REVIEWS OF MODERN PHYSICS VOLUME 25, NUMBER 4 OCTOBER, 1953

On Statistical Estimation in Physics

M. ANNIS,* W. CHESTON,† AND H. PRIMAKOFF

Washington University,‡ St. Louis, Missouri

The problem of the estimation of parameters determined statistically from physical measurements is discussed. Emphasis is placed on the fundamental role played by the prior probability distribution for the parameter. The validity of "maximum likelihood" estimation is examined with particular reference to the case of the estimation of a parameter which actually has an unique but (originally) unknown magnitude. Situations in which the prior probability distribution for the parameter is completely unknown are treated and a method is described for the calculation of this distribution from appropriate experimental data. Many examples are given throughout from the field of cosmic radiation.

I. INTRODUCTION

THE current paper is the outgrowth of an attempt by the authors to understand certain statistical considerations associated with the determination of elementary particle parameters on the basis of cosmic-ray measurements. We could not discover the answers to several questions which appeared relevant to the problem in any references easily available to physicists and were therefore forced to work out the conclusions below—many of our results are no doubt implicitly or perhaps even explicitly contained in the literature on probability and statistics. Nevertheless, we communicate these results in their present form in the hope that others may possibly find them useful, and with the intention of arousing further interest in the subject.

In physics, a situation frequently arises in which one desires to determine a physical quantity, which we shall call θ, characteristic of either a class of individual particles (e.g., the mean-life of a class of individual (unstable) particles of the same kind) or a class of systems of particles (e.g., the temperature of a class of stars); however, this physical quantity θ cannot in any sense be "measured" directly. Instead, one measures directly another set of n quantities, $x_1, x_2, x_3 \cdots x_n$, which we shall call x_i, which are not related to θ in a simple one-to-one fashion (i.e., θ is not uniquely determined by the set x_i). What we do know on the basis of some theoretical consideration is the probability that the set x_i lie in the interval dx_i (i.e., that, simultaneously, x_1 is between x_1 and x_1+dx_1, x_2 between x_2 and x_2+dx_2, etc.) for a given value of θ. We shall denote this probability function by $G(\theta; x_1, x_2 \cdots x_n) \times dx_1 dx_2 \cdots dx_n \equiv G(\theta; x_i)dx_i$. (In our notation, the quantity before the semi-colon is a parameter and the quantities after the semi-colon are variables.) However, having measured the set x_i directly, we should like to know the probability that θ lies in $d\theta$ for these given values of the set x_i. We shall denote this latter probability (the so-called posterior probability) by $H(x_1, x_2 \cdots x_n; \theta)d\theta \equiv H(x_i; \theta)d\theta$. The question now arises: "What relationship, if any, exists between the two probability functions $G(\theta; x_i)dx_i$ and $H(x_i; \theta)d\theta$?"

In order to answer the question posed above, we define another probability function (the so-called prior probability), $P(\theta)d\theta$. $P(\theta)d\theta$ is the probability that, in the type of experiment performed, θ lies in $d\theta$, independent of the values of the x_i. In the same manner, we may define the probability that, in a measurement of the set x_i, the x_i lie in dx_i, independent of the values of θ. We designate this latter probability by the expression $Q(x_1, x_2, \cdots x_n)dx_1 \cdots dx_n \equiv Q(x_i)dx_i$.

It is not superfluous to mention that all the four probability functions defined may be understood, in a physical context, in the sense of the corresponding, in principle observable, relative frequencies; the normalizability of these probability functions with respect to their variables may then be demanded. It should also be mentioned that our whole treatment can be generalized in a straightforward manner to the case of the existence of several parameters $\theta_1, \cdots, \theta_m$; the essential conclusions obtained below are also valid in the case of such a generalization.

II. CASE OF KNOWN PRIOR PROBABILITY DISTRIBUTION

An intimate relationship exists among the probability functions defined in Sec. I. To exhibit this relationship, we define a function $S(\theta, x_i)d\theta dx_i$ as the simultaneous probability that θ lies in $d\theta$ while the set x_i lie in dx_i. By the rules of combining probabilities, this simultaneous probability is the probability that the x_i lie in dx_i for given θ, multiplied by the probability that θ lies in $d\theta$, independent of the values of x_i; i.e.,

$$S(\theta, x_i)dx_i d\theta = P(\theta)d\theta \cdot G(\theta; x_i)dx_i. \tag{1a}$$

On the other hand, this simultaneous probability can also be expressed as the probability that θ lies in $d\theta$ for given values of the x_i multiplied by the probability that the x_i lie in dx_i, independent of θ; i.e.,

$$S(\theta, x_i)dx_i d\theta = Q(x_i)dx_i \cdot H(x_i; \theta)d\theta. \tag{1b}$$

[*] Present address: Institute of Physics, University of Padua, Padua, Italy.

[†] Present address: Department of Physics, University of Minnesota, Minneapolis, Minnesota.

[‡] Assisted by the joint program of the U. S. Office of Naval Research and the U. S. Atomic Energy Commission.

Combining the results of Eqs. (1a) and (1b), we may write a relationship between $G(\theta; x_i)dx_i$ and $H(x_i; \theta)d\theta$ usually attributed to Bayes, namely,

$$Q(x_i)H(x_i; \theta)dx_i d\theta = P(\theta)G(\theta; x_i)d\theta dx_i \qquad (2a)$$

or

$$H(x_i; \theta) = P(\theta)G(\theta; x_i)/Q(x_i). \qquad (2b)$$

Equation (2b) may be written in a slightly altered form if we remember the normalization of $H(x_i; \theta)$. Integrating both sides of Eq. (2a) over the variable θ, we note that

$$Q(x_i)dx_i \cdot \int H(x_i; \theta)d\theta = Q(x_i)dx_i = \left[\int P(\theta)G(\theta; x_i)d\theta\right] \cdot dx_i, \qquad (2c)$$

it being understood that all integrals contained in this paper extend over the complete range of the integration variable unless otherwise explicitly stated. Therefore, we see that

$$H(x_i; \theta) = \frac{P(\theta)G(\theta; x_i)}{\int P(\theta)G(\theta; x_i)d\theta}. \qquad (2d)$$

We can also write an expression analogous to Eq. (2c) for the probability function $P(\theta)$, via Eq. (2a). This is

$$P(\theta) = \int Q(x_i)H(x_i; \theta)dx_i. \qquad (2e)$$

One should now remember that $G(\theta; x_i)dx_i$ is usually known on the basis of theoretical considerations. It is to be noted, therefore, that any statements we make about $H(x_i; \theta)$ must depend upon our knowledge of the function $P(\theta)d\theta$. If this last function is known via some physical theory and/or a set of previously performed experiments, Eq. (2d) offers us an exact solution for the posterior probability function $H(x_i; \theta)d\theta$; i.e., we can state what the probability is that θ lies in $d\theta$ with our measured values of the set x_i. It is also obvious from Eq. (2d) that $H(x_i; \theta)d\theta$ cannot be determined if $P(\theta)d\theta$ is completely unknown. In many cases, it is, however, not necessary that $P(\theta)d\theta$ be completely known as a function of θ in order to make statements with significance in a probability sense. Such cases will be discussed in detail in subsequent sections of this paper.

We shall now examine more carefully the case where $P(\theta)d\theta$ is a known function of the parameter θ. The existence of the probability function $H(x_i; \theta)d\theta$ allows us here to answer the question, "What is the probability that the parameter θ lies in $d\theta$ for a given set of the x_i?" The answer to this question is given by exhibiting $H(x_i; \theta)$ as in Fig. 1(a) or 1(b).

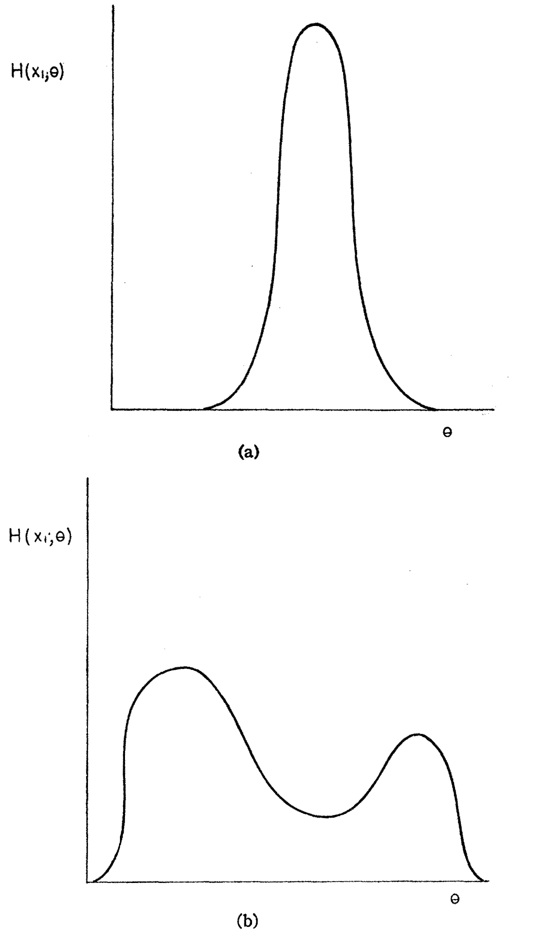

FIG. 1. The posterior probability function, $H(x_i; \theta)$, plotted as a function of θ, (a) in the case where a single estimate of θ might be given, and (b) in the case where no single estimate of θ can be quoted.

When $H(x_i; \theta)$ has the form shown in Fig. 1(a) (i.e., a single, well-defined peak), there is little doubt that one can quote a single value of θ which has "statistical significance." On the other hand, if $H(x_i; \theta)$ looks somewhat as in Fig. 1(b) with widely different θ's having comparable probabilities, it is perhaps unwise to quote a single value for θ. However, in a particular case, the information in Fig. 1(b) could conceivably still be valuable.

When $H(x_i; \theta)$ is well peaked as in Fig. 1(a), it is frequently useful to quote a single "statistically significant" value or estimate for θ, *viz.*, θ_{est}, with an appropriately defined standard deviation. (One should point out that this standard deviation is taken as the standard deviation of the posterior probability distribution about θ_{est} and hence is one measure of the statistical accuracy of the estimate.) Of course one can also easily answer any question concerning the probability of θ lying within a certain range about θ_{est}. This procedure is only useful if the value of θ_{est} and the standard deviation is not sensitive to the method used to calculate θ_{est} (assuming a "reasonable" method is used). A "reasonable" estimate of θ is one which lies somewhere near the probability maximum in Fig. 1(a). From Eq. (2b), it is evident that $H(x_i;\theta)$ might be well peaked if $P(\theta)$ is well peaked, even though $G(\theta;x_i)$ is not a well-peaked function of θ. Hence, a "reasonable" estimate of θ must take into account both $P(\theta)$ and $G(\theta;x_i)$.

