y's are functions of the x's that are nearly linear in that part of the space where $P(x)$ is large, Equation 4 may sometimes give an adequate approximation to the moment matrix $\boldsymbol{M}_y$, if one takes for T_{ij} the first partial derivatives $(\partial y_i/\partial x_j)$ evaluated at or near the mean $\langle \boldsymbol{x} \rangle$; this approximation is known to physicists as the propagation of errors.

If the x_i are statistically independent, then $\boldsymbol{M}_x$ is diagonal, and the variables of the y_j become linear functions of the variances of the x_i:

$$V_{yi} = (M_y)_{ii} = \sum_{j=1}^{N} (T_{ij})^2(M_x)_{jj} = \sum_{j=1}^{N} (T_{ij})^2 V_{xj}$$

or, in terms of standard deviations,

$$\sigma_{yi}{}^2 = \sum_{j=1}^{N} (T_{ij})^2\sigma_{xj}{}^2 \qquad 5.$$

Consider now N statistically independent random variables $x_1, \cdots, x_N$, each obeying the same distribution (the x_i might, for instance, represent N independent repetitions of a measurement of some quantity). The linear function

$$\xi = (1/N)\sum_{i=1}^{N} x_i$$

is known as the "sample mean"; it is a random variable and hence quite different in nature from a mean such as $\langle x\rangle$, which is a constant. By setting $T_{ij}=(1/N)$, $a_i=0$, and $\xi=y_1$, we get, from Equations 3 and 4,

$$\langle\xi\rangle = \langle x\rangle \tag{6.}$$

$$\sigma_\xi^2 = \frac{1}{N}\sigma_x^2 \tag{7.}$$

The standard deviation σ_ξ is a measure of the spread of the distribution of ξ. Since it decreases as $(N)^{-1/2}$, the probability density for ξ becomes more concentrated near the mean $\langle x\rangle$ as N increases. In other words, for large N, the sample mean ξ is very likely to be near $\langle x\rangle$. We can easily generalize these results to the case in which the ith "observation" consists of several random variables x_i, y_i, z_i, $\cdots$. The quantities

$$\xi = (1/N)\sum_{i=1}^{N} x_i, \qquad \eta = (1/N)\sum_{i=1}^{N} y_i, \qquad \zeta = (1/N)\sum_{i=1}^{N} z_i, \quad \text{etc.}$$

have the means $\langle\xi\rangle=\langle x\rangle$, $\langle\eta\rangle=\langle y\rangle$, $\langle\zeta\rangle=\langle z\rangle$, etc.; the moment matrix for the variables ξ, η, ζ, $\cdots$ is just $1/N$ times the moment matrix for x, y, z, $\cdots$. This can be shown by application of Equations 3 and 4.

1.3 *Multinomial and Poisson distributions.*—Let us illustrate the ideas developed in the preceding section with the help of a few well-known examples. Consider, first, a reaction which can go into r different channels, and has probability p_i of going into the ith channel

$$\left(\sum_{i=1}^{r} p_i = 1\right)$$

Suppose we observe N examples of the reaction; the probability of having n_1 going into the first channel, n_2 into the second, etc., is given by the multinomial distribution

$$P_{\text{mult}}(n_i) = \frac{N!}{\prod_{i=1}^{r} n_i.}\prod^{r} p_i^{n_i} \tag{8.}$$

We must consider P_{mult} as a function of only $(r-1)$ random variables, since the r variables n_i are functionally related,

$$\left(\sum_{i=1}^{r} n_i = N\right)$$

The form of P is derived by repeated application of the product rule and use of the additive property of probability. We have, for the means and moment matrix of the n_i,

$$\langle n_i\rangle = Np_i \tag{9.}$$

$$\langle(\delta n_i)(\delta n_j)\rangle = Np_i(\delta_{ij}-p_j), \qquad 1\le i,\ j\le r \tag{10.}$$

where

$$\delta n_i \equiv n_i - \langle n_i \rangle$$

These results are easily verified for the trivial case of $N=1$, i.e., a single observation; they are then directly generalized to any N by use of the results about repeated observations developed in the preceding section. The moment matrix is singular, since the n_i are linearly dependent.

In an experimental run of given duration, the probability of observing some reaction N times is usually assumed to take the Poisson form,

$$P_{\text{Pois}}(N) = \frac{A^N e^{-A}}{N!}$$

The Poisson law is an approximation to the binomial distribution which is good when the maximum possible value of N is very large compared with the expected value; even when this condition is satisfied, the Poisson law may not apply because the basic assumption of the statistical independence of the individual observations is not valid (as, for instance, in a counter experiment in which the resolving time is not negligible compared with the average time between counts).

The probability of observing n_1 reactions going into the first channel, n_2 into the second, etc., may be expressed as the product of the probability for observing a total of N reactions and the conditional probability, given N, of having a division into n_1, n_2, etc.:

$$\begin{aligned} P(n_1, n_2, \cdots, n_{r-1}, N) &= P_{\text{mult}}(n_1, n_2, \cdots, n_{r-1} \mid N) P_{\text{Pois}}(N) \\ &= \frac{N!}{\prod_{i=1}^{r} n_i!} \prod^{r} p_i^{n_i} \cdot \frac{A^N e^{-A}}{} \\ &= \prod_{i=1}^{r} \frac{(Ap_i)^{n_i} e^{-(Ap_i)}}{n_i!} \end{aligned}$$

we see here that P can also be written in the form of a product of Poisson distributions, as one would expect.

1.4 *Normal and χ^2 distributions.*—One can show, under quite general conditions, that the distribution of the sum of N random variables approaches the Gaussian or "normal" form as $N \rightarrow \infty$; this result, familiar to physicists from the theory of the random walk, is known in statistics as the central limit theorem.[1] In many applications in which N is large though finite, the Gaussian or normal distribution provides an adequate approximation.

The probability density for the most general normal distribution in n variables x_i can be written as

$$P(x) = C \cdot \exp\left[-(1/2)(x-a)^{\dagger} \cdot B \cdot (x-a)\right], \quad -\infty \leq x_i - \infty \qquad 11.$$

[1] For derivation of this and other results in the theory of statistics, the reader should consult the texts of Cramer (2) and of Kendall & Stuart (3).

here $\boldsymbol{x}$ and $\boldsymbol{a}$ are r-component column vectors and B is a positive definite symmetric n-by-n matrix (if B were not positive definite, the normalization integral would diverge). Since P is an even function of $(\boldsymbol{x}-\boldsymbol{a})$ we must have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} d^nx(x-a)P(x-a)=0 \qquad 12.$$

hence

$$\langle(\boldsymbol{x}-\boldsymbol{a})\rangle = 0, \quad \text{or} \quad \langle\boldsymbol{x}\rangle = \boldsymbol{a} \qquad 13.$$

We can obtain the moment matrix by differentiating Equation 12 with respect to $\boldsymbol{a}$; we obtain

$$\int\int\cdots\int d^nx[-I+(x-a)\cdot(x-a)^{\dagger}\cdot B]P(x-a)=0$$

or

$$-\boldsymbol{I}+\langle(\boldsymbol{x}-\boldsymbol{a})\cdot(\boldsymbol{x}-\boldsymbol{a})^{\dagger}\rangle\cdot\boldsymbol{B}=0$$

hence

$$\langle(\delta\boldsymbol{x})(\delta\boldsymbol{x})^{\dagger}\rangle = B^{-1}, \quad \text{where} \quad \delta\boldsymbol{x}=\boldsymbol{x}-\boldsymbol{a} \qquad 14.$$

Any set $y_1, \cdots, y_s$ of linear independent functions of the n variables is also normally distributed. We briefly outline the proof: For $s=n$, we are dealing with a transformation of variables; the quadratic form in the exponent of Equation 11 goes over into a quadratic form of the y_i; the Jacobian is constant and can therefore be absorbed into the normalization constant. If we are dealing with fewer than n variables y_i, they can always be considered a subset of a complete set of n variables. Such a subset of Gaussian variables is itself a set of Gaussian variables; to show this we must integrate the probability density over the remaining $(n-s)$ variables. Each integral can be carried out by completing the square in the exponent, and this procedure always leaves a quadratic form of the remaining variables in the exponent. Fortunately we do not need to carry out this procedure to obtain the probability density, say $Q(y)$, of the variables y_i, since the form of Q is completely determined by the means and the moment matrix of the y_i's. Let us set

$$\boldsymbol{y}=\boldsymbol{T}\cdot\boldsymbol{x}+\boldsymbol{b}$$

then

$$\langle\boldsymbol{y}\rangle=\boldsymbol{T}\cdot\boldsymbol{a}+\boldsymbol{b}$$

and

$$\langle(\delta\boldsymbol{y})(\delta\boldsymbol{y})^{\dagger}\rangle=\boldsymbol{T}\cdot B^{-1}\cdot\boldsymbol{T}^{\dagger}=\boldsymbol{H}^{-1}$$

hence we have

$$Q(\boldsymbol{y})=C'\exp\left[-(1/2)(\boldsymbol{y}-\langle\boldsymbol{y}\rangle)^{\dagger}\cdot\boldsymbol{H}\cdot(\boldsymbol{y}-\langle\boldsymbol{y}\rangle)\right]$$

