CTSL Internal Report #57
CALT-68-321

PROPERTIES OF THE MAXIMUM LIKELIHOOD METHOD*

Kirk McDonald

California Institute of Technology, Pasadena, California 91109

August 5, 1971

*Work supported in part by the U.S. Atomic Energy Commission.

In this paper I review some properties of the maximum likelihood method of fitting data. The sections on the bias of the method, and on tests for goodness of fit may be of particular interest, as this material is not generally known to physicists.

1. Introduction to the Maximum Likelihood Method.

Often in experiments we make observations of a quantity (or quantities), here labelled x, which are related to some parameter(s), a, by a probability distribution. That is, the probability of observing an event with x in the range dx is given by

$$P(a,x)dx \ .$$

We shall always consider P to be normalized

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

for all a. The problem is to estimate the parameter(s) a, from the observations of x.

One might think to use some sort of least squares fitting procedure. This requires casting the data into a function of a, which can then be fitted in the usual manner. The simplest way to do this is to divide the range of x into a definite set of bins Δx_i centered at x_i and setting

$$N_i = \text{number of events in } \Delta x_i / \text{total number of events.}$$

The N_i then form a set of data which may be compared with the function

$$F(a,i) = P(a,x_i)\Delta x_i$$

to predict a. The problem with this method is that the choice of the bin size is arbitrary. Further, one normally associates an error with each point in a least squares fit. What should we assign as the error on the number of events in a bin?

A method which avoids the question of binning is the method of moments. From the expression $P(a,x)$ we may calculate the various moments:

$$F_n(a) = \int x^n P(a,x)dx \quad .$$

If an experiment consists of N observations of x yielding x_i , i = 1...N, then we may estimate from the data:

$$F_n = \sum_{i=1}^{N} x_i^n/N \quad .$$

Of course, only the first N moments may be estimated in an experiment with N events. Equating the experimental and 'theoretical' moments we now have N equations for a. The problem is to decide which one gives the best estimate of a.

To overcome the difficulties seen above, R.A. Fisher proposed the Maximum Likelihood Method in 1912[1]. From the distribution $P(a,x)$ we form the quantity

$$L(a) = \prod_{i=1}^{N} P(a,x_i)$$

where the x_i are events observed in an experiment. Then L(a) is the probability (density) that an experiment produces the events x_i, assuming a to be the correct value for the parameter. L(a) is the likelihood function. The likelihood method consists of choosing as the best estimate of a that value which maximizes L(a).

The likelihood function, L(a), is a peculiar function. It is not a probability distribution--L(a)da is not the probability that a is the true parameter. To see this, consider another parametrization, say b, where $b = a^2$. If L were a probability distribution, then

$$L(b)db = L(b) \cdot 2ada \quad \text{or} \quad L(a) = 2aL(b) \ .$$

However, the definition of L implies $L(a) = L(b)$ for $b = a^2$.

It is also difficult to interpret the height of L. The more events in our experiment, the greater L is, but there is no relation between the height of L and, say, the goodness of our choice of parametrization as there is with χ^2. People say that if $L(a_1) = 2L(a_2)$ then a_1 is twice as likely as a_2, where 'likely' seems to be defined by this very statement.

In practice it is often more convenient to deal with $\ell n\ L(a)$ rather than L(a). These both attain their maxima at the same a, so either may be used to predict the value of a corresponding to maximum likelihood. However,

$$\ell n\ L(a) = \sum_i \ell n\ P(a,x_i)$$

which is much easier to deal with than the large product which makes up L.