More explicitly, any "reasonable" method of estimating θ should have a standard deviation, $\sigma(\theta_{\text{est}})$, not much larger than the minimum possible value of this quantity (see Eq. 7). $\sigma^2(\theta_{\text{est}})$ is defined by

$$\sigma^2(\theta_{\text{est}})=\int(\theta-\theta_{\text{est}})^2H(x_i;\theta)d\theta=\overline{\theta^2}-2\bar{\theta}\,\theta_{\text{est}}+\theta_{\text{est}}^2, \quad (3)$$

where

$$\int\theta^sH(x_i;\theta)d\theta\equiv\overline{\theta^s}.$$

1. One method of evaluating θ_{est} is to quote $\theta_{\text{est}}=\theta_{\text{m.p.}}$, where $\theta_{\text{m.p.}}$ is the value of θ, for which $H(x_i;\theta)$ has a maximum; i.e., $\theta_{\text{m.p.}}$ is the most probable single value of θ. The defining equation for $\theta_{\text{m.p.}}$ is

$$\left.\frac{\partial H}{\partial\theta}(x_i;\theta)\right]_{\theta=\theta_{\text{m.p.}}}=0 \quad (4)$$

and the standard deviation, from Eq. (3), is

$$\sigma^2(\theta_{\text{m.p.}})=\overline{\theta^2}-2\bar{\theta}\,\theta_{\text{m.p.}}+\theta_{\text{m.p.}}^2. \quad (5)$$

2. Using Eq. (3), it is possible to choose a value of θ_{est} which minimizes the standard deviation of the estimate. To do this, we differentiate Eq. (3) with respect to θ_{est} and set the derivative equal to 0, yielding a defining equation for θ_{est}. One finds

$$\theta_{\text{est}}=\bar{\theta} \quad (6)$$

with standard deviation

$$\sigma^2(\bar{\theta})=\overline{\theta^2}-\bar{\theta}^2. \quad (7)$$

3. To minimize the *relative* standard deviation, we solve Eq. (3) for $\sigma^2(\theta_{\text{est}})/\theta_{\text{est}}^2$, and set the derivative of this quantity with respect to θ_{est} equal to zero. We obtain

$$\theta_1=\overline{\theta^2}/\bar{\theta} \quad (8)$$

with relative standard deviation

$$\sigma^2(\theta_1)/\theta_1^2=1-\bar{\theta}^2/\overline{\theta^2}. \quad (9)$$

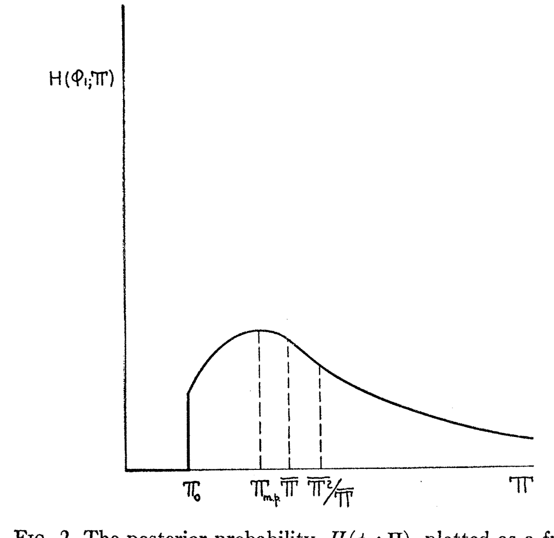

FIG. 2. The posterior probability, $H(\phi_i;\Pi)$, plotted as a function of $\Pi=pc\beta$. The relative position of the different estimates of Π is also shown.

Example from the Theory of Multiple Coulomb Scattering

One measures the n-projected angles of multiple scattering of a particle in the n plates of a multiplate cloud chamber and from this scattering data, one desires to find an estimate of $\Pi=pc\beta$ for the particle (p=momentum; $c\beta$=velocity), assuming that Π does not change while the particle is traversing the chamber. In this case the parameter θ becomes the physical quantity Π and the x_i become ϕ_i, the projected angle of multiple scattering in the ith plate.

From the theory of multiple Coulomb scattering, the probability that ϕ_i lies in $d\phi_i$, $f(\phi_i)d\phi_i$, is given as

$$f(\phi_i)d\phi_i=(2\pi GQ)^{-\frac{1}{2}}\exp\{-\phi_i^2/2GQ\}d\phi_i. \quad (10)$$

Equation (10) is an approximation, good only for $\phi_i^2<2GQ$, where G and Q are defined by Olbert.[1] However we shall assume that Eq. (10) holds for all ϕ_i, the resultant error in the final answer being less than 10 percent.

In the above approximation, G depends only on the material in the plates, while Q depends on the material in the plates, and in addition, is inversely proportional to Π^2. Hence we can define a quantity A independent of Π by the equation

$$A=\Pi(2GQ)^{\frac{1}{2}} \quad (11)$$

so that Eq. (10) becomes

$$f(\phi_i)d\phi_i=N_1\Pi\exp\{-\phi_i^2\Pi^2/A^2\}d\phi_i, \quad (12)$$

where N_1 is a normalization factor, independent of ϕ_i and Π.

Since the angles of multiple scattering in the n plates are statistically independent, it is evident that

$$\begin{aligned}G(\Pi;\phi_i)&=f(\phi_1)f(\phi_2)\cdots f(\phi_n)\\&=N_1^n\Pi^n\exp\left\{-\frac{\Pi^2}{A^2}\sum_{i=1}^{n}\phi_i^2\right\}. \quad (13)\end{aligned}$$

Let us now assume that the prior probability distribution for Π is a power law, i.e.,

$$P(\Pi)=\begin{matrix}0; & \Pi<\Pi_0\\ N_2\Pi^{-\gamma}; & \Pi\geqq\Pi_0,\end{matrix} \quad (14)$$

where N_2 is a normalization factor, and γ is an empirically determined constant ($\gamma\approx3$ in many cases). $H(\phi_i;\Pi)$ may then be evaluated from Eqs. (2*d*) and (13), (14) yielding

$$H(\phi_i;\Pi)=\begin{matrix}0; & \Pi<\Pi_0\\ N_3\Pi^{n-\gamma}\exp\left\{-\frac{\Pi^2}{A^2}\sum_{i=1}^{n}\phi_i^2\right\}; & \Pi\geqq\Pi_0.\end{matrix} \quad (15)$$

$H(\phi_i;\Pi)$ is plotted in Fig. 2 as a function of Π.

[1] S. Olbert, Phys. Rev. **87**, 319 (1952).

The most probable value of Π is, from Eqs. (15) and (4),

$$\Pi_{\text{m.p.}}=\begin{matrix}\Pi_0; & n\leqq\gamma\\ (m/2)^{\frac{1}{2}}A(\sum\limits_{i=1}^{n}\phi_i^2)^{-\frac{1}{2}}; & n>\gamma,\end{matrix}\tag{16}$$

where $m=n-\gamma$. Using Eq. (3), we find that the corresponding standard deviation is (for $\Pi_0\ll\Pi_{\text{m.p.}}$)

$$\sigma(\Pi_{\text{m.p.}})=\Pi_{\text{m.p.}}\{2+1/m-2(2/m)^{\frac{1}{2}}\Gamma(1+m/2)/\Gamma(\tfrac{1}{2}+m/2)\}^{\frac{1}{2}}.\tag{17}$$

Under the same assumption, the mean value of Π, i.e., the estimate of Π which minimizes the standard deviation, is (Eqs. (15) and (6))

$$\langle\Pi\rangle_{\text{Av}}=\frac{\Gamma(1+m/2)}{\Gamma(\frac{1}{2}+m/2)}A(\sum_{i=1}^{n}\phi_i^2)^{-\frac{1}{2}},\tag{18}$$

with standard deviation (Eq. 7)

$$\sigma(\langle\Pi\rangle_{\text{Av}})=\langle\Pi\rangle_{\text{Av}}\left\{\frac{m+1}{2}\frac{\Gamma^2(\frac{1}{2}+m/2)}{\Gamma^2(1+m/2)}-1\right\}^{\frac{1}{2}},\tag{19}$$

and the estimate of Π which minimizes the *relative* standard deviation is (Eqs. (15) and (8))

$$\Pi_1=\langle\Pi^2\rangle_{\text{Av}}/\langle\Pi\rangle_{\text{Av}}=\frac{m+1}{2}\frac{\Gamma(\frac{1}{2}+m/2)}{\Gamma(1+m/2)}A(\sum_{i=1}^{n}\phi_i^2)^{-\frac{1}{2}},\tag{20}$$

with standard deviation (Eq. (9))

$$\sigma(\Pi_1)=\Pi_1\left\{1-\frac{2}{m+1}\frac{\Gamma^2(1+m/2)}{\Gamma^2(\frac{1}{2}+m/2)}\right\}^{\frac{1}{2}}.\tag{21}$$

Each of these estimates is shown in Fig. 2. For all of these estimates

$$\operatorname*{Lim}_{m\to\infty}\sigma(\Pi_{\text{est}})=\Pi_{\text{est}}(2m)^{-\frac{1}{2}}.\tag{22}$$

(For $m=6$, Eq. (22) agrees to within a few percent with Eqs. (17), (19) and (21).)

From Eqs. (16) to (21) it is evident that for $m=(n-\gamma)>6$, and for $\Pi_0\ll\Pi_{\text{m.p.}}$; $\Pi_{\text{m.p.}}$, $\langle\Pi\rangle_{\text{Av}}$ and Π_1 are nearly the same. If $n\gg\gamma$, all of these estimates are effectively independent of $P(\Pi)$, the prior probability distribution in Π, so that uncertainties in detail about $P(\Pi)$ make little difference in the Π estimates.

We have mentioned previously that whenever it is desired to ascribe a numerical value to a statistically determined physical parameter such as θ, it is essential that at least some aspects of the nature of the prior probability function $P(\theta)d\theta$ be known. We have treated the case in which $P(\theta)d\theta$ is a known function of θ. It is, however, not necessary that $P(\theta)$ be completely known as a function of θ; on the other hand, one must at least know whether $P(\theta)$ is: (1) a continuous function of θ, (2) a linear combination of δ functions of θ, or (3) a single δ-function of θ. Put in another way, one should know: (1) whether all values of θ are possible, (2) whether only certain discrete values of θ are possible, or (3) whether θ has a certain unique but as yet unknown value. On the other hand, if $P(\theta)d\theta$ is completely unknown, steps must be taken, in general, to delimit at least some of its aspects. Such steps will now be described.

III. PROCEDURE FOR FINDING AN UNKNOWN PRIOR PROBABILITY DISTRIBUTION

In the previous section, we have assumed that the prior distribution in the physical parameter, $P(\theta)d\theta$, is known. This is often not the case. Indeed, in some situations this prior distribution may actually be the chief quantity which is sought. Before we outline the procedure for finding $P(\theta)$ under these circumstances, we shall describe the experimental work that must be done.

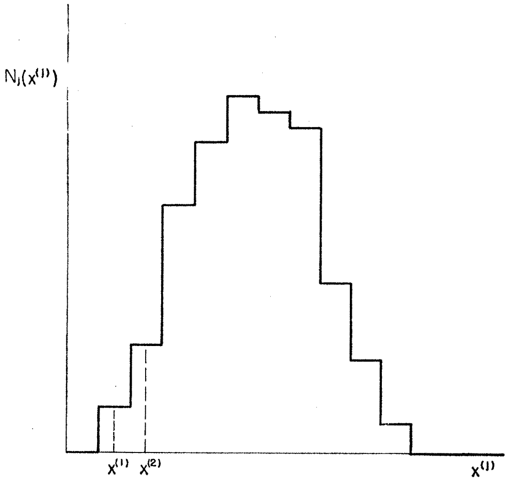

FIG. 3. A plot of $N_j(x^{(j)})$, the number of times the variable x had the measured value $x^{(j)}$, *versus* $x^{(j)}$.