We are often interested in the distribution of the quadratic form in the exponent of Equation 11. Let us first consider the case in which $\boldsymbol{a}=0$, $B=I$); then

$$P(x_i) = C \exp\left[-(1/2)\chi^2(x_i)\right] \qquad 15.$$

with

$$\chi^2 = \sum_{i=1}^{n} x_i^2 \qquad 16.$$

Here the x_i are independent Gaussian variables with mean $\langle x_i\rangle=0$ and standard deviation $\sigma_i=1$. We can consider χ to be the length of the radius vector in the n-dimensional state of the x_i. The volume element is then proportional to $\chi^{n-1}d\chi$, or $(\chi^2)^{(n/2)-1}\cdot d(\chi^2)$, if we want to express it in terms of the variable χ^2. We have, therefore,

$$P_n(\chi^2)d(\chi^2) \propto e^{-\chi^2/2}(\chi^2)^{(n/2)-1}d(\chi^2)$$

or

$$P_n(\chi^2) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} e^{-\chi^2/2}(\chi^2)^{(n/2)-1} \qquad 17.$$

the normalization constant follows directly from the definition of the function. One finds, for the mean and the standard deviation,

$$\langle\chi^2\rangle = n, \qquad \sigma_{\chi^2} = (2n)^{1/2} \qquad 18.$$

this last result can be shown most simply by appealing to the general results for sums of random variables developed in Section 1.2. If n is large, then P_n is nearly Gaussian; this follows from the central limit theorem, but can also be shown directly by asymptotic expansion of Equation 17.

If we now consider the general case of the n-dimensional Gaussian, Equation 11, we can show that the quantity

$$\chi^2 = (x - \boldsymbol{a})^{\dagger}\cdot B\cdot(x - \boldsymbol{a}) \qquad 19.$$

also has the probability density given by Equation 17: we can always construct a linear transformation of the variables x_i so that χ^2 will take on the form of Equation 16; since the Jacobian is constant, P will take the form of Equation 15, and the result follows.

2. Estimation of Parameters

We now return to the central problem: what can we say about nature after we have made some experimental observations? Very often the problem presents itself in the following way: we know the form of the statistical law which the observations must follow, but the law may contain some unknown or poorly known parameters; given some observations, what can we then say about these parameters? This problem of estimation of parameters has been treated at great length in the statistical literature [for an extensive treatment and bibliography, see Kendall & Stuart (3)]. We shall confine ourselves to a brief discussion of the underlying ideas involved in various different approaches, and then present some illustrative examples. In the following we shall denote the observations of a single experiment by x_i and the parameters by α. The joint probability density $L(x_i; \alpha)$ for the experiment is usually referred to as the likelihood function.[2]

We shall use as our "standard example" an experiment consisting of N independent observations $x_1, x_2, \cdots, x_N$ of some quantity α (say, for instance, the length of some object); suppose that the resolution of our measuring instrument is known to be Gaussian in form with a standard deviation σ. Then the likelihood function for the experiment has the form

$$L(x_i; \alpha) = c \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \alpha)^2\right] \qquad 20.$$

2.1 *Bayesian approach.*—The Bayesian postulates that at any stage his relative strength of belief in various possible values of α may be described by a probability density $P(\alpha)$. Such a probability does not have a frequency interpretation, but it may still be assumed to obey the axioms of probability distributions in general. The likelihood function can then be regarded as a conditional probability density $L(x_i | \alpha)$ for the x_i, given α. If our particular experiment yields a particularly set of observations x_i^{obs}, we can express our strength of belief after the experiment with the help of Bayes' theorem, Equation 1:

$$P_{\text{new}}(\alpha \mid x_i^{\text{obs}}) \propto L(x_i^{\text{obs}} \mid \alpha) \cdot P_{\text{old}}(\alpha) \qquad 21.$$

where P_{old} describes our strength of belief prior to the experiment, P_{new} that after the experiment. Here P_{new} is seen to be a conditional probability, since it depends on the given observations x^{obs}.[3] Equation 21 is the formal statement of Bayes' principle. It illumines the fact that different physicists faced with the same experimental observations may legitimately come to different conclusions: they have different "prior" knowledge and therefore assess the prior probability P_{old}, differently. How then can an experimenter present the results of his work in an "objective" fashion, that is, without introducing his own prior beliefs? One way, often used by physicists, is to present $L(x_i^{\text{obs}}; \alpha)$ as a function of α for his particular observations x_i^{obs}; it is then left up to each reader to apply Bayes' principle in his own way, that is, to put in his own knowledge or prejudice.

There is often a temptation to go one step farther and to try to give the conclusions of the experiment, assuming "complete ignorance" of α prior to the experiment. In order to do this we have to give P_{old}, corresponding to to "complete ignorance." There is, however, no universal prescription for

[2] We assume that α can take on a continuum of values. The x_i can be either discrete or continuous random variables; if all the x_i are discrete variables, then L is a probability rather than a probability density.

[3] The quantity P_{old} depends, of course, on previous knowledge and observations; it therefore seems best to abandon the rather misleading term of "a priori probability" which has often been used [see Jeffreys (1)].

doing this. One rule that tends to suggest itself is to say that ignorance corresponds to the assumption that P_{old} is constant over the interval in which α is defined. This rule leads to the following difficulty: Suppose we were to state the likelihood in terms of some other parameter β, which is some nonlinear function of α. Then we must require, for consistency, that the prior probability for β, say Q_{old}, is given by $P_{\text{old}}(\alpha)d\alpha = Q_{\text{old}}(\beta)d\beta = Q_{\text{old}}(\beta)(d\beta/d\alpha)d\alpha$. Since we cannot have both $P_{\bullet\text{ld}}$ and Q_{old} constant, we see that according to the above rule, ignorance of β leads to different results. Jeffreys (1) has tried to formulate rules which partially meet this and other objections. One may, however, question the necessity or even the desirability of giving a precise definition to a concept as vague as ignorance. The crucial point is this: An experiment that gives a "good" measurement of α is one for which $L(x_i^{\text{obs}}|\alpha)$ is a sharply peaked function of α; if we were relatively ignorant of α before the experiment, we imagine $P_{\text{old}}(\alpha)$ to be a relatively smooth function, and our new belief, $P_{\text{new}}(\alpha|x_i^{\text{obs}})$, dominated by the peak in the likelihood function.

In our standard example of N repeated measurements of α, $L(x_i^{\text{obs}};\alpha)$ is easily seen to be a Gaussian function of α with its peak at

$$(1/N)\sum_{i=1}^{N} x_i^{\text{obs}}$$

and its width given by the standard deviation $\sigma/N^{1/2}$. The width of L does, therefore, decrease with increasing N. Consider now the more general case of x_i independent observations, each of which has the the probability density $P(x_i;\alpha)$. Then the likelihood function is $L(x_i;\alpha) = \Pi_{i=1}^{N} P(x_i;\alpha)$. Now one can show that under quite general conditions $L(x_i;\alpha)$ is very likely to be approximately Gaussian in α for sufficiently large N.[4] Take, for example, the measurement of the lifetime of an unstable particle: the probability density for observing a time t_i for the ith event is $P(t_i;\tau) = \tau^{-1}e^{(t_i/\tau)}$, s● that the likelihood function is

$$L(t_i;\tau) = \tau^{-N}e^{-T/\tau} \qquad 22.$$

with

$$T \equiv \sum_{i=1}^{N} t_i, \qquad \text{for } 0 \leq t_i < \infty$$

(We ignore, for simplicity, effects due to the finite time of observation, uncertainty in t_i, etc.) For some particular value of T, say T_{obs}, and for large N we can expand Equation 22 asymptotically; we find that it is approximately a Gaussian with its maximum at $\tau_{\text{max}} = (1/N)T_{\text{obs}}$, and a width (standard deviation) equal to $\tau_{\text{max}}/N^{1/2}$. Re-expressed in terms of the decay rate λ, the likelihood function takes on the form

$$L'(t_i;\lambda) = \lambda^N \exp(-\lambda T)$$

[4] This result, though well known to physicists, is not found in this form in the standard texts on statistics, since anti-Bayesians attach no particular significance to the shape of $L(x_i^{\text{obs}};\alpha)$.

where $\lambda = 1/\tau$. The quantitities L and L' are of course quite different functions of their respective parameters, although they contain exactly the same information. For sufficiently large N, L' is also nearly Gaussian. If N is not very large, it turns out that $L'(\lambda)$ is more nearly Gaussian than $L(\tau)$; therefore, if we wish to characterize the shape of the likelihood function by the position of its maximum and its width, this is better done in terms of λ. Alternatively, one can often adequately characterize L by three values of the parameter: the value at which L has its maximum and the two values at which it has dropped by a factor of $(e)^{1/2}$, that is, the two points which correspond to $(\mu \pm \sigma)$ for a Gaussian of mean μ, standard deviation σ.