We can give a further demonstration that ℓnL is more useful than L. Suppose a_o is the true value of the parameter. Then $L(a_o) = \prod_i P(a_o,x_i)$ is the exact probability that events x_i are observed. Hence we may calculate the expectation value of any quantity $g(x_1 \ldots x_N)$ for an experiment of N events:

$$\langle g \rangle = \int \ldots \int g(x_1 \ldots x_N)\ L(a_o) \prod_i dx_i \ .$$

Consider $g = L(a)$

$$\langle L(a)\rangle = \int \cdots \int \prod_i P(a,x_i)\, P(a_o,x_i)\, dx_i = \left[\int P(a,x)\, P(a_o,x)\, dx\right]^N \quad .$$

We might expect $\langle L(a)\rangle$ to have a maximum at a_o. However,

$$d\langle L(a)\rangle/da\Big|_{a_o} \sim \frac{d}{da} \int P(a,x)\, P(a_o,x)dx\Big|_{a_o}$$

which is non-zero in general.

This might lead one to doubt that the maximum likelihood method works at all. If we call a^* that value of a which maximizes L(a) for a particular experiment, we should say the method works if

$$\langle a^* \rangle = a_o \quad .$$

This is indeed true for large N but the proof is not simple. I refer the reader to Cramér[2].

We can show that $\langle \ell n\, L(a)\rangle$ has a maximum at a_o.

$$\langle \ell n\, L(a)\rangle = \int \cdots \int \sum_i \ell n\, P(a,x_i) \prod_i P(a_o,x_i)dx_i$$

$$= N \int P(a_o,x)\, \ell n\, P(a,x)dx$$

$$\frac{d}{da} \langle \ell n\, L(a)\rangle = N \int \frac{P(a_o,x)}{P(a,x)} \frac{dP(a,x)}{da}\, dx \quad .$$

At $a = a_o$ this is

$$N \frac{d}{da} \int P(a,x)dx\Big|_{a_o} = 0$$

since $\int P(a,x)dx = 1$ for all a. Thus to get a picture of L(a) it is better to consider $e^{\langle \ell n\, L(a)\rangle}$ than $\langle L(a)\rangle$.

We close this section with some examples.

Consider the distribution

$$P(a,x) = (1 + a \cos x)/2\pi \quad .$$

This distribution arises in the azimuthal dependence of the scattering of a polarized spin 1/2 particle off a spin 0 target. An experiment consists of the observation of a set of scatters with angles x_i, $i = 1,..N$.

Then $$\ell n\, L(a) = \sum_i \ell n(1 + a \cos x_i) - N\, \ell n(2\pi)$$

$$\frac{d\, \ell n\, L(a)}{da} = \sum_i \frac{\cos x_i}{1 + a \cos x_i}$$

The value of a which maximizes this can be found by Newton's method or some other numerical technique. For an instructive comparison of this result with that of the method of moments (which provides a good first guess for Newton's method) see the thesis of Bruce Winstein[3].

Suppose a_o is the true value of a. We can calculate, for an experiment of N events,

$$\langle L(a)\rangle = \left[\frac{1}{2\pi} \int_o^{2\pi} (1 + a \cos x)(1 + a_o \cos x)dx\right]^N$$

$$= \left(1 + \frac{a\, a_o}{2}\right)^N$$

$$\langle \ell n\, L(a)\rangle = \frac{N}{2\pi} \int_o^{2\pi} (1 + a_o \cos x)\ell n\, (1 + a \cos x)dx$$

$$= N\left[\ell n\left(\frac{1 + \sqrt{1 - a^2}}{2}\right) + a_o \left(\frac{1 - \sqrt{1 - a^2}}{a}\right)\right]$$

Figure 1 shows $\exp(\langle \ell n\, L(a)\rangle)$ for $N = 50$ and $a_o = 0$ and 0.8. The functions have been normalized to 1.0 for comparison.