The experiment described in II (i.e., the observation of a set of the x_i for one particular value of θ) is repeated M times. If $P(\theta)$ is a function which is different from zero for more than one value of θ, then one expects that θ will not be unique in this set of M experiments. Each of the experiments will, in general, correspond to a different value of θ. Let us assume for simplicity that there is only one value of x in each of the M experiments (i.e., in *each* experiment $G(\theta;x_i)$ becomes $G(\theta;x)$); then the result of the experiments will look, for example, somewhat as illustrated in Fig. 3, where $N_j(x^{(j)})\Delta x$ is the number of times that the measured value x falls between $x^{(j)}-(\Delta x/2)$ and $x^{(j)}+(\Delta x/2)$, Δx being the difference between adjacent $x^{(j)}$'s. Since θ is not fixed during the M experiments, we do not expect $N_j(x^{(j)})$ to be well peaked unless $P(\theta)$ is well peaked. To convert the $N_j(x^{(j)})$ to a relative frequency distribution, we remember that

$$\sum_{j=1}^{\infty}N_j(x^{(j)})\Delta x=M,$$

so the distribution sought is simply

$$F_j(x^{(j)})=N_j(x^{(j)})/M,$$

where $F_j(x^{(j)})\Delta x$ is the relative frequency with which x lies between $x^{(j)}-\Delta x/2$ and $x^{(j)}+\Delta x/2$. In the limit of small Δx and large M, $F_j(x^{(j)})$ becomes $F(x)$, a continuous distribution in x. It is to be noted that the (prior) probability distribution for x, previously called $Q(x)$, must now be identified with the relative frequency distribution in x, $F(x)$.

The problem then is to find $P(\theta)$ having been given $F(x)$ from experiment and knowing the distribution function, $G(\theta;x)$, from theory. By use of Eq. (2c), we find

$$\lim_{\substack{M\to\infty\\ \Delta x\to 0}} F_j(x^{(j)}) = F(x) = Q(x) = \int P(\theta)G(\theta;x)d\theta. \quad \text{(23a)}$$

Equation (23a) is a *linear integral equation of the first kind*, and has a unique formal solution for the unknown function, $P(\theta)$, *viz.*,

$$P(\theta) = \int F(x')\{G(\theta;x')\}^{-1}dx', \quad \text{(23b)}$$

where

$$\int \{G(\theta;x')\}^{-1}G(\theta;x)d\theta = \delta(x-x'). \quad \text{(23c)}$$

Thus, in the case where $P(\theta)$ is unknown one must repeat the measurement of x (θ *not* fixed) to determine $F(x)=Q(x)$ and then use Eq. (23b) to calculate $P(\theta)$. Of course, in the special case where $P(\theta)$ is a linear combination of δ functions of θ or a single δ function of θ, Eq. (23b) will simply indicate that fact.

Comparison of Eqs. (23b) and (2e) (or of Eqs. (2c) and (2e)) should not lead one to the conclusion that $H(x';\theta) = \{G(\theta;x')\}^{-1}$; this conclusion is obviously wrong, if only because $H(x';\theta)$ depends on $P(\theta)$ (see Eq. (2b)) and $\{G(\theta;x')\}^{-1}$ does not. More explicitly, if $H(x';\theta)$ were equal to $\{G(\theta;x')\}^{-1}$, Eq. (23c) would indicate that

$$\int H(x';\theta)G(\theta;x)d\theta = \delta(x-x'),$$

and this last is impossible since H and G, being both probability distributions, are everywhere positive as functions of θ for any x', x. Thus the reasons for the validity of (23b) and (2e) are quite different; the former follows from (23a) as a consequence of the δ function relation between $G(\theta;x)$ and $\{G(\theta;x)\}^{-1}$, the latter on the other hand follows from (2a) as a consequence of the interconnected definitions of $Q(x)$, $P(\theta)$, $G(\theta;x)$, $H(x;\theta)$ in Eqs. (2).

Example from the Theory of Multiple Coulomb Scattering

Many particles are observed to reach the end of their range in one of the plates of a multiplate cloud chamber. The problem is to estimate the masses and relative frequencies of occurrence of the different particles. We shall assume that each particle is observed to penetrate n plates before reaching the end of its range.

Following Annis *et al.*[2] we define a variable, η_i, by the relationship

$$\eta_i \equiv \phi_i R_i^{\alpha}, \quad \text{(24)}$$

where ϕ_i is the projected angle of scattering in the ith plate, and R_i is the residual range of the particle in the ith plate. The exponent $\alpha=0.553$ for all materials, and is defined by the equation

$$R/mc^2 = A_Z(\Pi/mc^2)^{1/\alpha}, \quad \text{(25)}$$

where $\Pi = pc\beta$ as before, m is the mass of the particle, R is the range of a particle with given Π, and A_Z is a constant for a given scattering material. Equation (25) allows us to write the distribution in η_i in a form analogous to Eq. (11);

$$f(\eta_i) \cong (2\pi)^{-\frac{1}{2}}\rho^{-1}\exp\{-\eta_i^2/2\rho^2\}, \quad \text{(26)}$$

where ρ is a constant for a given particle, independent of the residual range, and depending on the mass of the particle through the relation $\rho \sim m^{1-\alpha}$. From Eq. (26) it is evident that ρ has the dimensions of η_i.

Let us now define the mean square value of the η_i by the equation

$$s^2 = n^{-1}\sum_{i=1}^{n}\eta_i^2, \quad \text{(27)}$$

and let $G(\rho;s)ds$ be the probability for given ρ (i.e., given m), that s lies in ds. Then,

$$G(\rho;s)ds = \int\cdots\int f(\eta_1)\cdots f(\eta_n)d\eta_1\cdots d\eta_n, \quad \text{(28)}$$

the region of integration in Eq. (28) being over those values of the η_i for which Eq. (27) holds. The integral in Eq. (28) can be readily performed and the result is

$$G(\rho;s) = B\rho^{-n}s^{n-1}\exp\{-(1/2)ns^2/\rho^2\}, \quad \text{(29)}$$

where B is a constant independent of ρ and s.

Let us now suppose that we have observed M particles reach the end of their range in the chamber and we calculate s for each of these particles. If $F_i(s^{(i)})\ \Delta s$ is defined as the relative frequency with which s lies between $s^{(i)}-(\Delta s/2)$ and $s^{(i)}+(\Delta s/2)$, the (normalized) $F_i(s^{(i)})$ function might look somewhat as shown in Fig. 4. In this case, Eq. (23a) has approximately the form (if $M\to\infty$ and $\Delta\rho\to 0$)

$$F_i(s^{(i)}) = \sum_{j=1}^{\infty} P_j(\rho^{(j)})G(\rho^{(j)};s^{(i)})\Delta\rho, \quad \text{(30)}$$

where $P_j(\rho^{(j)})\Delta\rho$, is the (unknown) probability that ρ lie between $\rho^{(j)}-\Delta\rho/2$ and $\rho^{(j)}+\Delta\rho/2$. Here $\Delta\rho$ is at our disposal, and since ρ has the dimensions of s, let us take $\Delta\rho=\Delta s$. It is evident from Fig. 4, that the right-hand side of Eq. (30) is zero for $i>N$. This implies that $P_j(\rho^{(j)})<0$ for some values of j (since $G(\rho^{(j)};s^{(i)})$ is always positive), but since $P_j(\rho^{(j)})$ is a probability, $P_j(\rho^{(j)}) \geqq 0$ for all j. The contradiction arises because of the experimental approximation that $F_i(s^{(i)})=0$ for $i>N$. In other words, we can say immediately that $P_j(\rho^{(j)})=0$ for $\rho^{(j)}<s^{(1)}$ and for $\rho^{(j)}>s^{(N)}$ to the approximation considered here, and the sum in Eq. (30) becomes a finite sum over those values of $\rho^{(j)}$ between $s^{(1)}$ and $s^{(N)}$.

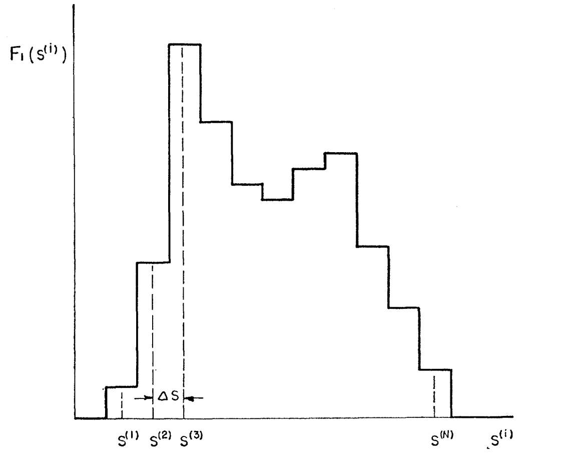

FIG. 4. The form that the data might take in a measurement of the multiple scattering of many particles. The relative frequency $F_i(s^{(i)})$ of different values of $s^{(i)}$ is plotted *versus* the measured $s^{(i)}$.

[2] Annis, Bridge, and Olbert, Phys. Rev. **89**, 1216 (1953).

If one places

$$\Delta\rho=\Delta s,$$

$$F_i(s^{(i)})=f_i, \tag{31}$$

$$P_j(\rho^{(j)})=p_j,$$

and

$$(\Delta\rho)G(\rho^{(j)};s^{(i)})=g_{ji},$$

Eq. (30) becomes

$$f_i=\sum_{j=1}^{N} p_j g_{ji};\quad i=1,2,\cdots,N, \tag{32}$$

where f_i and the g_{ji} are known and the p_j are unknown. There are N equations in Eq. (32) with N unknowns, so that we can solve immediately for the p_j. Explicitly,

$$p_j=|g_{ji}^{-1}|\sum_{i=1}^{N} f_i G_{ji}, \tag{33}$$

where $|g_{ji}|$ is the determinant of the g_{ji} and G_{ji} is the cofactor of g_{ji}. In addition, the "statistical errors" assigned to the p_j are related to the "statistical errors" in the f_i's by the equation

$$\epsilon^2(p_j)=|g_{ji}|^{-2}\sum_{i=1}^{N} G_{ji}^2\epsilon^2(f_i), \tag{34}$$

the $\epsilon^2(f_i)$ being conventionally calculated from the experimentally determined f_i (i.e., $\epsilon^2(f_i)=f_i$).

Our final solution has then the form shown in Fig. 5. The horizontal "errors" are simply the channel widths $\Delta\rho$ and the vertical "errors" are the $\epsilon(p_j)$ calculated from the $\epsilon(f_i)$ by Eq. (34).