Likelihood functions that depend on several parameters α_λ may often be adequately described by giving (a) the point where L has its maximum, and (b) the shape of L near that maximum; if L is roughly Gaussian near its maximum we can specify the shape in terms of the moment matrix,[5] that is, the inverse of the matrix $[-(\partial^2 \ln L)/(\partial\alpha_\lambda \partial\alpha_\mu)]_{\alpha_{\max}}$. Physicists often refer to this moment matrix for the parameters as the error matrix. If L is relatively large in several distinct regions of the space of the α_λ, we can describe it by giving the positions and heights of all the important maxima and the corresponding error matrices. Here L may be large on or near the boundary of the "physical region," that is, the region in which the α_λ are physically meaningful; even in that case it may be expedient to give the position and height of the maximum, which may be outside the physical region, and the error matrix.

Bayes' principle, Equation 21, leads to a simple and consistent method of combining independent experiments: Suppose a second experiment is described by a likelihood function L' and has yielded a set of observations y_i^{obs}, then we must again revise P_{new}, the belief we had after the first experiment, to obtain P_{newest}:

$$\begin{aligned} P_{\text{newest}}(\alpha \mid y_i^{\text{obs}}, x_i^{\text{obs}}) &\propto L'(y_i^{\text{obs}} \mid \alpha) P_{\text{new}}(\alpha \mid x_i) \\ &\propto L'(y_i^{\text{obs}} \mid \alpha) L(x_i^{\text{obs}} \mid \alpha) P_{\text{old}}(\alpha) \end{aligned}$$

Had we considered the two experiments as one composite experiment, we would have formed $L'L$, the joint probability density for the set of observation x_i^{obs}, y_i^{obs}, and would therefore have arrived at the same result. This rule for combining experiments must be used with caution, because of possible uncertainties common to the experiments. Thus, for instance, in the determination of the lifetime of a particle we have to know the mass in order to calculate the proper times t_i; several experiments using the same mass determination in their analysis are then subject to a common uncertainty.

The combination of several experiments becomes particularly simple if all the likelihood functions are Gaussian functions of the parameters, since

[5] Since $L(x_i \mid \alpha_\lambda)$ is not a probability density for αr_λ, the term "moment matrix" is here used in a somewhat extended sense.

a product of Gaussians is itself a Gaussian. (This is one reason for trying to pick the parameters in a way that makes the likelihood functions appear nearly Gaussian.) Consider the case of two likelihood functions with peaks at $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ and error matrices $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$, respectively:

$$L_1 \propto \exp\left[-\tfrac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_1)^{\dagger}\cdot \boldsymbol{\varepsilon}_1^{-1}\cdot(\boldsymbol{\alpha} - \boldsymbol{\alpha}_1)\right]$$
$$L_2 \propto \exp\left[-\tfrac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_2)^{\dagger}\cdot \boldsymbol{\varepsilon}_2^{-1}\cdot(\boldsymbol{\alpha} - \boldsymbol{\alpha}_2)\right]$$

Then, if we form the product and regroup the terms appropriately, we find

$$L_1L_2 \propto \exp\left[-\tfrac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{1,2})^{\dagger}\cdot \boldsymbol{\varepsilon}_{1,2}^{-1}\cdot(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{1,2})\right. \qquad 23.$$

where

$$\boldsymbol{\alpha}_{1,2} = \boldsymbol{\varepsilon}_{1,2}(\boldsymbol{\varepsilon}_1^{-1}\cdot \boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon}_2^{-1}\cdot \boldsymbol{\alpha}_2)$$

and

$$\boldsymbol{\varepsilon}_{1,2}^{-1} = \boldsymbol{\varepsilon}_1^{-1} + \boldsymbol{\varepsilon}_2^{-1}$$

The extension to more than two experiments is immediate. For a single parameter, Equation 23 reduces to the familiar inverse-square weighting rule; with $\epsilon_1 = \sigma_1^2$ and $\epsilon_2 = \sigma_2^2$ we get

$$\alpha_{1,2} = \sigma_{1,2}^2\left(\frac{1}{\sigma_1^2}\alpha_1 + \frac{1}{\sigma_2^2}\alpha_2\right)$$

2.2 *Anti-Bayesian approach—point estimation.*—The problem of estimation can be put in the following way. Let us try to find some function α^* of the random variables x_i which is likely to represent a "good estimate" of the true value of the parameter α. Since $\alpha^*(x_i)$ is a function of random variables, it is itself a random variable; that means that if we were to consider a series of experiments performed under identical conditions, α^* would take on a series of values. The probability density of any such function can be derived from that of the x_i, that is, from the likelihood function. A particular $\alpha^*(x_i)$ is considered to furnish a good estimate if its probability density $g(\alpha^*;\,\alpha)$ is highly concentrated near α for all admissible values of α. The result of a particular experiment which has yielded the particular observations x_i^{obs} is then "summarized" in $\alpha_{\text{obs}}^* = \alpha^*(x_i^{\text{obs}})$. We can thus avoid use of Bayes' principle, by simply confining ourselves to the statement of α_{obs}^* and $g(\alpha^*;\,\alpha)$; but we do this at the cost of not saying anything about the true value of the parameter. If we did want to conclude something about the true value of α, we would have to say that since the probability density of α^*, $g(\alpha^*;\,\alpha)$, is strongly peaked about α, α_{obs}^* is likely to be close to the true value of α; therefore, the true value of α is likely to be close to α_{obs}^*. We are thus consciously or unconsciously applying Bayes' principle to the distribution of $\alpha^*(x_i)$ rather than directly to the distribution of the observations x_i, that is, the likelihood function.

Two principal criteria are usually applied in picking a "good" estimator α^*: (*a*) the σ of α^* should be as small as possible; (*b*) the "bias," i.e., the difference between $\langle\alpha^*\rangle$ and α, should be small compared with the σ. A third criterion, simplicity of the form of α^*, may also come into play. Consider our standard example of N repeated observations x_i of some quantity α. We might try, for simplicity, some linear function of the x_i say

$$\alpha^*(x_i) = \sum_{i=1}^{N} \mathrm{w}_i x_i$$

The bias and the σ of α^* depend on the particular choice of the weights w_i; it is easy to show that the choice $w_i = 1/N$ leads to zero bias and the smallest σ. Is it possible to find some other, nonlinear, estimator that is even better? The answer to this and similar questions is contained in a fundamental theorem[6] that puts a lower limit on the standard deviation of any unbiased estimator:

$$\frac{1}{\sigma_\alpha{}^{*2}} \leq \left\langle \left(\frac{\partial \ln L}{\partial \alpha} \right)^2 \right\rangle = - \left\langle \frac{\partial^2 \ln L}{\partial \alpha^2} \right\rangle \qquad 24.$$

If the likelihood function is of the form $L(x_i;\ \alpha) = \Pi \overset{N}{=} P(x_i;\ \alpha)$, the Equation 24 takes on the form

$$\frac{1}{\sigma_\alpha{}^{*2}} \leq N \int dx P(x) \left(\frac{\partial \ln P}{\partial \alpha} \right)^2 = - N \int dx P(x) \frac{\partial^2 \ln P}{\partial \alpha^2} \qquad 25.$$

An estimator that attains the minimum σ—that is, one for which the equality sign in Equation 24 (or Eq. 25) holds—is said to be an efficient estimator. Such efficient estimators exist only for a very limited class of problems: the likelihood function has to have a form such that

$$\frac{\partial \ln L(x_i;\ \alpha)}{\partial \alpha} = [f(x_i) - \alpha] k(\alpha) \qquad 26.$$

the function $\alpha^* = f(x_i)$ is then an efficient estimator of α. For the likelihood function of our standard example, Equation 20, one finds

$$\frac{\partial \ln L}{\partial \alpha} = \left(\frac{1}{N} \sum_{i=1}^{N} x_i - \alpha \right) \frac{N}{\sigma^2}$$

Therefore $\alpha^* = (1/N) \sum x_i$ is indeed an efficient estimator of α—there are no unbiased estimators with smaller σ. Similarly one readily verifies that $\tau^* = (1/N) \sum t_i$ is an efficient estimator for the lifetime problem, Equation 22.

One method of estimation with a number of desirable properties is known as the maximum-likelihood (m.l.) method: the m.l. estimator of α is that value of α—say α^*—for which the likelihood function takes on its greatest value. This method has such a close apparent similarity to the Bayesian approach that it is often confused with it; for in the Bayesian approach we would say that the most likely value of α must be near the point where L has its maximum, as long as our prior probability is reasonably flat. The intent of the m.l. method is, however, quite different from that of the Bayesian approach: The value α^*, at which $L(x_i;\ \alpha)$ has its maximum, is considered a random variable. In order to see whether α^* is a good estimator, we have to

[6] For proofs of this and other theorems quoted in this section, consult Cramer (2) or Kendall & Stuart (3).

examine its probability distribution; we pay no attention to the shape of L for our particular experiment. It can be shown that the m.l. estimator is the efficient estimator whenever one exists (this follows essentially from Eq. 26). Under very general conditions, the m.l. estimator is asymptotically unbiased, Gaussian, and efficient; that is, for sufficiently large N, the distribution of α^* is approximately Gaussian with its mean at α, and its σ is nearly equal to the smallest possible one.