It is perhaps useful to have an example involving 2 parameters. Consider

$$P(x) = (1 + a\cos x + b\sin x)/2\pi \quad .$$

For an experiment

$$\ell n\, L(a,b) = \sum_{i=1}^{N} \ell n(1 + a\cos x_i + b\sin x_i) - N\,\ell n(2\pi) \quad ,$$

leading to the maximum likelihood conditions:

$$\sum_i \frac{\cos x_i}{1 + a\cos x_i + b\sin x_i} = 0$$

$$\sum_i \frac{\sin x_i}{1 + a\cos x_i + b\sin x_i} = 0 \quad .$$

To get a feel for the shape of $L(a,b)$ we again consider $\langle \ell n\, L(a,b)\rangle$. It is convenient to rewrite

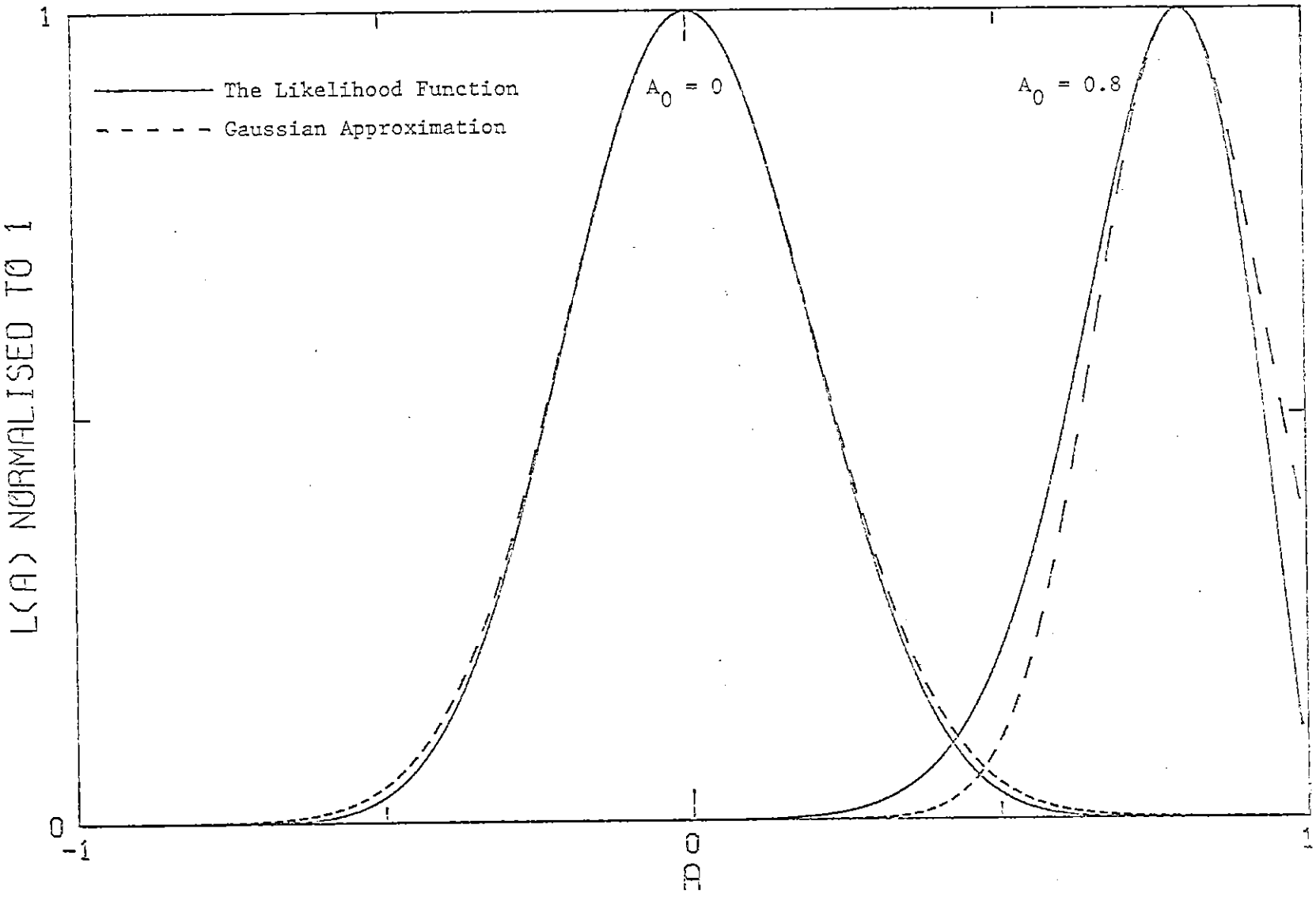
$$P(x) = 1 + A_o\cos(x - \Xi_o)$$

where $a_o = A_o\cos\Xi_o$ $b_o = A_o\sin\Xi_o$ are the true values of the parameters. Now