Example: The Measurement of Mean Lives of Unstable Particles

M individual unstable particles are observed to decay in the gas of a very large cloud chamber. It is desired to find the relative abundances and the mean lives of the different species of unstable particles present. In this case, $F_i(t^{(i)})\Delta t$ is the observed relative frequency with which the particle life span t lies between $t^{(i)}-\Delta t/2$ and $t^{(i)}+\Delta t/2$, $P_j(\tau^{(j)})\Delta\tau$ is the unknown relative abundance of particles with mean life τ between $\tau^{(j)}-\Delta\tau/2$ and $\tau^{(j)}+\Delta\tau/2$, and

$$G(\tau;t)=\tau^{-1}\exp\{-t/\tau\}. \tag{35}$$

Then, exactly as in the previous example, we have, approximately,

$$F_i(t^{(i)})=\sum_{j=1}^{N} P_j(\tau^{(j)})G(\tau^{(j)};t^{(i)})\Delta\tau, \tag{36}$$

whence, setting $f_i=F_i(t^{(i)})$, $P_j(\tau^{(j)})=p_j$, $(\Delta\tau)G(\tau^{(j)}, t^{(i)})=g_{ji}$, Eq. (33) follows as before (analogous to the previous case, one can take $\Delta\tau=\Delta t\equiv$life-span channel width in the $F_j(t^{(j)})$ histogram). If now, for example, $p_j\Delta\tau=P_j(\tau^{(j)})\Delta\tau$ is "large" (≈ 1) for only one $\tau^{(j)}$, one has reliable evidence of the existence of only a single species of unstable particles with mean-life near $\tau^{(j)}$.

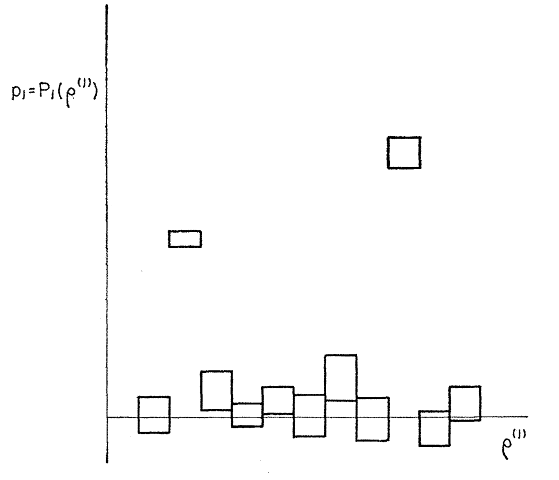

FIG. 5. A possible result of the calculation described in the text concerning the spectrum of masses incident on a cloud chamber. The relative frequency of occurrence of different values of $\rho^{(j)}\sim m_j^{-(1-\alpha)}$ is plotted as a function of $\rho^{(j)}$. Note that the result is consistent with two discrete mass values.

IV. CASE OF COMPLETELY UNKNOWN AND OF DELTA-FUNCTION TYPE PRIOR PROBABILITY DISTRIBUTION

In this section we shall treat two related problems. In both problems, a measurement of a set of the x_i is made on a single system which is one of a class of systems, the value of the parameter θ being fixed for the set of the x_i. It is desired to estimate the value of θ corresponding to the set x_i. The difference between the problems lies in our knowledge of the (prior) probability, $P(\theta)d\theta$, that θ lies in $d\theta$. Problems of type 1 are those in which we have no prior knowledge whatever about the value of θ for any of the systems. Problems of type 2 are those in which we know that in *each* system of the class of systems considered, the value of θ is fixed and the same, i.e., $P(\theta)d\theta=\delta(\theta-\theta_0)d\theta$; however the numerical value of θ_0 is unknown.

The occurrence of situations in which $P(\theta)=\delta(\theta-\theta_0)$ is generally accepted by physicists as a valid description of the properties of a certain class of systems (e.g., a certain type of elementary particles) about which it is known that several parameters other than θ (i.e., ψ, $\chi\cdots$) have the same value for all the systems in the class. The systems are then assumed to be identical in every way and are therefore characterized by the same value of the θ parameter ($=\theta_0$). Thus in the example of elementary particles, θ may be the mean life of the particular type of particles, ψ the mass of this type, χ the charge, etc.

Problem of Type 1

A μ meson enters a multiplate cloud chamber and scatters through angles ϕ_i upon traversing the n plates of the chamber. It is desired to find the product $\Pi=pc\beta$ for the meson. Here Π is the same for every member of the set ϕ_i; however, we do not know the (prior) probability that Π lies in $d\Pi$ independent of the data on the particular scattering angles ϕ_i.

Problem of Type 2

Measurements are made of the life spans (t_i) of a number, N, of "μ mesons" selected by the parent particles (π mesons) from which they are born, the daughter particles (electrons) into which they decay, etc., i.e., the identification of these particles as "μ mesons" is independent of the measurements of the t_i. It is desired to find the mean life, τ, of the μ mesons. Here we know that $P(\tau)d\tau=\delta(\tau-\tau_0)d\tau$ by our selection or sorting procedure but we do not know the numerical value of τ_0.

We first consider problems of type 1, i.e., problems where no independent information whatever is available about $P(\theta)$. The first thing that is evident is that *in* *this case there is no rigorous way to find* $H(x_i;\theta)$, i.e., there is no way to estimate θ and to quote a standard deviation in this estimate which may not be grossly in error. To justify this statement, we need only examine Eq. (2b) for $H(x_i;\theta)$. Since we have no information about $P(\theta)$, we must make some assumption about the form of $P(\theta)$ in order to estimate θ (e.g., we might take $P(\theta)=\text{constant}$). Having made this assumption, it is then a simple matter to use one of the methods described in Sec. II to estimate θ, and then to evaluate the standard deviation in this estimate. However, $P(\theta)$ is here actually unknown, and may have a form which makes our estimate of θ differ from the true θ by an amount large compared to the quoted standard deviation (e.g., $P(\theta)$ may be sharply peaked for values of θ at which $G(\theta;x_i)$ is flat and small.) *Hence, we see immediately that there is a certain degree of arbitrariness in any procedure one adopts in the case where* $P(\theta)$ *is completely unknown; this arbitrariness can eventually be removed only if enough individual measurements of the set* x_i *are made so that, as in the previous section, an integral equation for* $P(\theta)$, *Eq. (23), can be set up and solved.*

Let us, however, consider a possible, yet ultimately arbitrary, procedure, in the case in which $P(\theta)$ is completely unknown and only one measurement of the set x_i is available. We suggest, essentially following Laplace, that in this case, the least unreasonable choice for $P(\theta)$ is $P(\theta)=\text{constant}$. We further suggest that one take θ_{est} as $\theta_{\text{m.p.}}$, defined by (see Eq. (4))

$$\left.\frac{\partial H(x_i;\theta)}{\partial\theta}\right]_{\theta=\theta_{\text{m.p.}}}=\text{const}\left.\frac{\partial G(\theta;x_i)}{\partial\theta}\right]_{\theta=\theta_{\text{m.p.}}}=0, \quad (37)$$

and that one quote a "statistical error" or "figure of merit" for the estimate, $\epsilon(\theta_{\text{m.p.}})$, given by (see below—Eq. (47) *et seq.*)

$$\epsilon(\theta_{\text{m.p.}})=\left\{-\left.\frac{\partial^2}{\partial\theta^2}\log G(\theta;x_i)\right]_{\theta=\theta_{\text{m.p.}}}\right\}^{-\frac{1}{2}}. \quad (38)$$

In view of the necessary arbitrariness of this or any other procedure for estimating θ when $P(\theta)$ is completely unknown, it is always well to indicate explicitly that $P(\theta)$ has been taken as constant in the derivation of such an estimate.

One reason for suggesting the estimate of Eq. (37) is that this estimate has an interesting property, observed by many authors and by R. A. Fisher,[3] in particular, which allows one to resolve a certain ambiguity in the problem. The ambiguity in question is the following.

In any physical problem, the choice of θ is usually not unique. There may be many different physical quantities, related one to another in a simple one-to-one fashion, which might be chosen as θ. For example, in the case of radioactive decay of a group of nuclei of the same species, θ can be taken as either the mean life τ, or the reciprocal mean life, $\lambda=1/\tau$. If one assumes $P_1(\tau)$ is a constant, this is not consistent with taking $P_2(\lambda)$ constant. Indeed,

$$P_1(\tau)=P_2(\lambda)\left|\frac{d\lambda}{d\tau}\right|=P_2(\lambda)\tau^{-2}.$$

Hence, $P_1(\tau)=\text{constant}$ and $P_2(\lambda)=\text{constant}$ are two very different physical assumptions.

If one now takes τ as θ, and the corresponding $P_1(\tau)$ as constant, one can find a $\tau_{\text{m.p.}}$. If one takes λ as θ, and the corresponding $P_2(\lambda)$ as constant, one can similarly find $\lambda_{\text{m.p.}}$. However, it turns out that $\lambda_{\text{m.p.}}=1/\tau_{\text{m.p.}}$, so that the arbitrariness in the choice of the parameter θ (with the corresponding $P(\theta)$ always taken as constant) does not lead to any ambiguity in the estimate $\theta_{\text{m.p.}}$. This property holds in general and is perhaps the strongest reason for adopting the $\theta_{\text{m.p.}}$ estimate.

To prove the last remark we let $\theta=\theta(\phi)$ be any function of ϕ, where ϕ is a new parameter. We must then prove that $\theta_{\text{m.p.}}$ is equal to $\theta(\phi_{\text{m.p.}})$, where $\theta_{\text{m.p.}}$ is calculated assuming $P_\theta(\theta)=\text{constant}$, and $\phi_{\text{m.p.}}$ is calculated assuming $P_\phi(\phi)=\text{constant}$.

$\theta_{\text{m.p.}}$ is defined by Eq. (37), i.e.,

$$\left.\frac{\partial G(\theta;x_i)}{\partial\theta}\right]_{\theta=\theta_{\text{m.p.}}}=0, \quad (39)$$

and $\phi_{\text{m.p.}}$ is defined, similarly, by

$$\left.\frac{\partial G(\theta(\phi);x_i)}{\partial\phi}\right]_{\phi=\phi_{\text{m.p.}}}=0. \quad (40)$$

But

$$\frac{\partial G}{\partial\phi}(\theta(\phi);x_i)=\frac{\partial G}{\partial\theta}(\theta(\phi);x_i)\cdot\frac{d\theta(\phi)}{d\phi}.$$

If $d\theta(\phi)/d\phi$ does not equal zero, we can write Eq. (40) as

$$\left.\frac{\partial G}{\partial\theta}(\theta(\phi);x_i)\right]_{\phi=\phi_{\text{m.p.}}}=0, \quad (41)$$

so that from Eqs. (39) and (41), it is evident that

$$\theta(\phi_{\text{m.p.}})=\theta_{\text{m.p.}}, \quad (42)$$

which is to be proved.

We now treat problems of type 2, ($P(\theta)=\delta(\theta-\theta_0)$ with θ_0 unknown). Almost all physical problems eventually reduce to this type since, when the physical situation is sufficiently well understood, a large number of particles (or systems of particles) all with the same θ can be isolated. The physicist then conducts his further investigations on these "completely sorted" particles or systems of particles only.[4]

This problem is no different in principle from the one treated in Sec. III. We define $G(\theta;x)dx$ as the proba-

[3] Fisher, R. A., *Contributions to Mathematical Statistics* (John Wiley and Sons, Inc., New York, 1950).