In spite of the desirable properties just mentioned, the m.l. method is inadequate for many problems encountered in physics. For instance, in the partial-wave analysis of scattering processes we often find that the likelihood function has several maxima of comparable height. In that case, the various theorems about the m.l. method are of no help. If one were to follow the prescription of the m.l. method, one would report only the largest maximum of L; this is clearly an unsatisfactory way of summing up the conclusions of the experiment. Even when L possesses only one maximum, the m.l. estimate is not guaranteed to be nearly unbiased or nearly efficient when the number of observations is small.

2.3 *Confidence intervals.*—Suppose we have an estimator α^* and its probability density $g(\alpha^*; \alpha)$. Then we can, in general, construct two functions, $\alpha_1(\alpha^*)$ and $\alpha_2(\alpha^*)$, such that there is a fixed probability, say $(1-\epsilon)$, that the interval (α_1, α_2) includes the true value of α; i.e., if the experiment were repeated many times, the confidence interval (α_1, α_2) would include the true value of the parameter a fraction $(1-\epsilon)$ of the time. In our standard example, we know that the estimator $\alpha^* = (1/N) \sum x_i$ obeys the Gaussian distribution with mean α and $\sigma_{\alpha^*} = \sigma/N^{1/2}$; we can state that the probability that the interval $(\alpha_1 = \alpha^* - \sigma_{\alpha^*}, \alpha_2 = \alpha^* + \sigma_{\alpha^*})$ includes the true value of α is 68 percent (the area under a Gaussian up to σ). We could, therefore, report the result of our experiment by giving $\alpha_1^{\text{obs}} = \alpha^*_{\text{obs}} - \sigma_{\alpha^*}$ and $\alpha_2^{\text{obs}} = \alpha_{\text{obs}}^* + \sigma_{\alpha^*}$, that is, a confidence interval with a 68 percent confidence coefficient. We cannot assert without recourse to Bayes' principle that the true value of α has a probability of 68 percent of lying in the particular interval; we have formulated a rule by which we get the "right answer" (the true value inside the interval) 68 percent of the time, but we cannot say whether or not we have obtained the right "answer" in our particular experiment.

In principle we can construct a confidence interval even when the distribution of α^* is not Gaussian. We define $f_1(\alpha)$ and $f_2(\alpha)$ by

$$\int_{-\infty}^{f1} g(\alpha^*; \alpha) d\alpha^* = \frac{\epsilon}{2}$$

$$\int_{f2}^{\infty} g(\alpha^*; \alpha) d\alpha^* = \frac{\epsilon}{2}$$

If the observed value of the estimator is α^*_{obs}, the corresponding values α_1, α_2 specifying the confidence interval are given by the equations $\alpha_{\text{obs}}^* = f_1(\alpha_1)$, $\alpha_{\text{obs}}^* = f_2(\alpha_2)$ (the interval may be broken into several pieces if f_1 and f_2 are not monotonic functions of α). Different estimators for a given problem clearly lead to different confidence intervals. A good estimator—one with a relatively small σ—will, in general, lead to a shorter confidence interval than a poor estimator. The shorter interval has the same probability of including the true value as the longer one, because it fluctuates less.

It is generally not practical to calculate exact confidence intervals; however, Bartlett has given a good method of approximation and applied it to the problem of lifetime determination (4).

The confidence-interval (or confidence region) method is probably not so useful in really complicated problems in which the likelihood function contains many parameters and has a number of relative maxima. In any case, it is not a method suited to the combination of the results from several experiments.

3. Method of Least Squares

When the random variables x_i in an experiment are distributed according to the Gaussian law and their moment matrix G_{ij} is known, then the likelihood function takes on the form

$$L \propto \exp\left[-(1/2)\chi^2(x;\alpha)\right]$$

with

$$\chi^2 = \left[x - f(\alpha)\right]^\dagger \cdot G^{-1} \cdot \left[x - f(\alpha)\right]$$

Finding the minimum (or minima) of χ^2 is clearly equivalent to finding the maximum (or maxima) of L. Thus, in the case of Gaussian variables, the maximum-likelihood estimate leads directly to the well-known least-squares method. (Note that if the variables x_i are independent, then G is diagonal and χ^2 is a sum of squares.) The least-squares criterion may provide a reasonable method of estimation even if the x_i are not exactly Gaussian, and even if their moment matrix is only approximately known. It therefore has a vast number of applications in experimental physics to problems such as orbit-fitting, curve-fitting, mensuration and surveying, the determination of atomic constants (5), kinematic analysis of particle reactions (6), and analysis of angular distributions. We shall briefly review the method and a few of its applications.

3.1 *Linear problem.*—Let us first assume that the n quantities f_i are linear functions of the r parameters α_λ. Then χ^2 is a quadratic function of the parameters:

$$\chi^2 = (\xi - F\cdot\alpha)^\dagger \cdot G^{-1} \cdot (\xi - F\cdot\alpha)$$

where we have set

$$f = f_0 + F\cdot\alpha$$
$$x = f_0 + \xi$$

We also have, by assumption,

$$\langle\xi\rangle = F\cdot\alpha$$
$$\langle(\delta\xi)(\delta\xi)^\dagger\rangle = G$$

The values, say $\boldsymbol{\alpha}^*$, of the parameters that minimize χ^2 are obtained by setting the first derivatives of χ^2 equal to zero:

$$\nabla_\alpha \chi^2 = -2\boldsymbol{F}^\dagger \cdot \boldsymbol{G}^{-1} \cdot (\boldsymbol{\xi} - \boldsymbol{F} \cdot \boldsymbol{\alpha}^*) = 0$$

hence

$$\boldsymbol{\alpha}^* = \boldsymbol{H}^{-1} \cdot \boldsymbol{F}^\dagger \cdot \boldsymbol{G}^{-1} \cdot \boldsymbol{\xi}$$

where

$$\boldsymbol{H} = \boldsymbol{F}^\dagger \cdot \boldsymbol{G}^{-1} \cdot \boldsymbol{F}$$

Note that

$$\langle \boldsymbol{\alpha}^* \rangle = \boldsymbol{H}^{-1} \cdot \boldsymbol{F}^\dagger \cdot \boldsymbol{G}^{-1} \cdot \boldsymbol{F} \cdot \boldsymbol{\alpha} = \boldsymbol{\alpha} \qquad 27.$$

hence $\boldsymbol{\alpha}^*$ is an unbiased estimator.

The matrix $\boldsymbol{H}$ is nonsingular if, and only if, the f_i are linearly independent functions of the α_λ. If there is a linear dependence among the f_i, one clearly cannot determine all the parameters from the data. The estimates α_λ^* are linear functions of the Gaussian variables ξ_i, and are therefore also Gaussian; the moment matrix of the α^* is particularly simple:

$$\begin{aligned}(\delta \boldsymbol{\alpha}^*) \cdot (\delta \boldsymbol{\alpha}^*)^\dagger &= \boldsymbol{H}^{-1} \cdot \boldsymbol{F}^{+} \cdot \boldsymbol{G}^{-1} \cdot (\delta \boldsymbol{\xi}) \cdot (\delta \boldsymbol{\xi})^{+} \cdot \boldsymbol{G}^{-1} \cdot \boldsymbol{F} \cdot \boldsymbol{H}^{-1} \\ &= \boldsymbol{H}^{-1} \cdot \boldsymbol{F}^{+} \cdot \boldsymbol{G}^{-1} \cdot \boldsymbol{G} \cdot \boldsymbol{G}^{-1} \cdot \boldsymbol{F} \cdot \boldsymbol{H}^{-1} \\ &= \boldsymbol{H}^{-1} \cdot \boldsymbol{H} \cdot \boldsymbol{H}^{-1} \\ &= \boldsymbol{H}^{-1} \qquad 28.\end{aligned}$$

We can write $\chi^2(\boldsymbol{\alpha})$ as a sum of its value at the minimum, $\chi^2(\boldsymbol{\alpha}^*)$, and a term quadratic in $(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)$. One finds, after some algebra,

$$(\boldsymbol{\xi} - \boldsymbol{F} \cdot \boldsymbol{\alpha})^\dagger \cdot \boldsymbol{G}^{-1} \cdot (\boldsymbol{\xi} - \boldsymbol{F} \cdot \boldsymbol{\alpha}) = (\boldsymbol{\xi} - \boldsymbol{F} \cdot \boldsymbol{\alpha}^*)^\dagger \cdot \boldsymbol{G}^{-1} \cdot (\boldsymbol{\xi} - \boldsymbol{F} \cdot \boldsymbol{\alpha}^*) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^\dagger \cdot \boldsymbol{H} \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)$$