$$\langle \ell n\, L(A,\Xi)\rangle = \frac{N}{2\pi}\int (1 + A_o\cos(x - \Xi_o))\,\ell n(1 + A\cos(x - \Xi))dx$$

$$= N\left[\ell n\left(\frac{1 + \sqrt{1 - A^2}}{2}\right) + A_o\cos(\Xi - \Xi_o)\left(\frac{1 - \sqrt{1 - A^2}}{A}\right)\right] \quad .$$

This function is illustrated in Figure 2 for $A_o = 0.8$, $\Xi_o = 0$ and $N = 200$. The parameters are called $P_{\parallel}$ and $P_{\perp}$ corresponding to a and b; also P_o is the same as A_o.



THE LIKELIHOOD FUNCTION FOR P(X) = 1 + AxCOS(X)

Figure 1

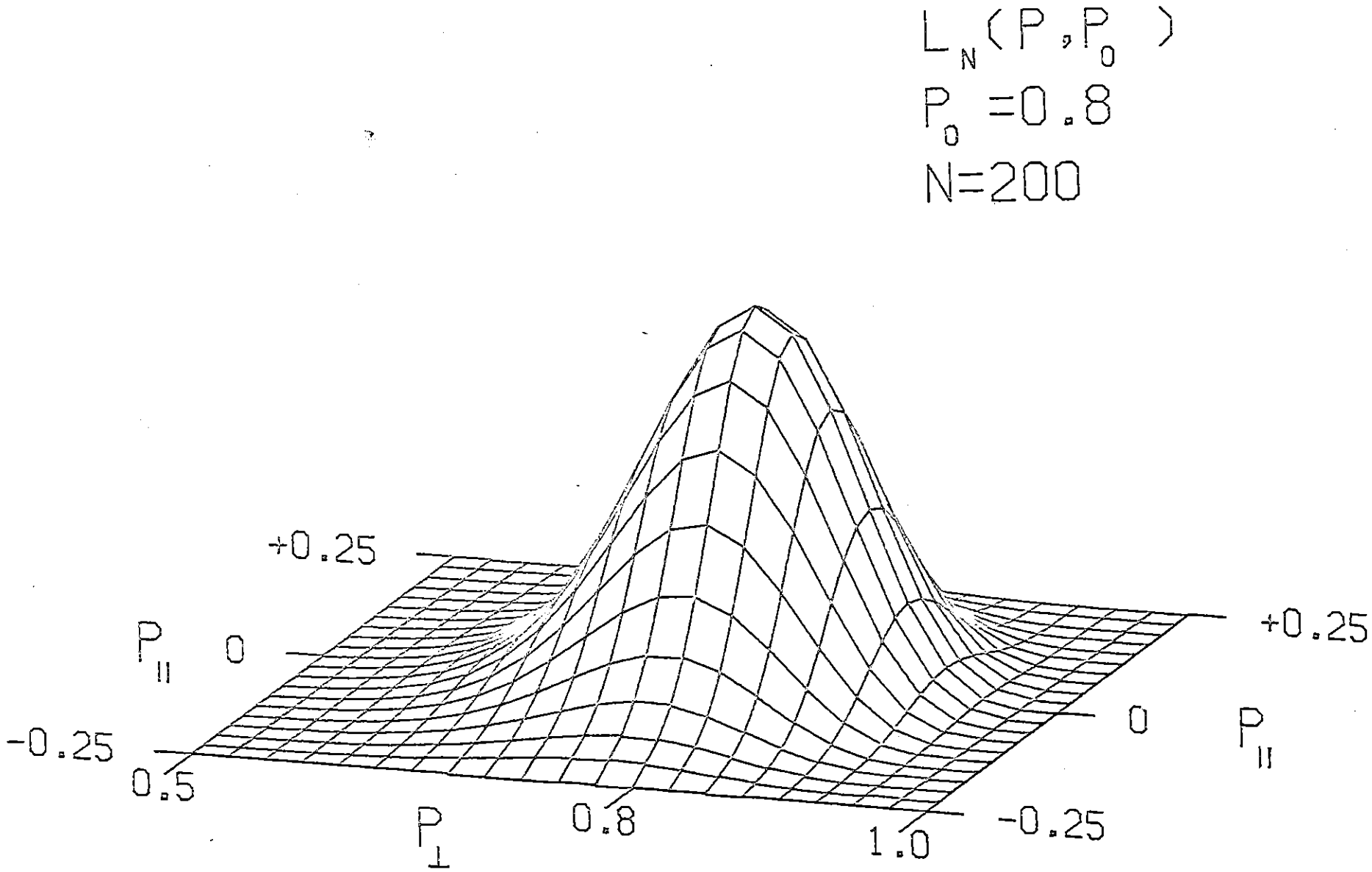

$L_N(P, P_0)$
$P_0 = 0.8$
N=200
+0.25
$P_{\parallel}$ 0
-0.25
0.5
$P_{\perp}$
0.8
1.0
-0.25
0
+0.25
$P_{\parallel}$

Figure 2

2. The Likelihood Function for Observations in a Limited Range.

Suppose we conduct an experiment in which x is observed over only a limited range, say, x_1 to x_2. The correct application of the maximum likelihood method requires our 'theoretical' probability distribution to be normalized for the interval (x_1, x_2). Thus

$$P(a,x)_{NEW} = P(a,x) / \int_{x_1}^{x_2} P(a,x)dx \quad .$$

We illustrate this using the example of Section 1,

$$P(a,x) = (1 + a \cos x)/2\pi \quad ,$$

for observations restricted to $(\mathbb{X}_1, \mathbb{X}_2)$. The new probability distribution is:

$$P(a,x) = (1 + a \cos x)/(\mathbb{X}_2 - \mathbb{X}_1 + a(\sin\mathbb{X}_1 - \sin\mathbb{X}_2) \quad .$$