[4] A general comment can be made regarding this "completely sorted" situation where $P(\theta)=\delta(\theta-\theta_0)$ (θ_0 unknown). Using Eq. (23a), one has $Q(x)=\int P(\theta)G(\theta;x)d\theta=G(\theta_0;x)$, so that if $Q(x)$ is found by the procedure described in Sec. III, a previously unknown functional form of $G(\theta;x)$ (the probability of finding x, knowing θ) may be determined. This method is clearly followed in deducing physical laws from experimental data. On the other hand, it is quite evident from Eq. (23a) that if nothing is known concerning $P(\theta)$, a previously unknown $G(\theta;x)$ cannot be deduced from the experimental data summarized by $Q(x)$. Unfortunately, this latter situation is usually encountered when the "sorting" is incomplete.

bility, for fixed θ, that the measured quantity x lies in dx. (For simplicity, we again assume that the set x_i measured in any single experiment has only one member, x.) Exactly as before, we measure a large number of x's, $x^{(1)}, \cdots, x^{(M)}$, and so can define a distribution in x, $F(x)=Q(x)$. In this case, however, we know that θ is unique, and all we wish to find is this unique value of θ.

The solution for $P(\theta)$ is (from Eq. (23b))

$$P(\theta)=\int F(x)\{G(\theta;x)\}^{-1}dx, \tag{43}$$

and, of course, we would hope that this solution is consistent with the original assumption that $P(\theta)$ is a δ function about the unique value of $\theta(\theta=\theta_0)$. Moreover θ_0 can now be immediately found. If, on the other hand, our solution is not consistent with the δ function assumption, we immediately suspect a systematic error of some kind.

The rigorous solution to the problem, given by Eq. (43), suffers however from several disadvantages:

1. It is rather laborious for practical calculation. The steps that have to be carried out are summarized in the examples done in Sec. III.

2. It is difficult to assign a "statistical error" or "figure of merit" to the estimate of θ. There is strictly no meaningful probability statement that can be made about any estimate in the case where $P(\theta)$ is known to be a single δ function, e.g., the statement, "θ has a probability of 0.1 to be greater than, say, θ_1" is obviously meaningless since this probability, by hypothesis, is either 1 or 0. However, it is still useful to have some number which is a "figure of merit" for the estimate. This number cannot have a relation to any statements about probability (=relative frequency), but instead must indicate, in some well-defined but ultimately arbitrary sense, the experimenter's estimate of the over-all precision which may be assigned to the value of θ.

There is, however, another approach to the present problem, which, as we shall see, is free from both of these disadvantages. In this approach one defines, essentially following R. A. Fisher,[3] the "likelihood function," $G(\theta;x_1x_2\cdots x_M)$, as

$$\begin{aligned}G(\theta;x_1\cdots x_M)&\\ &\equiv G(\theta;x^{(1)})G(\theta;x^{(2)})\cdots G(\theta;x^{(M)}),\\ &\qquad (x_i\equiv x^{(i)});\end{aligned} \tag{44}$$

$G(\theta;x_1\cdots x_M)$ is formally identical with the previously defined function which gives the probability. that for given θ, the set $x_i=x^{(i)}$ lies in dx_i. This likelihood function, considered as a function of θ with the x_i given, will, as we show later, have a very sharp maximum at some value of θ, $\theta=\theta_l$, if M is large enough. θ_l is then taken as the "maximum likelihood" estimate of the actual unique value of θ, θ_0. More precisely, θ_l is defined by the equation

$$\left.\frac{\partial G}{\partial\theta}(\theta;x_1\ldots x_M)\right]_{\theta=\theta_l}=0, \tag{45}$$

formally, the same equation which defines the estimate, $\theta_{\text{m.p.}}$, calculated from the corresponding $H(x_i;\theta)$ with $P(\theta)$ taken as constant (see Eq. (4)). A justification of this procedure of estimating θ_0 by θ_l, based on physical arguments, may now be attempted as follows:

Suppose that the likelihood function $G(\theta;x_1,\cdots,x_M)$ becomes an infinitely sharply peaked function in θ as M approaches infinity, i.e., suppose

$$\lim_{M\to\infty} G(\theta;x_1\ldots x_M)\sim\delta(\theta-\theta');$$

$$\theta'=\lim_{M\to\infty}\text{ function of }x_1\ldots x_M. \tag{46}$$

(In all of the examples treated in this paper, Eq. (46) is obviously satisfied.) Put in another way, Eq. (46) states that for sufficiently large M, the value of θ, $\theta=\theta'(x_1\cdots x_M)$, is determined uniquely by a set of values of the x_j. It is evident from Eq. (45) that θ_l as defined by Eq. (45) is then the same as θ' defined in Eq. (46), at least in the case where M is sufficiently large. Hence, if Eq. (46) holds, the likelihood function gives (for finite M):

(1) a relatively easy way to find the estimate $\theta_l(x_1,\cdots,x_M)$, and

(2) a well-defined but nevertheless ultimately arbitrary "figure of merit" or "statistical error" for this estimate. For with M large but finite, the "width" of the likelihood function tells us how close we are to the optimum situation where $G(\theta;x_1\cdots x_M)$ becomes an infinitely sharply peaked function of θ. We therefore suggest that this "figure of merit" be defined as the quantitative measure of the width of $G(\theta;x_1\cdots x_M)$, i.e., we define[5]

$$\begin{aligned}\epsilon(\theta_l)&\equiv\left\{\left.-\frac{\partial^2}{\partial\theta^2}\log G(\theta;x_1\ldots x_M)\right]_{\theta=\theta_l}\right\}^{-\frac{1}{2}}\\ &=\left\{\left.-G^{-1}\frac{\partial^2G}{\partial\theta^2}\right]_{\theta=\theta_l}\right\}^{-\frac{1}{2}}.\end{aligned} \tag{47}$$

This definition of $\epsilon(\theta_l)$ is motivated by the fact that for large but finite M, $G(\theta;x_1\cdots x_M)$ is well approximated

[5] In Eq. (47) we have implicitly assumed that the set x_i is given exactly by the measurements. However, if any x_i is itself uncertain as a consequence of the limited precision of the x_i measurement, then the $\epsilon(\theta_l)$ has an additional contribution. This contribution is discussed in Sec. V and shown to be small with proper design of the experiment, in cases of practical interest.

(as in all the examples of this paper) by

$$G(\theta;x_1\ldots x_M)\cong G(\theta_l;x_1\ldots x_M)$$
$$\times\exp\left\{\frac{\partial^2}{\partial\theta^2}\log G(\theta;x_1\ldots x_M)\right]_{\theta=\theta_l}\cdot\tfrac{1}{2}(\theta-\theta_l)^2\Bigg\}. \quad (48)$$

The expression for θ_l given by Eq. (45) (with the "statistical error" given by Eq. (47)) is (for M large but finite) a good approximation to θ' when Eq. (46) is true. *However, Eq. (46) must hold in every case where θ actually has a unique value $\theta=\theta_0$, with, in addition, $Lim_{M\rightarrow\infty}\theta'(x_1\cdots x_M)$ being just equal to θ_0.* For, suppose Eq. (46) with $\theta'=\theta_0$ did not hold as $M\rightarrow\infty$. This would mean that, as $M\rightarrow\infty$, θ was not uniquely determined as θ_0 by the set of the x_i, and this is counter to our hypothesis that each member of the set of the x_i obeys the probability distribution $G(\theta;x)$ with $\theta=\theta_0$. Hence as $M\rightarrow\infty$, the likelihood function $G(\theta;x_1\cdots x_M)$ must approach an infinitely sharply peaked function of θ centered at $\theta'(x_1,\cdots,x_M)=\theta_0$, so that $\theta_l(x_1\cdots x_M)\rightarrow\theta_0$ as $M\rightarrow\infty$.

As we have mentioned previously, the physical quantity one chooses as θ is not unique; thus there are many quantities related in a one-to-one fashion any one of which can be chosen as the parameter θ. The previous discussion shows that the maximum likelihood estimate of θ, θ_l given by Eq. (45) is independent of this lack of uniqueness. However, $\epsilon(\theta_l)$, as given by Eq. (47), will in general depend in a sensitive way on the choice of the parameter θ; thus from Eqs. (47), (39), (40), we have

$$\epsilon(\theta_l)=\epsilon(\phi_l)\left|\frac{d\theta(\phi)}{d\phi}\right]_{\phi=\phi_l}\Bigg|. \quad (49)$$

Example: Measurement of Mean Life of a Moving Particle in a Cloud Chamber; the Mean Life Assumed Unique

In this example we assume that by means of an experimental sorting procedure we have isolated a set of identical particles each of which has the same mean life, $1/\lambda=1/\lambda_0$. If a particle has mean life $1/\lambda$, then

$$\exp\{-\lambda t\}\lambda dt \quad (50)$$

is the probability that it will decay in dt at t. However, if the observing chamber is of finite size, and if T is the maximum value of t that can be observed, then the probability of observing a decay in dt at $t\leqq T$ becomes

$$G(\lambda;t)dt=\exp\{-\lambda t\}(1-\exp\{-\lambda T\})^{-1}\lambda dt, \quad (51)$$

and the corresponding likelihood function is ($t^{(j)}\equiv t_j$ is the life span of the jth particle and T_j is the maximum value of $t^{(j)}$ that could be observed)

$$G(\lambda;t_1\cdots t_M)=\prod_{j=1}^{M}\exp\{-\lambda t^{(j)}\}(1-\exp\{-\lambda T_j\})^{-1}\lambda. \quad (52)$$

The maximum likelihood estimate of λ is now given by (Eq. (45)),

$$1/\lambda_l=M^{-1}\sum_{j=1}^{M}t^{(j)}+M^{-1}\sum_{j=1}^{M}T_j(\exp\{\lambda_l T_j\}-1)^{-1}. \quad (53)$$

Equation (53) can be solved for λ_l by a process of iteration. In the case where $T_j\gg 1/\lambda_l$, the solution for λ_l becomes

$$1/\lambda_l\cong t_{\text{mean}}+M^{-1}\sum_{j=1}^{M}T_j(\exp\{T_j/t_{\text{mean}}\}-1)^{-1}, \quad (54)$$

where

$$t_{\text{mean}}\equiv M^{-1}\sum_{j=1}^{M}t^{(j)}.$$

For the "figure of merit," Eq. (47) gives

$$\epsilon(\lambda_l)=\lambda_l(M-\sum_{j=1}^{M}(\lambda_l T_j)^2\exp\{-\lambda_l T_j\}/[1-\exp\{-\lambda_l T_j\}]^2)^{-\frac{1}{2}}, \quad (55)$$

where the effect, for given M, of the finite T_j in increasing $\epsilon(\lambda_l)$ is clearly shown.[6]

Example

If the decay of the individual particles is detected by a counter with background rate b, then $G(\lambda;t)dt$, the probability per particle of observation of a count in dt, is

$$G(\lambda;t)dt=(b+\lambda\exp\{-\lambda t\})(bT+1-\exp\{-\lambda T\})^{-1}dt;$$
$$\int_0^T G(\lambda;t)dt=1, \quad (56)$$

whence

$$G(\lambda;t_1\cdots t_M)=\prod_{j=1}^{M}(b+\lambda\exp\{-\lambda t^{(j)}\})(bT+1-\exp\{-\lambda T\})^{-1}. \quad (57)$$

($T_j=T=$maximum possible observed value of $t^{(j)}\equiv t_j$; T is determined from the time at which the counting rate is reduced to a value near the background rate.) The maximum likelihood estimate of λ is, if $b\ll\lambda_l\exp(-\lambda_l T)$,

$$1/\lambda_l\cong t_{\text{mean}}+\frac{T\exp\{-\lambda_l T\}}{bT+1-\exp\{-\lambda_l T\}}$$
$$+\frac{b}{\lambda_l^2}M^{-1}\sum_{j=1}^{M}(1-\lambda_l t_j)\exp\{\lambda_l t_j\}; \quad (58)$$

and since for large M,

$$M^{-1}\sum_{j=1}^{M}f(t_j)\approx\int_0^T f(t)G(\lambda_l;t)dt \quad \text{(where } f(t) \text{ is any function of } t\text{)}, \quad (59)$$

we obtain

$$1/\lambda_l\cong t_{\text{mean}}+\frac{T\exp\{-\lambda_l T\}}{bT+1-\exp\{-\lambda_l T\}}$$
$$+\frac{bT}{1-\exp\{-\lambda_l T\}}\left(\frac{T}{2}-\lambda_l^{-1}\right). \quad (60)$$

For small b, $\epsilon(\lambda_l)$ is given in terms of λ_l again by Eq. (55).