Let us consider a linear transformation from the set of n variables $\boldsymbol{\xi}$ to the set consisting of the r variables $\boldsymbol{\alpha}^*$ and some $(n-r)$ additional variables, say $\boldsymbol{y}$. The probability density for the new set has the form

$$\begin{aligned}L(\boldsymbol{\alpha}^*, \boldsymbol{y}; \boldsymbol{\alpha}) \propto \exp&\left[-\tfrac{1}{2}(\boldsymbol{\xi} - \boldsymbol{F} \cdot \boldsymbol{\alpha}^*)^\dagger \cdot \boldsymbol{G}^{-1} \cdot (\boldsymbol{\xi} - \boldsymbol{F} \cdot \boldsymbol{\alpha}^*)\right] \\ \cdot \exp&\left[-\tfrac{1}{2}(\boldsymbol{\alpha}^* - \boldsymbol{\alpha})^\dagger \cdot \boldsymbol{H} \cdot (\boldsymbol{\alpha}^* - \boldsymbol{\alpha})\right] \qquad 29.\end{aligned}$$

The second exponential factor is evidently just the probability density of the $\boldsymbol{\alpha}^*$ (except for a normalization constant). The first factor must therefore be the probability density for the remaining $(n-r)$ variables $\boldsymbol{y}$. It follows from the general results of Section 1.4 that $\chi^2(\boldsymbol{\alpha}^*)$—that is, the quadratic form in the exponent—obeys the χ^2 distribution for $(n-r)$ degrees of freedom. We have assumed up to now that we knew the form of the basic distribution, that is, the likelihood function. If our theory is incorrect, we would be quite likely to get a poor fit: there will be no set of values $\boldsymbol{\alpha}^*$ such that x_i will be close to $f_i(\boldsymbol{\alpha}^*)$ for all n observations. In such a case one may obtain an improbably large value of χ^2. A very large value of χ^2 is therefore taken as an indication that the original hypothesis (i.e., the original likelihood function) may have been incorrect. We can not "prove" that the original hypothesis was wrong; we cannot even attach a Bayesian probability to the correctness of the hypothesis unless some other hypotheses present themselves to which we would attach nonvanishing probability. In other words, an unexpectedly large value of χ^2 will motivate us to search for possible alternative hypotheses.

Very often the mistake does not lie in the overall formulation, but in one or a few of the n observations. One can often establish which ones of the observations are in error by an examination of the "residuals" $\rho_i \equiv [x_i - f_i(\alpha^*)]$; their expected values are zero; their moment matrix is readily calculated:

$$\begin{aligned}\langle(\delta\boldsymbol{\varrho})(\delta\boldsymbol{\varrho})^\dagger\rangle &= \langle\delta(\boldsymbol{\xi} - \boldsymbol{F}\cdot\boldsymbol{\alpha}^*)\delta(\boldsymbol{\xi} - \boldsymbol{F}\cdot\boldsymbol{\alpha}^*)^\dagger\rangle \\ &= (\boldsymbol{I} - \boldsymbol{F}\cdot\boldsymbol{H}^{-1}\cdot\boldsymbol{F}^\dagger\cdot\boldsymbol{G}^{-1})\cdot\langle(\delta\boldsymbol{\xi})(\delta\boldsymbol{\xi})^\dagger\rangle\cdot(\boldsymbol{I} - \boldsymbol{G}^{-1}\cdot\boldsymbol{F}\cdot\boldsymbol{H}^{-1}\cdot\boldsymbol{F}^\dagger) \\ &= (\boldsymbol{I} - \boldsymbol{F}\cdot\boldsymbol{H}^{-1}\cdot\boldsymbol{F}^\dagger\cdot\boldsymbol{G}^{-1})\cdot\boldsymbol{G}\cdot(\boldsymbol{I} - \boldsymbol{G}^{-1}\cdot\boldsymbol{F}\cdot\boldsymbol{H}^{-1}\cdot\boldsymbol{F}^\dagger) \\ &= \boldsymbol{G} - \boldsymbol{F}\cdot\boldsymbol{H}^{-1}\cdot\boldsymbol{F}^\dagger\end{aligned}$$

Hence the σ of ρ_i is given by

$$\sigma_{\rho_i} = \left[G_{ii} - \sum_{\lambda,\mu}^{r} F_{i\lambda}F_{i\mu}(H^{-1})_{\lambda\mu}\right]^{1/2}$$

An observation whose residual ρ_i is large in magnitude compared to its standard deviation σ_{ρ_i} may be considered suspect. This method must be applied with caution; it is most likely to be useful when the number of degrees of freedom $(n-r)$ is large and the number of mistakes small. In the extreme case of only one degree of freedom, one can show that the quantity $|\rho_i|/\sigma_{\rho_i}$ is the same for all n variables; in this case the residuals contain no more information than the value of χ^2. A more detailed discussion of this method and its application to the problem of the determination of the atomic constants is found in (5).

3.2 *Polynomial fit.*—Let us consider the following simple example: we want to fit a polynomial,

$$y(x) = \sum_{\lambda=1}^{r} \alpha_\lambda x^{\lambda-1}$$

to a set of n observations y_i made at a series of points x_i. Let us assume that the observations are independent and that the ith one has a standard deviation σ_i. Then we have

$$\alpha_\lambda^* = \sum_{\mu=1}^{r} (H^{-1})_{\lambda\mu} Y_\mu$$

where

$$H_{\lambda\mu} = \sum_{i=1}^{n} (x_i)^{\lambda+\mu-2}/\sigma_i^2$$

$$Y_\mu = \sum_{i=1}^{n} y_i(x_i)^{\mu-1}/\sigma_i^2$$

We also find

$$\langle(\delta\alpha_\lambda^*)(\delta\alpha_\mu^*)\rangle = (H^{-1})_{\lambda\mu}$$

$$\chi^2(\alpha^*) = \sum_{i=1}^{n}\left[y_i - \sum_{\lambda=1}^{r}\alpha_\lambda^*(x_i)^{\lambda-1}\right]^2\Big/\sigma_i^2 = \sum_{i=1}^{n}\left(\frac{y_1}{\sigma_i}\right)^2 - \sum_{\lambda=1}^{r}\alpha_\lambda^* Y_\lambda$$

$$\rho_i = y_i - \sum \alpha_\lambda^*(x_i)^{\lambda-1}$$

$$\sigma_{\rho_i} = \left[\sigma_i^2 - \sum_{\lambda,\mu=1}^{r}(x_i)^{\lambda+\mu-2}(H^{-1})_{\lambda\mu}\right]^{1/2} \qquad 30.$$

Equation 30 gives a useful check on the numerical computation. We can also obtain the uncertainty in the fitted curve $y^*(x) = \sum\alpha_\lambda^* x^{\lambda-1}$:

$$\langle\delta y^*(x_a)\delta y^*(x_b)\rangle = \sum_{\lambda,\mu=1}^{r}(x_a)^{\lambda-1}(x_b)^{\mu-1}(H^{-1})_{\lambda\mu}$$

The matrix $\boldsymbol{H}$ will not be singular as long as at least r of the coordinates x_i are distinct. However, one may run into one of two numerical difficulties:

(*a*) The differences between various values of x_i are small compared with their average; in that case $\boldsymbol{H}$ may become very nearly singular; this difficulty is circumvented by making a translation along the x axis, say $x' = x - a$, such that the average value of the x_i' is small or zero.

(*b*) The σ of one of the variables may be much smaller than all the others; then its contribution to $H_{\lambda\mu}$ dominates, and $\boldsymbol{H}$ may therefore become nearly singular; in cases of that type one can always find a rearrangement of the equations for α_λ which avoids the difficulty.

3.3 *Optimum properties.*—One can show that the least-squares method yields efficient estimates of the parameters α_λ if the variables x_i are Gaussian and the $f_i(\alpha_\lambda)$ are linear functions of the α_λ. Even if the variables x_i are not Gaussian, the least-squares method gives the most efficient unbiased linear estimates, i.e., α_λ^*, which are linear functions of the observations [see (3) or (5)]. The fact that the σ's of the estimates are minimum when G_{ij}, the correct moment matrix for the x_i, is used has the following important practical consequence: A small error in G will have only a second-order effect on the standard deviations of the α_λ^*. One can sometimes obtain a very great simplification in the numerical computation by neglecting some small correlations. In any case, G is generally known only approximately. One can very often improve one's knowledge of G by a study of the distribution of χ^2 and the residuals over a series of experiments.