Note that in practice $\mathbb{X}_1$ and $\mathbb{X}_2$ could be different for each event. For an experiment of N observations x_i, the maximum likelihood condition is

$$\sum_{i=1}^{N} \frac{\cos x_i}{1 + a \cos x_i} = \frac{N(\sin\mathbb{X}_1 - \sin\mathbb{X}_2)}{\mathbb{X}_2 - \mathbb{X}_1 + a(\sin\mathbb{X}_1 - \sin\mathbb{X}_2)} \quad .$$

3. The Effect of Uncertainty in the Data.

Suppose each observation of x_i in an experiment is associated with an error σ_i, presumed to be random. Then the probability of y_i being the value which should have been observed in the case of infinite precision is

$$\frac{1}{\sqrt{2\pi}\,\sigma_i} e^{-(y_i - x_i)^2/2\sigma_i^2} .$$

The likelihood function would then be

$$L(a) = \prod_i P(a,x_i,\sigma_i)$$

where

$$P(a,x,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \int P(a,y) e^{-(y-x)^2/2\sigma^2}\, dy .$$

Consider our example, $P(a,x) = (1 + a \cos x)/2\pi$. Calculation yields an effective distribution

$$P(a,x,\sigma) = (1 + a \cos x\, e^{-\sigma^2/2})/2\pi$$

assuming $\sigma << 2\pi$ so that the limits of integration can be extended to infinity. We can use this result to estimate a correction to the maximum likelihood value of a in the case where the error in N is ignored. We have

$$a_{\text{CORRECTED}} = a\, e^{\sigma^2/2}$$

where σ is the average error. Thus an average error of 10° would cause only a 1% correction.