Example

The very fact that physical situations exist in which a "sorting" procedure can be performed to separate particles with a unique value of θ implies that workers in different laboratories can carry out the same sorting procedure, and then estimate the unique value of θ. The question arises, "How must one proceed in order to combine these estimates of θ from the different laboratories to form a 'combined estimate' and what is the 'statistical error' or 'figure of merit' of this 'combined estimate'?"

We will show how one proceeds in a simple case and quote the results for one other simple case.

Consider that N different laboratories have carried out the same sorting procedure to isolate a group of particles and have then estimated the unique value of the reciprocal mean life λ_0. Calling the jth maximum-likelihood estimate of λ_0, $(\lambda_l)_j$, and

[6] See Fretter, May, and Nakada, Phys. Rev. **89**, 168 (1953); W. L. Alford and R. B. Leighton, Phys. Rev. **90**, 622 (1953); M. S. Bartlett, Phil. Mag. **44**, 249 (1953).

assuming that the maximum measurable time T is infinite, we find the likelihood function for the jth laboratory is (see Eqs. (52) and (53))

$$G_j(\lambda;t_k)=\lambda^J\exp\{-\lambda\sum_{k=1}^{J}t_k\}, \tag{61}$$

where J is the number of particles seen to decay by the jth laboratory, i.e., $J=J(j)$.

From Eq. (61)

$$1/(\lambda_l)_j=J^{-1}(j)\sum_{k=1}^{J(j)}t_k \tag{62}$$

with "statistical error,"

$$\epsilon[(\lambda_l)_j]=J^{-\frac{1}{2}}(j)(\lambda_l)_j.$$

The over-all likelihood function is

$$\begin{aligned}G(\lambda;(\lambda_l)_j)&=\prod_{j=1}^{N}\lambda^{J(j)}\exp\{-\lambda\sum_{k=1}^{J(j)}t_k\}\\&=\lambda^M\exp\{-\lambda\sum_{j=1}^{N}J(j)/(\lambda_l)_j\},\end{aligned} \tag{63}$$

where

$$M\equiv\sum_{j=1}^{N}J(j). \tag{64}$$

It is evident from Eq. (63) that λ_l, the combined maximum likelihood estimate of λ_0, is given by

$$1/\lambda_l=M^{-1}\sum_{j=1}^{N}J(j)/(\lambda_l)_j \tag{65}$$

and

$$\epsilon(\lambda_l)=M^{-\frac{1}{2}}\lambda_l.$$

In practice, however, one does not in general know the values of the $J(j)$, hence it is useful to express our result for λ_l and $\epsilon(\lambda_l)$ in terms of the $(\lambda_l)_j$ and $\epsilon[(\lambda_l)_j]$. From Eqs. (63–65) we obtain

$$1/\lambda_l=\frac{\sum_{j=1}^{N}(\lambda_l)_j/\{\epsilon[(\lambda_l)_j]\}^2}{\sum_{j=1}^{N}\{(\lambda_l)_j\}^2/\{\epsilon[(\lambda_l)_j]\}^2} \tag{66}$$

and

$$\{\epsilon(\lambda_l)\}^{-2}=\lambda_l^{-2}\sum_{j=1}^{N}\{(\lambda_l)_j\}^2/\{\epsilon[(\lambda_l)_j]\}^2.$$

Referring now to the problem of combining mass values from N different laboratories, we call $(\rho_l)_j$ the maximum likelihood estimate of the parameter ρ_0 by the jth laboratory (according to example of Sec. III, $\rho_0\sim(m)^{-0.447}$), and the corresponding "statistical error." A calculation similar to the one carried out above gives

$$\rho_l^2=\frac{\sum_{j=1}^{N}\{(\rho_l)_j\}^4/\{\epsilon[(\rho_l)_j]\}^2}{\sum_{j=1}^{N}\{(\rho_l)_j\}^2/\{\epsilon[(\rho_l)_j]\}^2} \tag{67}$$

and

$$\{\epsilon(\rho_l)\}^{-2}=\rho_l^{-2}\sum_{j=1}^{N}\{(\rho_l)_j\}^2/\{\epsilon[(\rho_l)_j]\}^2,$$

where ρ_l is the combined maximum likelihood estimate of ρ_0 and $\epsilon(\rho_l)$ is the corresponding "statistical error."

We note from Eqs. (66) and (67) that the rule for combining estimated mass values is not the same as the rule for combining estimated reciprocal mean lives. In general, the rule for combining the $(\theta_{\text{est}})_j=(\theta_l)_j$ depends on the functional form of the corresponding $G(\theta;x_1,\cdots,x_{J(j)})$.

V. EFFECT ON THE ESTIMATES OF IMPERFECT EXPERIMENTAL PRECISION IN THE DIRECTLY MEASURED QUANTITIES

It sometimes arises in practice that one can easily derive from theory an expression for $G'(\theta;y_1,\cdots,y_N)\equiv G'(\theta;y_i)$, where $G'(\theta;y_i)dy_i$ is the probability that the quantities y_i lie in dy_i for given θ, but that the y_i cannot be "measured exactly." However, one can "measure exactly" a set of quantities x_i which are related to the y_i through another probability distribution. For instance one may have

$$\begin{aligned}g(y_i;x_i)dx_i=(2\pi)^{-N/2}\prod_{j=1}^{N}\sigma_j^{-1}\\\times\exp\{-(x_j-y_j)^2/2\sigma_j^2\}dx_j;\end{aligned} \tag{68}$$

i.e., each of the x_j may be distributed statistically in Gaussian fashion about the y_j. In this case, the likelihood function becomes

$$\begin{aligned}G(\theta;x_1\ldots x_N)&\equiv G(\theta;x_i)\\&=\int\cdots\int G'(\theta;y_i)g(y_i;x_i)dy_i.\end{aligned} \tag{69}$$

Let us now expand $G'(\theta;y_i)$ in a Taylor series about the values $y_1=x_1$, $y_2=x_2$, etc.:

$$\begin{aligned}G'(\theta;y_i)=G'(\theta;y_i=x_i)+\sum_{j=1}^{N}(y_j-x_j)\frac{\partial G'(\theta;x_i)}{\partial x_j}\\+\tfrac{1}{2}\sum_{j,k=1}^{N}(y_j-x_j)(y_k-x_k)\frac{\partial^2G'(\theta;x_i)}{\partial x_j\partial x_k}+\ldots;\end{aligned} \tag{70}$$

and let us introduce the additional assumption that $g(y_i;x_i)$, considered as a function of the y_i, is sharply peaked for $y_i\approx x_i$, whereas $G'(\theta;y_i)$ is relatively slowly varying in the region of $y_i\approx x_i$. Then, substituting Eq. (70) into Eq. (69) and neglecting terms above 2nd order in (y_j-x_j) in Eq. (70), we get

$$G(\theta;x_i)=G'(\theta;x_i)+\tfrac{1}{2}\sum_{j=1}^{N}\sigma_j^2\frac{\partial^2G'(\theta;x_i)}{\partial x_j^2}+\ldots. \tag{71}$$

Equation (71) can be used in Eq. (2) to derive the posterior probability $H(\theta;x_i)$ or it can be used directly to estimate θ from the exactly measured x_i in the case when θ is independently known to be unique (see Sec. IV). If θ is unique the maximum likelihood estimate of θ, θ_l, is given by

$$\left[\frac{\partial G'(\theta;x_i)}{\partial\theta}+\tfrac{1}{2}\sum_{j=1}^{N}\sigma_j^2\frac{\partial^3G'(\theta;x_i)}{\partial\theta\partial x_j^2}+\ldots\right]_{\theta=\theta_l}=0, \tag{72}$$

while if θ obeys a known prior distribution law, $P(\theta)$, θ_l is given by

$$\left[\frac{\partial H'(\theta;x_i)}{\partial\theta}+\tfrac{1}{2}\sum_{j=1}^{N}\sigma_j^2\frac{\partial^3 H'(\theta;x_i)}{\partial\theta\partial x_j^2}+\ldots\right]_{\theta=\theta_l}=0, \quad (73)$$

where

$$H'(\theta;x_i)=P(\theta)G'(\theta;x_i)\Big/\int P(\theta)G'(\theta;x_i)d\theta.$$

The relative "error" in θ_l, $\epsilon(\theta_l)/\theta_l$ will not be changed appreciably if the correction terms in Eqs. (72) and (73) are not large, i.e., if σ_j^2/x_j^2 is not large. If the correction terms are large, the formulas in (72) and (73) are inadequate; therefore in either case it is of little use to derive from them a new formula for $\epsilon(\theta_l)$. In this situation, the problem must be re-examined and improved methods to measure the y_i devised.

Example from the Theory of Multiple Coulomb Scattering. "Noise Level Scattering"

Let us consider again the example treated at the end of Sec. II. We shall now assume that each of the n measured angles of "apparent" multiple scattering ϕ_i obeys a Gaussian distribution about the corresponding "real" angle of multiple scattering ψ_i, with standard deviation σ. The magnitude of σ is a measure of the "noise level scattering." In this case (see Eq. (73) and Eqs. (13) and (15)),

$$\begin{aligned}&\theta\rightarrow\Pi\\&y_i\rightarrow\psi_i\\&x_i\rightarrow\phi_i\\&\sigma_j\rightarrow\sigma=\text{constant}\end{aligned} \quad (74)$$

and

$$H'(\Pi;\phi_i)=\text{const}\Pi^m\exp\left\{-\frac{\Pi^2}{A^2}\sum_{i=1}^{n}\phi_i^2\right\},$$

where $m=n-\gamma$ as before, and we assume $\Pi_0\ll\Pi_{\text{m. p.}}$ Substituting in Eq. (73), and remembering that the second term in Eq. (73) is small, we find

$$\Pi_{\text{m. p.}}=(m/2)^{\frac{1}{2}}A\{\sum_{i=1}^{n}\phi_i^2-(n-2\gamma)\sigma^2\}^{-\frac{1}{2}}. \quad (75)$$

Comparison of Eq. (16) with Eq. (75) shows that the estimate of $\Pi=pc\beta$ is increased when one takes into account the noise-level scattering.