3.4 *Nonlinear problem.*—In most applications the functions $f_i(\alpha_\lambda)$ are at best only approximately linear. In the nonlinear case the least-squares estimator is no longer unbiased or efficient. However, if the nonlinearities are small within the part of the space of the parameters in which the likelihood function is large, then various properties deduced for the linear case may still apply to good approximation. It is often adequate to represent f in the form

$$f_i(\alpha_\lambda) \approx f_i(\alpha_\lambda^0) + \sum_{\lambda=1}^{r}(\alpha_\lambda - \alpha_\lambda^0)\left(\frac{\partial f_i}{\partial\alpha_\lambda}\right)_{\alpha^0} \qquad 31.$$

where α_λ^0 is a point near the minimum of χ^2 found by some method of approximation. If the first approximation is not adequate, one can often find the minimum by a simple iteration procedure: Find the minimizing values α_λ' assuming the form 31; re-expand $f_i(\alpha_\lambda)$ about α_λ', find the new minimum, etc. Occasionally more powerful minimizing methods may be necessary.

3.5 *Lagrange multipliers.*—In the least-squares problem we try to "fit" n observations x_i with n functions, f_i, of r parameters. There are therefore $(n-r)$ functional relations among the f_i. It is often convenient to reformulate the problem in a mathematically equivalent way: We introduce the $(n-r)$ constraints $C_k(f_i)$ explicitly with the help of the Lagrange multipliers λ_k. We rewrite χ^2 in the form

$$\chi^2(\boldsymbol{x}; \boldsymbol{f}, \boldsymbol{\lambda}) = (\boldsymbol{x} - \boldsymbol{f})^\dagger \cdot \boldsymbol{G}^{-1} \cdot (\boldsymbol{x} - \boldsymbol{f}) + 2\boldsymbol{\lambda}^\dagger \cdot \boldsymbol{C}(\boldsymbol{f})$$

The least-squares solution is then given by the stationary point of χ^2 as a function of the n variables f_i and the $(n-r)$ variables λ_k. Sometimes the constraints will be functions not only of the f_i but also of some additional s unknowns, say y_u. One can then either reduce the number of constraints by eliminating the y_u, or find the stationary point of χ^2 as a function of the set of variables $\boldsymbol{f}$, $\boldsymbol{\lambda}$, and $\boldsymbol{y}$. Let us assume that the C_k are linear functions of the f_i and y_u:

$$\boldsymbol{C}(\boldsymbol{f}, \boldsymbol{y}) = \boldsymbol{C}_0 + \boldsymbol{U}^\dagger \cdot \boldsymbol{f} + \boldsymbol{V}^\dagger \cdot \boldsymbol{y}$$

One obtains three sets of linear equations on setting to zero the first derivatives of χ^2 with respect to $\boldsymbol{f}$, $\boldsymbol{y}$, and $\boldsymbol{\lambda}$:

$$\begin{aligned} -\boldsymbol{G}^{-1} \cdot (\boldsymbol{x} - \boldsymbol{f}^*) + \boldsymbol{U} \cdot \boldsymbol{\lambda}^* &= 0, \\ \boldsymbol{V} \cdot \boldsymbol{\lambda}^* &= 0, \\ \boldsymbol{C}_0 + \boldsymbol{U}^\dagger \cdot \boldsymbol{f}^* + \boldsymbol{V}^\dagger \cdot \boldsymbol{y}^* &= 0 \end{aligned}$$

The most straightforward way of solving these equations is to eliminate $\boldsymbol{f}^*$, using the first and last set, then using the result and the second set to eliminate $\boldsymbol{\lambda}^*$ and solve for $\boldsymbol{y}^*$; back-substitution then gives $\boldsymbol{\lambda}^*$ and $\boldsymbol{f}^*$. One obtains

$$\begin{aligned} \boldsymbol{y}^* &= -\boldsymbol{K}^{-1} \cdot \boldsymbol{J} \cdot (\boldsymbol{C}_0 + \boldsymbol{U}^\dagger \cdot \boldsymbol{x}) \\ \boldsymbol{\lambda}^* &= \boldsymbol{H}^{-1} \cdot (\boldsymbol{C}_0 + \boldsymbol{U}^\dagger \cdot \boldsymbol{x} + \boldsymbol{V}^\dagger \cdot \boldsymbol{y}^*) \\ \boldsymbol{f}^* &= \boldsymbol{x} - \boldsymbol{E} \cdot \boldsymbol{\lambda}^* \end{aligned}$$

here

$$\boldsymbol{E} \equiv \boldsymbol{G} \cdot \boldsymbol{U}, \qquad \boldsymbol{H} \equiv \boldsymbol{U}^\dagger \cdot \boldsymbol{G} \cdot \boldsymbol{U}, \qquad \boldsymbol{J} \equiv \boldsymbol{V} \cdot \boldsymbol{H}^{-1}, \qquad \boldsymbol{K} \equiv \boldsymbol{V} \cdot \boldsymbol{H}^{-1} \cdot \boldsymbol{V}^\dagger$$

It may happen that $\boldsymbol{H}$ is singular or nearly so, although the overall set of equations has a well-defined solution; in that case one must rearrange the equations. If, for instance, G is diagonal and one particular x_i has a much larger σ than the others, $\boldsymbol{H}$ will be nearly singular; one can avoid this difficulty by grouping the equations for the corresponding f_i with the set for the unknowns y_u [see (7)].

The $\boldsymbol{y}^*$, $\boldsymbol{\lambda}^*$, $\boldsymbol{f}^*$ are linear functions of the variables $\boldsymbol{x}$, and their moment matrices are readily calculated; one finds

$$\langle(\delta \boldsymbol{f}^*)(\delta \boldsymbol{f}^*)^\dagger\rangle = \boldsymbol{G} - \boldsymbol{E}\cdot(\boldsymbol{H}^{-1} - \boldsymbol{J}^\dagger\cdot\boldsymbol{K}^{-1}\cdot\boldsymbol{J})\cdot\boldsymbol{E}^\dagger$$

$$\langle(\delta \boldsymbol{f}^*)(\delta \boldsymbol{y}^*)^\dagger\rangle = -\,\boldsymbol{E}\cdot\boldsymbol{J}\cdot\boldsymbol{K}^{-1}$$

$$\langle(\delta \boldsymbol{y}^*)(\delta \boldsymbol{y}^*)^\dagger\rangle = \boldsymbol{K}^{-1}$$

The moment matrix of the residuals is given by

$$\langle\delta(\boldsymbol{f} - \boldsymbol{x})\cdot\delta(\boldsymbol{f} - \boldsymbol{x})^\dagger\rangle = \boldsymbol{E}\cdot(\boldsymbol{H}^{-1} - \boldsymbol{J}^\dagger\cdot\boldsymbol{K}^{-1}\cdot\boldsymbol{J})\cdot\boldsymbol{E}^\dagger$$

If the constraints are nonlinear functions, one can usually find the least-squares solution by expanding them to first order in $\boldsymbol{f}$ and $\boldsymbol{y}$ and iterating.

The method described above is widely applied to the kinematical analysis of interactions observed in bubble chambers (6–8); the x_i are the measured directions and momenta of the observed charged tracks at an interaction vertex, the f_i the corresponding quantities satisfying the constraints of energy and momentum balance. If some of the directions or momenta are unmeasured (e.g., if an unobserved neutral particle takes part in the interaction), the corresponding variables may be introduced as unknowns y_u, or they may be eliminated by using a subset of the energy-momentum constraints. The "measured" momenta and directions are estimates based on an analysis of the stereo images of the tracks. These estimates are usually the result of least-squares fits. Thus, the overall analysis of an interaction or a chain of interactions usually proceeds through a series of least-squares fits. An alternative approach is to make an overall fit to the "primitive" measurements by introducing a sufficient number of variables, unknowns, and constraints. The breakdown of the overall problem into a series of steps has several advantages: It allows one to deal with problems of smaller dimension which are more manageable from the computational viewpoint; it also makes it easier to track down mistakes and inconsistencies. On the other hand, the breakdown usually necessitates neglect of some correlations, and therefore involves some loss of information. There are in general a number of ways in which a complex analysis problem can be broken down into a series of phases. The detailed choice clearly involves a considerable amount of judgment and experience.

The kinematical analysis serves two distinct purposes: (*a*) It yields estimates of the unmeasured momenta and directions and improved estimates of the measured momenta and directions; these are used to calculate quantities of physical interest such as center-of-mass angles, or the invariant masses of groups of particles participating in the interaction. (*b*) The relative goodness of fit may help to decide among different possible interactions when the masses of the participating particles are not known. If hypothesis *A* gives an improbably large value of χ^2 and hypothesis *B* gives a χ^2 value of the order of the expected value or less, then, by Bayes' principle, one would consider *B* more likely to be correct than *A* unless one has other strong reasons to prefer *A*. If neither χ^2 is very large, one cannot make a decision with any confidence on the basis of kinematics alone. Formulating a χ^2 criterion must be done with care; choosing consistently the hypothesis with the smaller χ^2 may lead to quite erroneous results: (*a*) Reaction A may occur much more frequently than B—there might even be a selection rule against B; taking always the smaller χ^2 would lead one to conclude erroneously that the selection rule is violated in cases in which by chance B happened to lead to a smaller χ^2. (*b*) Even if reactions A and B occur with comparable rates, A might be more likely to "fake" B (that is, lead to a small χ^2 for B) than B to simulate A; a division based on the smaller χ^2 would then lead to too large a fraction of events assigned to B. Thus, before one can attempt a division of a sample of ambiguous events, one must have some idea of the relative rates of the possible reactions and some knowledge of how likely it is that one reaction will simulate another.