Note that uncertainty in the data can affect the value of the maximum likelihood estimate of a, as well as the uncertainty in a.

4. The Error of the Maximum Likelihood Method.

Suppose a true value, a_o, of the parameters exists, and the likelihood method predicts a* from an experiment. What is the error associated with a*? What is the value for $\langle (a^* - a_o)^2 \rangle$? We give a rather intuitive argument which gives an answer (of order 1/N, N = number of events) which is, however, correct with the neglect of terms of order $1/N^2$.

We consider the second moment of a in $\exp(\langle \ell n\, L(a) \rangle)$. Expanding this about a_o:

$$\langle \ell n\, L(a) \rangle = \langle \ell n\, L(a_o) \rangle + (a - a_o) \frac{d}{da} \langle \ell n\, L(a_o) \rangle + 1/2\, (a - a_o)^2 \frac{d^2}{da^2} \langle \ell n\, L(a_o) \rangle + \ldots$$

The first derivative vanishes at a_o as shown in Section 1. Thus, with the neglect of higher order terms, $\exp(\langle \ell n\, L(a) \rangle)$ is Gaussian. Our intuitive assumption is that the width of this Gaussian is a good estimate of $\langle (a^* - a_o)^2 \rangle$:

$$\langle (a^* - a_o)^2 \rangle = -1 / \frac{d^2}{da^2} \langle \ell n\, L(a_o) \rangle \quad .$$

For a single experiment, we estimate a_o by a^*, and the error on our estimate as

$$\sigma_a^2 = -1 / \frac{d^2}{da^2} \ell n\, L(a^*) \quad .$$

Cramér[2] gives a proof that the above estimate of the error is in fact a lower bound on $\langle (a^* - a_o)^2 \rangle$ attained only in certain cases for finite experiments. It is correct in the limit of large N for most practical cases.

Figure 1 shows $\langle \ell n\, L(a) \rangle$ for the distribution $(1 + a \cos x)/2\pi$. The

dashed curves are Gaussians of widths given by the above expression.

We now consider the extension to the case of several parameters, a_i, $i = 1...M$. We now are interested in all of the moments

$$\langle (a_i^* - a_{oi})(a_j^* - a_{oj}) \rangle \quad .$$

Following the argument for 1 parameter, we expand (using the notation $\vec{a}$ for a_i, $i = 1..,M$),

$$\langle \ell n\ L(\vec{a}) \rangle = \langle \ell n\ L(\vec{a}_o) \rangle - 1/2 \sum_{ij} (a_i - a_{oi})(a_j - a_{oj})\ A_{ij} + \ldots$$

where

$$A_{ij} = - \frac{\partial^2}{\partial a_i\ \partial a_j} \langle \ell n\ L(\vec{a}_o) \rangle \quad .$$

If A_{ij} is diagonal then clearly we should take

$$\langle (a_i^* - a_{oi})^2 \rangle = (A_{ii})^{-1} = (A^{-1})_{ii} \quad .$$

For A_{ij} non-diagonal, the reader will appreciate that the second equality is the correct generalization:

$$\langle (a_i^* - a_{oi})(a_j^* - a_{oj}) \rangle = (A^{-1})_{ij} \quad .$$

Again in practice, we estimate

$$A_{ij} = - \frac{\partial^2}{\partial a_i\ \partial a_j} \ell n\ L(\vec{a}^*) \quad .$$

To get a feel for the magnitude of the errors from the basic distribution

$P(\vec{a},x)$ recall

$$\langle \ell n\ L(\vec{a})\rangle = N \int P(\vec{a}_o,x)\ \ell n\ P(\vec{a},x)\ dx \quad .$$

Then

$$\langle A_{ij}\rangle = -N \int P(\vec{a}_o,x)\ \frac{\partial^2}{\partial a_i\ \partial a_j}\ \ell n\ P(\vec{a}_o,x)\ dx \quad ,$$

or

$$\langle A_{ij}\rangle = N \int \frac{1}{P(\vec{a}_o,x)}\ \frac{\partial P(\vec{a}_o,x)}{\partial a_i}\ \frac{\partial P(\vec{a}_o,x)}{\partial a_j}\ dx \quad ,$$

where we have used the fact $\int P(\vec{a},x)\ dx = 1$.