The formula given in Eq. (75) does not at first glance agree with the formula given by Olbert.[1] (We have σ^2, where Olbert has $2\sigma^2$.) The reason for this apparent discrepancy is a difference in the definition of σ from the one used by Olbert. A possible method of measuring the σ we have defined in Eq. (68) is given below.

Allow N tracks of very great momentum to penetrate the plates of the cloud chamber. (The momentum must be great enough so that the real scattering is negligible compared to the noise-level scattering.) In this case the distribution in scattering angles is given by Eq. (68) with $y_i\rightarrow\psi_i\cong 0$ and $x_i\rightarrow\phi_i$; i.e.,

$$g(\sigma;\phi_i)d\phi_i=(2\pi)^{-N/2}\sigma^{-N}\exp\{-(1/2\sigma^2)\sum_{j=1}^{N}\phi_j^2\}\prod_{j=1}^{N}d\phi_j \quad (76)$$

is the probability that for fixed (unique but unknown) σ each of the ϕ_i lie in $d\phi_i$. σ can now be estimated from $g(\sigma;\phi_i)$ by the method of maximum likelihood (Sec. IV), and is given by

$$\sigma_l=\{N^{-1}\sum_{j=1}^{N}\phi_j^2\}^{\frac{1}{2}}. \quad (77)$$

VI. THE "ESTIMATE DISTRIBUTION"

Suppose an experiment is performed in which the information is available that the parameter θ is unknown but fixed and unique ($\theta=\theta_0$) for the set $x_i\equiv x^{(i)}$. The likelihood function for the experiment may then be constructed, namely $G(\theta;x_i)dx_i$, and an estimate of the parameter $\theta_{\text{est}}=f_{\text{est}}(x_1,\cdots,x_N)$, e.g., the maximum likelihood estimate $\theta_l=\theta_{\text{est}}$, may be calculated. We now ask "What is the probability that θ_{est} lies in $d\theta_{\text{est}}$ if the value of the parameter is assumed to be θ_0?" This question may be answered by constructing the function $R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}$ which is the probability that θ_{est} lies in $d\theta_{\text{est}}$ for θ assumed to be θ_0; $R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}$ is given by Eq. (78):

$$R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}=\int\ldots\int dx_1\ldots dx_N G(\theta_0;x_1\ldots x_N), \quad (78)$$

where the integration over the set dx_i is carried out under the restriction imposed by the equation

$$\theta_{\text{est}}\leqq f_{\text{est}}(x_1\ldots x_N)\leqq\theta_{\text{est}}+d\theta_{\text{est}}. \quad (79)$$

Thus,

$$\begin{aligned}R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}\\=\Big[\int\ldots\int dx_1\ldots dx_N G(\theta_0;x_1\ldots x_N)\\\times\delta(\theta_{\text{est}}-f_{\text{est}}(x_1\ldots x_N))\Big]d\theta_{\text{est}}.\end{aligned} \quad (80a)$$

The probibility function $R(\theta_0;\theta_{\text{est}})$ will be called the "estimate distribution" and can be used to answer questions of the type: "What will be the probability that θ_{est} is, e.g., equal to or greater than some θ_a if the value of the parameter θ is assumed to be θ_0?" This probability, $J(\theta_0,\theta_a)$, is obtained directly from the estimate distribution $R(\theta_0;\theta_{\text{est}})$, and is

$$J(\theta_0,\theta_a)=\int_{\theta_a}^{\infty}R(\theta_0,\theta_{\text{est}})d\theta_{\text{est}}, \quad (80b)$$

with $J(\theta_0,-\infty)=1$.

Using Eq. (80a), we find averages over the estimate distribution can be expressed as follows:

$$\begin{aligned}\langle\theta_{\text{est}}\rangle&\equiv\int\theta_{\text{est}}R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}\\&=\int f_{\text{est}}(x_1\ldots x_N)G(\theta_0;x_1\ldots x_N)dx_1\ldots dx_N,\end{aligned} \quad (81a)$$

$$\begin{aligned}\langle(\theta_{\text{est}}-\theta_0)^2\rangle&\equiv\int(\theta_{\text{est}}-\theta_0)^2R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}\\&=\int(f_{\text{est}}(x_1\ldots x_N)-\theta_0)^2\\&\quad\times G(\theta_0;x_1\ldots x_N)dx_1\ldots dx_N,\end{aligned} \quad (81b)$$

etc. It can then be proved[7] (under suitable restrictions)

[7] Cramer, H., *Mathematical Methods oj Statistics* (Princeton University Press, Princeton, 1946), Chapters 32 and 33.

that for $\theta_{\text{est}}=f_{\text{est}}(x_1, \cdots, x_N)=\theta_l$, θ_{est} "converges in probability" to the true value of $\theta=\theta_0$, i.e.

$$\lim_{N\to\infty}\langle\theta_l\rangle\to\theta_0. \tag{81c}$$

One can also show[7] (under suitable restrictions) that

$$\langle(\theta_l-\theta_0)^2\rangle<\langle(\theta_{\text{est}}-\theta_0)^2\rangle \tag{81d}$$

for any fixed finite N and any $\theta_{\text{est}}\neq\theta_l$. Equations (81c) and (81d) constitute additional powerful arguments in favor of maximum likelihood estimation in the case of an unique but unknown parameter.

Suppose a single charged particle which stops in a multiplate cloud chamber has a certain angle of multiple scattering in each of the plates of the chamber. From these angles of multiple scattering (and the range) one can estimate the mass of the particle: θ_{est} = mass estimate = a definite function of the multiple-scattering angles. One can then calculate the probability that a particle of known mass, e.g., a proton, appears to exhibit a mass equal to or greater than the above θ_{est}, i.e., appears to scatter as little or less than the observed particle. This calculation involves constructing the estimate distribution for θ_{est} (Eqs. (80)) and hence the estimate distribution for the angles of multiple scattering. However, questions of the inverse type—"What is the probability that a particle which scatters the observed amount is, say, a proton?"—are meaningless since this probability is, by hypothesis, one or zero.

As a further illustration of the use of the estimate distribution, suppose a "sorting" has been made in two experiments with the result that for each experiment the information is available that the parameter θ has an unique value (which may be different, in general) in the two experiments. Then the simultaneous probability that the estimate of θ from the first experiment lies in $d\theta_{\text{est}}^{(1)}$ and that the estimate of θ from the second experiment lies in $d\theta_{\text{est}}^{(2)}$ for a unique value of θ, θ_0, assumed the same in both experiments, is

$$R(\theta_0;\theta_{\text{est}}^{(1)},\theta_{\text{est}}^{(2)})d\theta_{\text{est}}^{(1)}d\theta_{\text{est}}^{(2)}$$
$$=R(\theta_0;\theta_{\text{est}}^{(1)})d\theta_{\text{est}}^{(1)}R(\theta_0;\theta_{\text{est}}^{(2)})d\theta_{\text{est}}^{(2)}.$$

Example

Suppose two experiments are performed on the life spans of two sets of unstable particles. In addition, it is assumed on the basis of the sorting procedure used that within each experiment the particles are identical. It is now desired to know if the results of the two experiments are consistent with the assumption that the mean life of *both sets* of particles is the *same*, in other words, if the *two sets* of particles are the *same*. The probability in experiment 1 that the estimate of τ lies in $d\tau_{\text{est}}^{(1)}$ for a unique mean life assumed equal to τ_0 is given by Eq. (80a). We shall take this estimate to be

$$\tau_{\text{est}}^{(1)}=\tau_l^{(1)}=t_{\text{mean}}^{(1)}\equiv N^{-1}\sum_{i=1}^{N}t_i^{(1)}$$

(See Eq. (54) with $T_j\to\infty$), where N is the number of life spans measured in experiment 1. Then

$$R(\tau_0;\tau_{\text{est}}^{(1)})d\tau_{\text{est}}^{(1)}=\Big[\int dt_1^{(1)}\cdots dt_N^{(1)}\prod_{i=1}^{N}\tau_0^{-1}$$
$$\times\exp\{-t_i^{(1)}/\tau_0\}\delta(\tau_{\text{est}}^{(1)}-N^{-1}\sum_{j=1}^{N}t_j^{(1)})\Big]d\tau_{\text{est}}^{(1)}$$

and by use of the integral representation of the δ function

$$\delta(x-a)=(2\pi)^{-1}\int_{-\infty}^{\infty}\exp\{i\alpha(x-a)\}d\alpha,$$

the integral over $dt_1^{(1)}\cdots dt_N^{(1)}$ may be performed yielding

$$R(\tau_0;\tau_{\text{est}}^{(1)})d\tau_{\text{est}}^{(1)}$$
$$=\frac{i^N}{2\pi}\left(\frac{N\tau_{\text{est}}^{(1)}}{\tau_0}\right)\int_{-\infty}^{\infty}\frac{\exp\{i\beta\}d\beta}{\{\beta-iN\tau_{\text{est}}^{(1)}/\tau_0\}^N}\left(\frac{d\tau_{\text{est}}^{(1)}}{\tau_{\text{est}}^{(1)}}\right).$$

Finally, via the calculus of residues,[8]

$$R(\tau_0;\tau_{\text{est}}^{(1)})d\tau_{\text{est}}^{(1)}$$
$$=\frac{N^N}{(N-1)!}\left(\frac{\tau_{\text{est}}^{(1)}}{\tau_0}\right)^N\exp\left\{\frac{-N\tau_{\text{est}}^{(1)}}{\tau_0}\right\}\frac{d\tau_{\text{est}}^{(1)}}{\tau_{\text{est}}^{(1)}}. \tag{82}$$

In a similar manner, the estimate distribution for the second experiment, again for a unique mean life assumed equal to τ_0, is

$$R(\tau_0;\tau_{\text{est}}^{(2)})d\tau_{\text{est}}^{(2)}$$
$$=\frac{M^M}{(M-1)!}\left(\frac{\tau_{\text{est}}^{(2)}}{\tau_0}\right)^M\exp\left\{\frac{-M\tau_{\text{est}}^{(2)}}{\tau_0}\right\}\frac{d\tau_{\text{est}}^{(2)}}{\tau_{\text{est}}^{(2)}},$$

where M is the number of life spans measured in experiment 2. Thus

$$R(\tau_0;\tau_{\text{est}}^{(1)},\tau_{\text{est}}^{(2)})d\tau_{\text{est}}^{(1)}d\tau_{\text{est}}^{(2)}$$
$$=\frac{N^N}{(N-1)!}\frac{M^M}{(M-1)!}\frac{(\tau_{\text{est}}^{(1)})^N(\tau_{\text{est}}^{(2)})^M}{\tau_0^{N+M}}$$
$$\cdot\exp\left\{\frac{-N\tau_{\text{est}}^{(1)}+M\tau_{\text{est}}^{(2)}}{\tau_0}\right\}\frac{d\tau_{\text{est}}^{(1)}}{\tau_{\text{est}}^{(1)}}\frac{d\tau_{\text{est}}^{(2)}}{\tau_{\text{est}}^{(2)}}, \tag{83}$$

so that the simultaneous probability for experiment 1 to yield a value of τ_{est} between $\tau_{\text{est}}^{(1)}-\epsilon(\tau_{\text{est}}^{(1)})$ and $\tau_{\text{est}}^{(1)}+\epsilon(\tau_{\text{est}}^{(1)})$, and experiment 2 a value of τ_{est} between $\tau_{\text{est}}^{(2)}-\epsilon(\tau_{\text{est}}^{(2)})$ and $\tau_{\text{est}}^{(2)}+\epsilon(\tau_{\text{est}}^{(2)})$ is

$$\int_{\tau_{\text{est}}^{(2)}-\epsilon(\tau_{\text{est}}^{(2)})}^{\tau_{\text{est}}^{(2)}+\epsilon(\tau_{\text{est}}^{(2)})}\int_{\tau_{\text{est}}^{(1)}-\epsilon(\tau_{\text{est}}^{(1)})}^{\tau_{\text{est}}^{(1)}+\epsilon(\tau_{\text{est}}^{(1)})}R(\tau_0;\tau_{\text{est}}^{(1)\prime},\tau_{\text{est}}^{(2)\prime})$$
$$\times d\tau_{\text{est}}^{(1)\prime}d\tau_{\text{est}}^{(2)\prime}. \tag{84}$$