Let us mention one other type of application of the least-squares method with Lagrange multipliers. Suppose we want to determine the geometrical parameters describing a bubble chamber and its associated camera. We may do this by measuring the positions of a series of reference marks and the positions of the images of these reference marks. The true positions of a reference mark and its image are constrained to lie on a light ray. The constraints are then functions of the coordinates of the reference marks and their images, as well as unknown parameters of the optical elements. In general, measurements of some of the optical parameters are necessary; they must be included in the least-squares sum. The least-squares equations will be nonsingular only if the measurements are adequate to determine the unknowns. There must clearly be at least as many constraints as unknowns, but this condition is not sufficient; if, for instance, all reference marks are in a plane parallel to the plane containing the lenses, one can learn nothing about the lens separation from the stereo images.

4. Analysis of Angular Distributions

4.1 *Single-parameter case.*—Let us first consider the simple case of the decay of a spin $\frac{1}{2}$ particle such as a Λ into a spin 0 and a spin $\frac{1}{2}$ particle. The angular distribution is of the form

$$P(x;\alpha) = \tfrac{1}{2}(1+\alpha x), \qquad \text{for } -1 \le x \le 1$$

Here x is the cosine of the angle of one of the decay particles with respect to the direction of polarization of the decaying particle; α is the product of the degree of polarization and the parity-nonconservation parameter. We find, for the first two moments,

$$\langle x\rangle = \int_{-1}^{1} dx\cdot\tfrac{1}{2}(1+\alpha x)\cdot x = \tfrac{1}{3}\alpha$$

$$\langle(\delta x)^2\rangle = \langle x^2\rangle - \langle x\rangle^2 = \tfrac{1}{9}(3-\alpha^2)$$

Suppose we have a sample of N decays, i.e., N cosines x_i. The simplest estimate of the parameter α is then one proportional to the sample mean:

$$\alpha^* = \frac{3}{N}\sum_{i=1}^{N} x_i$$

This estimate is constructed to be unbiased, i.e., $\langle\alpha^*\rangle = \alpha$. Let us call α^* the moment estimator. For the variance of α^* we find

$$\langle(\delta\alpha^*)^2\rangle = \frac{9}{N}\langle(\delta x)^2\rangle = \frac{1}{N}(3-\alpha^2) \qquad 32.$$

The minimum possible variance is given by Equation 25.

$$\begin{aligned}\langle(\delta\alpha^*)^2\rangle_{\min} &= \left[N\int_{-1}^{1}dx\cdot\frac{1}{P}\left(\frac{\partial P}{\partial\alpha}\right)^2\right]^{-1} \\ &= \left[N\int_{-1}^{1}dx\cdot\frac{1}{2}\,\frac{x^2}{1+\alpha x}\right]^{-1} \\ &= \left[\frac{N}{2\alpha^3}\left(\ln\frac{1+\alpha}{1-\alpha}-2\alpha\right)\right]^{-1} \\ &= \frac{1}{N}\left[\frac{1}{3}+\frac{1}{5}\alpha^2+\frac{1}{7}\alpha^3\cdots\right]^{-1} \qquad 33.\end{aligned}$$

Comparison of Equations 32 and 33 shows that the moment estimator is very nearly efficient in this case, as long as $|\alpha|$ is not very close to unity. It should be noted that α^* can range between -3 and 3, although the parameter α is meaningful only in the interval $-1 \le \alpha \le 1$; a value of α^* greater than unity would be taken as evidence that the true α is probably close to unity, at least if N is not very small. The moment estimator leads to a very simple rule for combining experiments:

$$\alpha_{\text{comb}} = \frac{3}{N_1+N_2}\sum_{i=1}^{N_1+N_2} x_i = \frac{1}{N_1+N_2}(N_1\alpha_1^* + N_2\alpha_2^*)$$

(here α_1^*, α_2^* are the moment estimates for two experiments, and N_1, N_2 the number of events).

The maximum-likelihood (m.l.) estimator has some advantages and disadvantages compared with the moment estimator. It approaches efficiency as N becomes very large, hence for sufficiently large N and $|\alpha|$ close to unity it may be appreciably more efficient than the moment estimator. For small values of N, however, the m.l. estimator will be distinctly biased towards small values of $|\alpha|$, since it must be confined to the interval $-1 \le \alpha_{\text{m.l.}}^* \le 1$. One should therefore not attempt to combine directly the m.l. estimates obtained for two low-statistics experiments; one should instead multiply the two corresponding likelihood functions and derive a new m.l. estimate from the product.

Let us next take a slightly more general case; consider a distribution of the form

$$P(x;\alpha) = [1+\alpha f(x)]g(x)$$

Let us further assume that f is an odd function of the variables x in the sense that

$$\int f(x)g(x)dx = 0$$

and

$$\int f^3(x)g(x)dx = 0$$

Then we have

$$\begin{aligned}\langle f\rangle &= \alpha f_2\\ \langle f^2\rangle &= f_2\\ \langle(\delta f)^2\rangle &= f_2(1 - \alpha^2 f_2)\end{aligned}$$

where

$$f_2 \equiv \int f^2(x)g(x)dx \qquad 34.$$

The moment estimator for α is then

$$\alpha^* = \frac{1}{f_2}\,\frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

the variance is

$$\langle(\delta\alpha^*)^2\rangle = \frac{1}{N}\left(\frac{1}{f_2} - \alpha^2\right)$$

In the case of scattering of polarized spin $\frac{1}{2}$ particles, we have a distribution of the form

$$(1 + P_{\text{inc}}\cdot P'(\theta)\cdot\sin\phi)$$

where P_{inc} is the degree of polarization of the incoming particle, P' is the analyzing power of the scattering process, and θ and ϕ are the scattering angle and azimuth of the scattered particle. If P' is known, we can estimate angle P_{inc} by the moment method; evidently $f(\theta, \phi) = P'(\theta) \sin\phi$.

As another example, consider an experiment to determine the magnetic moment of the Λ: A sample of Λ's with a polarization p_Λ initially in the x direction is decaying in a magnetic field oriented in the z direction. The direction of polarization will precess at a frequency ω proportional to the Λ magnetic moment. The distribution of the decay proton is then of the form

$$P(\boldsymbol{n}, t; \omega) \approx [1 + a_\Lambda p_\Lambda(n_x + \omega t n_y)]\,\frac{1}{\tau_\Lambda}\exp(-t/\tau_\Lambda) \qquad 35.$$

Here $\boldsymbol{n}$ is the direction of the decay proton in the Λ rest frame, t the decay time, a_Λ the decay asymmetry parameter, and τ_Λ the Λ lifetime; we assume $\omega\tau_\Lambda \ll 1$. We can estimate ω by the moment method,

$$\omega^* = \frac{1}{Nf_2}\sum_{i=1}^{N} f(\boldsymbol{n}_i, t_i)$$

with

$$f(n, t) = a_\Lambda p_\Lambda n_y t$$

Note that the term in n_x in Equation 35 has no effect since it is orthogonal to the other two terms. If the efficiency for detecting decays is nearly independent of the time t and the direction $\boldsymbol{n}$, then $\langle t^2\rangle \approx 2\tau_\Lambda{}^2$ and $\langle n_y{}^2\rangle \approx \frac{1}{3}$. Hence we obtain

$$f_2 \approx \tfrac{2}{3}(a_\Lambda p_\Lambda \tau_\Lambda)^2$$

In many practical situations, the functions $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are so complicated that it is hard to calculate the integral for f_2, Equation 34. If N is not too small, one can obtain a reasonable approximation to f_2 by taking the sample average of f^2; that is,

$$f_2 \approx \frac{1}{N}\sum_{i=1}^{N} f^2(x_i)$$

This method can be used even if the exact form of g is not known. However, the method will be biased unless $\int f(\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x}=0$.