We now illustrate the estimation of errors using the examples of Section 1. First we consider the distribution $(1 + a\ \cos x)/2\pi$. For an experiment

$$\sigma^2 = 1/\sum_i \frac{\cos^2 x_i}{(1 + a^*\ \cos x_i)^2} \quad ,$$

the expectation for σ is given by

$$\sigma^2 = \frac{2\pi}{N} \Big/ \int_0^{2\pi} \frac{\cos^2 x\ dx}{1 + a_o\ \cos x} = \frac{1 - a_o^2 + \sqrt{1 - a_o^2}}{N} \quad .$$

For $a_o = 0$, $\sigma = \sqrt{2/N}$, while for $a_o = 1$, $\sigma = 0$! The latter limit is a result of the fact that the likelihood function varies rapidly with a_o for a_o near 1.

The second example is the distribution

$$(1 + a\ \cos x + b\ \sin x)/2\pi = (1 + A\ \cos(x - \overline{\Delta}))/2\pi \quad .$$

When working with experimental data it is probably easier to use the parametrization with a and b. However, the "error matrix", A_{ij}, is not diagonal, even in the limit of large N. The parametrization A and X is unwieldy to handle experimentally, but yields simpler understanding of the errors as A_{ij} is diagonal in this case. Evaluating the integrals for true parameters A_o and $\mathcal{X}_o$,

$$\sigma_A = \sqrt{\frac{1 - A_o^2 + \sqrt{1 - A_o^2}}{N}}$$

$$A_o \sigma_{\mathcal{X}} = \sqrt{\frac{1 + \sqrt{1 - A_o^2}}{N}} \quad .$$

The errors exhibit rotational invariance. Note also that in the limit of $A_o = 1$, $\sigma_A = 0$ but $\sigma_{\mathcal{X}_o} = \sqrt{1/N}$.

5. The Bias of the Likelihood Method

So far we have been tacitly assuming that if a true parameter, a_o, exists, then the likelihood estimate a^*, obeys $\langle a^* \rangle = a_o$. This is true only asymptotically[2]. Steve Yellin has derived an expression for the bias for the case of a finite number of events[4]. Writing

$$\langle a_i^* \rangle = a_{oi} + f_i(\vec{a}_o) \quad ,$$

(for the case of several parameters), he finds

$$f_i(a_o) = -\frac{1}{2N} \int \frac{dx}{P(\vec{a}_o,x)} \sum_j (A^{-1})_{ij} \frac{\partial P(\vec{a}_o,x)}{\partial a_j} \sum_{mn} (A^{-1})_{mn} \frac{\partial^2 P(\vec{a}_o,x)}{\partial a_m \partial a_n} \quad ,$$

plus terms of order $1/N^2$. A is now given by

$$A_{ij} = \int \frac{dx}{P(\vec{a}_o,x)} \frac{\partial P(\vec{a}_o,x)}{\partial a_i} \frac{\partial P(\vec{a}_o,x)}{\partial a_j} \quad .$$

Note that if P is linear in its parameters, the bias vanishes (neglecting order $1/N^2$).

As an example of a bias occurring in the likelihood method, consider a Gaussian distribution

$$P_x = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu^2)/2\sigma^2} \quad .$$

The likelihood estimates of μ and σ are:

$$\mu^* = \sum_i x_i/N$$

$$\sigma^* = \sqrt{\sum_i (x_i - \mu^*)^2/N} \quad .$$

However, it is well known that the best estimate of σ is $\sqrt{\sum_i (x_i - \mu^*)^2/(N-1)}$.