If now this last expression calculated from the observed $t_i^{(1)}$, $t_i^{(2)}$ (see Eqs. (54) and (55)) is "very much less than one," whatever the choice of τ_0, it is not "likely" that the mean life of *both* sets of particles is actually the *same*. If, on the other hand, a value of τ_0 can be found for which this expression is not "very much less than one," it is "likely" that *both* sets of particles have the *same* mean life, τ_0. (From Eq. (83) we find that in the optimum situation in which $\tau_{\text{est}}^{(1)}=\tau_{\text{est}}^{(2)}$, $\epsilon(\tau_{\text{est}}^{(1)})=\epsilon(\tau_{\text{est}}^{(2)})$, and τ_0 is taken as $\tau_{\text{est}}^{(1)}$, Eq. (84) gives approximately one-half.) Obviously, both of the preceding qualitative statements can be made quantitative by a definite (though ultimately arbitrary) numerical specification of "very much less than one," and "likely."

Example

Another interesting case of the use of the estimate distribution involves the following problem previously considered and solved by Sard and Sard.[9] In a cosmic-ray counting experiment the observed number of coincidences in a certain time interval is x,

[8] In this problem, the integral on the right-hand side of Eq. (78) may be performed directly because of the simple relationship for τ_{est}, namely

$$\tau_{\text{est}}=(1/N)\sum_{i=1}^{N}t_i.$$

In this case

$$R(\tau_0;\tau_{\text{est}})d\tau_{\text{est}}=\tau_0^{-N}\exp\{-N\tau_{\text{est}}/\tau_0\}\int dt_1\cdots dt_N,$$

where the integral is the volume in N-dimensional t space contained between the two hypersurfaces defined by τ_{est} and $\tau_{\text{est}}+d\tau_{\text{est}}$. This volume is proportional to $(\tau_{\text{est}})^{N-1}d\tau_{\text{est}}$, so that

$$R(\tau_0;\tau_{\text{est}})=A\left(\frac{\tau_{\text{est}}}{\tau_0}\right)^N\tau_{\text{est}}^{-1}\exp\{-N\tau_{\text{est}}/\tau_0\}$$

in agreement with Eq. (82). The normalization constant A may be determined by requiring that $\int R(\tau_0;\tau_{\text{est}})d\tau_{\text{est}}=1$. This simple method of evaluating the right-hand side of Eq. (78) will work whenever $G(\theta_0;x_i)$ can be written as a function of θ_0 and θ_{est} only.

[9] A. Sard and R. Sard, Rev. Sci. Instr. **20**, 526 (1949).

the calculated number of chance coincidences in the same time interval is β, and the parameter θ to be estimated is the mean number of true coincidences in the time interval. It then follows that the probability distribution for x, given θ, i.e., the likelihood function, is Poissonian:[9]

$$G(\theta;x)=\frac{(\theta+\beta)^x}{x!}\exp\{-(\theta+\beta)\}. \tag{85}$$

The maximum likelihood estimate of θ, θ_l, is then given by

$$\left.\frac{\partial G(\theta;x)}{\partial\theta}\right]_{\theta=\theta_l}=0;\quad\text{whence, }\theta_l=x-\beta. \tag{86}$$

By use of Eqs. (78), (79), the estimate distribution with $\theta_{\text{est}}=\theta_l=x-\beta$, is given as

$$R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}=\int G(\theta_0;x)dx$$

with

$$\theta_{\text{est}}+\beta\leqq x\leqq\theta_{\text{est}}+\beta+d\theta_{\text{est}},$$

so that

$$\begin{aligned}R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}&=G(\theta_0;\theta_{\text{est}}+\beta)d\theta_{\text{est}}\\&=\frac{(\theta_0+\beta)^{\theta_{\text{est}}+\beta}}{(\theta_{\text{est}}+\beta)!}\exp\{-(\theta_0+\beta)\}d\theta_{\text{est}}.\end{aligned} \tag{87}$$

Thus the Poissonian property of $G(\theta_0;x)$ implies that $R(\theta_0;\theta_{\text{est}})$ with $\theta_{\text{est}}=\theta_l$ is also Poissonian in the variable $\theta_{\text{est}}+\beta$. Further the "statistical error" or "figure of merit" assigned by Eq. (47) to θ_l is given by

$$\epsilon^2(\theta_l)=x, \tag{88}$$

so that we can write

$$\theta_0\approx(x-\beta)\pm x^{\frac{1}{2}}$$

in agreement with the results of Sard and Sard.[9] It is also possible to calculate (for $\theta_{\text{est}}=\theta_l=x-\beta$)

$$\begin{aligned}\langle\theta_{\text{est}}\rangle&=\langle\theta_{\text{est}}+\beta\rangle-\beta\\&\equiv\int(\theta_{\text{est}}+\beta)R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}-\beta\\&=\int(\theta_{\text{est}}+\beta)\left\{\frac{(\theta_0+\beta)^{\theta_{\text{est}}+\beta}}{(\theta_{\text{est}}+\beta)!}\exp\{-(\theta_0+\beta)\}\right\}\\&\qquad\times d\theta_{\text{est}}-\beta=\theta_0,^{10}\end{aligned} \tag{89}$$

and similarly

$$\langle\theta_{\text{est}}{}^2\rangle\equiv\int\theta_{\text{est}}{}^2R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}=\theta_0{}^2+(\theta_0+\beta), \tag{91}$$

so that

$$\langle\theta_{\text{est}}{}^2\rangle-\langle\theta_{\text{est}}\rangle^2=\theta_0+\beta=\langle\theta_{\text{est}}\rangle+\beta. \tag{92}$$

Equation (92) can be used as a check on the internal consistency of the coincidence counting data. Specifically, Eq. (92) gives (since $\theta_{\text{est}}=\theta_l=x-\beta$)

$$M^{-1}\sum_{\gamma=1}^{M}(x^{(\gamma)}-\beta)^2-\{M^{-1}\sum_{\gamma=1}^{M}(x^{(\gamma)}-\beta)\}^2=M^{-1}\sum_{\gamma=1}^{M}x^{(\gamma)}, \tag{93}$$

where $x^{(\gamma)}$ is the number of coincidences during a definite time interval γ (say, one day).

VII. CONCLUSION

In this paper we have attempted to find a relationship between the probability functions $H(x_i;\theta)$ and $G(\theta;x_i)dx_i$ defined in Sec. I. In other words, we have attempted to answer the question, "How does one proceed to estimate a physical parameter θ, given a set x_i of physical data related to θ through a theoretically derived and/or experimentally confirmed probability function $G(\theta;\,x_i)dx_i$; in addition, what statistical significance does this estimate and any quoted 'error' possess?"

The above problem has, of course, been treated in the literature in great detail. In fact, R. A. Fisher[3] has actually constructed a "maximum likelihood" theory of estimation essentially independent of the prior probability $P(\theta)$. On the other hand, in the present paper we attempt to formulate a treatment of estimation consistent with a partial knowledge of $P(\theta)$. Specifically:

1. If $P(\theta)d\theta$ is known initially, the procedure to follow is given in Sec. II. If a single measurement of the set x_i is made, the corresponding value of θ can be estimated and meaningful probability statements can be made concerning this estimate, θ_{est}; e.g., we can give the probability that θ lies between $\theta_{\text{est}}+\sigma(\theta_{\text{est}})$ and $\theta_{\text{est}}-\sigma(\theta_{\text{est}})$.
2. If $P(\theta)d\theta$ is initially completely unknown, one can always find it on the basis of sufficient experimental data using the procedure in Sec. III.
3. If $P(\theta)d\theta$ is initially completely unknown and *only one measurement of the set x_i is made*, there is an intrinsic degree of arbitrariness in any estimation procedure. Nevertheless we suggest in Sec. IV that in this case one take all values of θ as having the same prior probability and that one quote an estimate $\theta_{\text{m.p.}}$ formally equivalent to the "maximum likelihood" estimate. The corresponding "statistical error" in $\theta_{\text{m.p.}}$ as well as $\theta_{\text{m.p.}}$ itself, however, are now quantities about which one can make no statements which are meaningful in the sense of probability = relative frequency.
4. If $P(\theta)d\theta$ is initially known to be a single δ function about an unknown value of θ [$P(\theta)=\delta(\theta-\theta_0)$ where θ_0 is unknown], and if one measurement of the set x_i is made, the procedure is similar to that in (3) above, but we believe that its justification rests on firmer ground (Sec. IV). One can here quote an estimate of θ, the "maximum likelihood" estimate, and a "figure of merit" (or "statistical error") for this estimate (see also Sec. V). However, it is still intrinsic to the problem that no really meaningful probability statement (in the sense probability = relative frequency) can be made about the "statistical error" associated with the estimate. On the other hand, this "statistical error," and indeed the estimate itself, possesses a significance in the sense of a suitable statement of the relative degree of belief which can be objectively assigned to the possible θ values; we feel that such a significance is greater when $P(\theta)=\delta(\theta-\theta_0)$ with θ_0 unknown, than in the corresponding case just above where nothing is known about $P(\theta)$ (see also Sec. VI).

One of the authors (M.A.) wishes to thank Professor Bruno Rossi for very helpful discussions. We also wish to thank Professor Robert Sard for his helpful comments.

[10] The estimate distribution $R(\theta_0;\theta_{\text{est}})d\theta_{\text{est}}$ permits the definition of the so-called "unbiased" estimate through the relation

$$\begin{aligned}\theta_0=\langle\theta_{\text{est}}\rangle&\equiv\int d\theta_{\text{est}}\theta_{\text{est}}R(\theta_0;\theta_{\text{est}})\\&=\int f_{\text{est}}(x_1\cdots x_N)G(\theta_0;x_1\cdots x_N)dx_1\cdots dx_N,\end{aligned} \tag{90}$$

any θ_{est} satisfying the condition in Eq. (90) being called "unbiased." Thus any θ_{est} is unbiased provided that its average over the estimate distribution is equal to the true (or assumed) value θ_0 of the parameter θ. Eq. (89) shows that the $\theta_{\text{est}}=\theta_l$ of the present example ($N=1!!$) is unbiased. (It is to be noted that Eq. (81c) proves the "unbiasedness" in general of maximum likelihood estimates only in the limit $N\rightarrow\infty$.)