4.2 *General case.*—Let us assume that the general angular distribution is given by $P(\boldsymbol{x};\, \boldsymbol{a})$; the variables $\boldsymbol{x}$ represent all the angles in the problem—in thc case of an interaction chain, not only the primary production angles, but also those of the subsequent interactions or decays; the reaction or scattering amplitudes are assumed to be expressed in terms of the parameters α. The total expected number of events, ν, is also a function of α. Let us define $R(\boldsymbol{x};\, \alpha) = \nu(\alpha)P(\boldsymbol{x};\, \alpha)$; the function R specifies the number of events expected in an element of the angular space. We can now write the likelihood function for the experiment in one of two slightly different forms,

$$\begin{aligned}\mathcal{L}(N, \boldsymbol{x}_i;\, \alpha) &= \frac{1}{N!}e^{-\nu(\alpha)}\prod_{i=1}^{N} R(\boldsymbol{x}_i;\, \alpha)\\ &= \frac{e^{-\nu}\nu^N}{N!}\cdot L(\boldsymbol{x}_i;\, \alpha)\end{aligned} \qquad 36.$$

where

$$L(\boldsymbol{x}_i;\, \alpha) = \prod_{i=1}^{N} P(\boldsymbol{x}_i;\, \alpha)$$

and

$$\int P(x,\, \alpha)dx = 1$$

The generalized likelihood function $\mathcal{L}$ describes the probability distribution in N, the total number of events, as well as in the N sets of continuous variables $\boldsymbol{x}_i$; on the other hand, L is just the probability density describing the shape of the angular distribution. Whether to deal with the generalized function $\mathcal{L}$ or the "usual" likelihood L is primarily a question of convenience. In problems in which the shape of the distribution is of primary interest, one would in general use L; it contains one free parameter less than $\mathcal{L}$, because it is independent of the total rate. On the other hand, removing one parameter from P makes it a somewhat more complicated function of the remaining variables.

Once the likelihood function, either L or $\mathcal{L}$, has been formulated, one generally makes a numerical search for regions in the parameter space in which L (or $\mathcal{L}$) is relatively large.[7]

4.3 *Histogram method.*—When the number of events is very large, one can reduce the amount of computation by splitting up the angular space into s intervals. Let $p_k(\alpha)$ be the expected number of events in the kth interval and n_k the observed number. The likelihood function is then a product of Poisson distributions:

$$\mathcal{L}(n_k;\ \alpha) = \prod_{k=1}^{s} \frac{(p_k)^{n_k} e^{-p_k}}{n_k!} \qquad 37.$$

where

$$p_k(\alpha) = \int_{\Delta k} R(x;\ \alpha) dx$$

The grouping of events into histogram intervals always results in some loss of information, but this loss will be quite small as long as the variation of $R(x;\ \alpha)$ over each interval is relatively small for all admissible values α. In general it is more convenient to deal with the logarithm of the likelihood function than with the likelihood function itself. Let us define w by

$$w(n_k;\ \alpha) = \sum_{k=1}^{s} \left[n_k \ln p_k(\alpha) - p_k(\alpha) \right] \qquad 38.$$

the quantity w is the logarithm of $\mathcal{L}$ to within an additive constant.

If $n_k \gg 1$ for all histogram intervals, then the differences $(n_k - p_k)$ are expected to be small compared with n_k. Let us write

$$p_k = n_k(1 + \epsilon_k), \qquad \text{with } \epsilon_k \equiv \frac{p_k - n_k}{n_k}$$

Expansion of the logarithm in Equation 37 in powers of ϵ gives

$$\begin{aligned} w(n_k;\ \alpha) &\approx \sum_{k=1}^{s} n_k(\ln n_k - 1) - \frac{1}{2} \sum_{k=1}^{s} n_k \epsilon_k^2 \\ &= \sum_{k=1}^{s} n_k(\ln n_k - 1) - \frac{1}{2} \sum_{k=1}^{s} \left[\frac{p_k(\alpha) - n_k}{n_k^{1/2}} \right]^2 \end{aligned} \qquad 39.$$

The first term in Equation 38 is independent of α; the second term is minus one-half the least-squares sum that one obtains when one approximates the σ for n_k by $(n_k)^{1/2}$.

Many scattering experiments do not contain enough information to allow one to determine the phase shifts or scattering amplitudes. In that case, one has to content oneself with determining the coefficients of the spherical harmonics entering into the description of the angular distribution; $p_k(\alpha)$ is then linear in the parameters α, and the least-squares approximation, Equation 39, leads to linear equations for the α. When the least-squares approximation is not justified (that is, when $n_k \gg 1$ does not hold for some inter-

[7] For a brief description of some search techniques, see Rosenfeld & Humphrey (8).

vals), one should maximize the exact expression for w, Equation 38; in that case the least-squares approximation will generally provide a good starting point for an iterative solution.

In a counter experiment one always deals with likelihood functions of the form of Equation 37, which are functions of the discrete variables n_k, the number of counts registered by a particular counter or counter combination.

4.4 *Detection efficiency.*—In a bubble chamber experiment, the probability of detecting an event will in general depend on the angles $\boldsymbol{x}$ as well as on some additional variables $\boldsymbol{y}$ specifying the position of the interaction vertex (or vertices). Let us assume that the detection probability is given by $e(\boldsymbol{x}, \boldsymbol{y})$ and the distribution of $\boldsymbol{y}$ by $Q(\boldsymbol{y})$. Our likelihood function, Equation 36, must then be modified by replacing R by R' and ν by ν', where

$$R'(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\alpha}) = R(\boldsymbol{x}; \boldsymbol{\alpha})e(\boldsymbol{x}, \boldsymbol{y})Q(\boldsymbol{y})$$

and ν' is the integral of R'. The quantity R can in general be written as a sum of products:

$$R(\boldsymbol{x}; \boldsymbol{\alpha}) = \sum_{a=1}^{A} U_a(\boldsymbol{\alpha})g_a(\boldsymbol{x})$$

The logarithm of $\mathcal{L}$ can then be written in the form

$$W = -\sum_{a=1}^{A} e_a U_a(\boldsymbol{\alpha}) + \sum_{i=1}^{N} \ln R(\boldsymbol{x}_i; \boldsymbol{\alpha}) \qquad 40.$$

where

$$e_a = \int g_a(\boldsymbol{x})e(\boldsymbol{x}, \boldsymbol{y})Q(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \qquad 41.$$

we have dropped terms in W independent of the parameters $\boldsymbol{\alpha}$. We see that the detection efficiency enters the problem only through the constants e_a defined by the above integral, Equation 41; they can be calculated once and for all for a given experiment.

In practice it is often quite difficult to evaluate the integrals for e_a. An alternative, simpler method consists in "weighting" the individual events with the reciprocal of the detection efficiency. One defines a new quantity W':

$$W' = -\int R(\boldsymbol{x}; \boldsymbol{\alpha})d\boldsymbol{x} + \sum_{i=1}^{N} \frac{1}{e_i} \ln R(\boldsymbol{x}_i; \boldsymbol{\alpha}) \qquad 42.$$

where

$$e_i = e(\boldsymbol{x}_i, \boldsymbol{y}_i)$$

Maximization of W' leads to asymptotically unbiased estimates of the parameters $\boldsymbol{\alpha}$ [this can be shown by expanding W' to second order about the true values of the parameters and showing that $\langle(\partial W'/\partial\alpha)\rangle\alpha_{\text{true}} = 0$]. Since W' is not simply related to the true likelihood function of the problem, the inverse error matrix is not given by $-(\partial^2 W'/\partial\alpha_\lambda\partial\alpha_\mu)$. One can show (again by use of a Taylor expansion) that the error matrix is given asymptotically by

$$\langle(\delta\alpha^*)(\delta\alpha^*)^\dagger\rangle = H^{-1}\cdot H'\cdot H^{-1}$$

where

$$H_{\lambda\mu} = -\frac{\partial^2 W'}{\partial\alpha_\lambda\partial\alpha_\mu}$$

and

$$H_{\lambda\mu}' = \sum_{i=1}^{N}\left(\frac{1}{e_i R_i}\right)^2 \frac{\partial R_i}{\partial\alpha_\lambda}\,\frac{\partial R_i}{\partial\alpha_\mu}$$

There is some loss of information involved in the use of this method of "weighting." This loss is serious if the function $e(x, y)$ is close to zero for some values of x and y (this difficulty can usually be avoided by a suitable choice of fiducial volume).

LITERATURE CITED

1. Jeffreys, H., *Theory of Probability*, 3rd ed. (Oxford Univ. Press, Oxford, England, 447 pp., 1961)
2. Cramér, H., *Mathematical Methods of Statistics* (Princeton Univ. Press, Princeton, N. J., 575 pp., 1946)
3. Kendall, M. G., and Stuart, A., *The Advanced Theory of Statistics*, **2** (Charles Griffin and Co., Ltd., London, 676 pp., 1961)
4. Bartlett, M. S., *Phil. Mag.*, **44**, 249, 1407 (1953)
5. Cohen, E. R., Crowe, K. M., and Dumond, J. W. M., *Fundamental Constants of Physics* (Interscience Publ., Inc., New York, 287 pp., 1957)

6a. Berge, J. P., Solmitz, F. T., and Taft, H. D., *Rev. Sci. Instr.*, **32**, 538–48 (1961)

6b. Böck, R., *Kinematics Analysis of Bubble Chamber Events* (CERN 61-29, Data Handling Division, 1961) (Unpublished)

7. Solmitz, F. T., *Alvarez Memo 187* (Lawrence Radiation Lab., 1960) (Unpublished)
8. Rosenfeld, A. H., and Humphrey, W. E., *Ann. Rev. Nucl. Sci.*, **13**, 103 (1963)