6. Goodness of Fit

The question arises whether the parametrization we used to fit the data was a good choice. As we have remarked earlier, the size of the likelihood function itself does not give any precise clue as to the goodness of the fit. If we have two forms of parametrization we can say that the one which has the larger likelihood function at maximum is better, but how much better is a little vague.

As an example, suppose an experiment is performed sampling a distribution of the form $1 + a_o \sin x$ but we mistakenly try to fit the results to $1 + a \cos x$. Then the likelihood function we construct is

$$\langle \ell n\, L(a) \rangle = N/2\pi \int_0^{2\pi} (1 + a_o \sin x)\, \ell n(1 + a \cos x)\, dx$$

$$= N\, \ell n\left(\frac{1 + \sqrt{1 - a^2}}{2}\right) \quad ,$$

which is independent of a_o! The likelihood function reaches its maximum at $a = 0$. If $a_o = 0$ our fit is good, but if $a_o = 1$ it is very bad. However, our likelihood function is the same in both cases. Thus it can yield no information as to the goodness of the fit.

The classic test for goodness of fit is, of course, Pearson's χ^2 test. A good review of its properties is given by Cochran[5]. To apply it to the fit for a probability distribution one must divide the range of x into bins of equal probability. The number of bins should be such that there are at least 5 to 10 events observed in each bin. The advantage of the χ^2 test is that (at least for large samples) it is independent of the form of the parametrization (non-parametric, to the mathematicians) and also it is not

sensitive to the fact that the values of the parameters have been determined by the data itself. Its disadvantage for the present case is that the data must be binned rather arbitrarily. In addition, the number of bins will usually be small and the properties of the χ^2 test for small samples are not as well established as its common usage might indicate.

What is desired is a test of goodness of fit more suited to the nature of a probability distribution. Such tests, in their ideal form, are known to mathematicians as non-parametric statistics, of which the χ^2 test is only one example. These tests seem to be more familiar to psychologists than to physicists, and there is even a textbook on this subject for behavioral scientists[6].

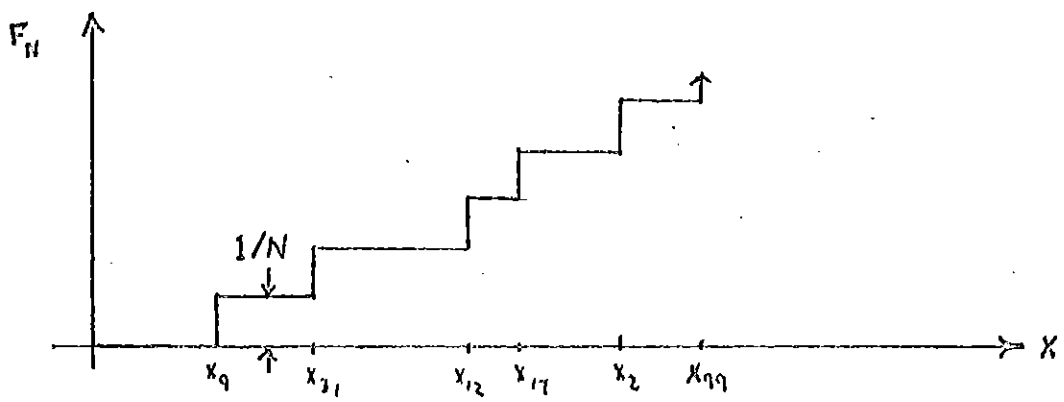
The basis for the tests is something called the "empirical distribution function". For an experiment with events x_i, $i = 1...N$, it is defined by

$$F_N(x) = 1/N \sum_i \varepsilon(x - x_i)$$

where

$$\varepsilon(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases}$$

It looks like



F_N is a step function with a step of height $1/N$ occurring at each data point. F_N is limited to the interval $[0,1]$.

If $P(a,x)$ is the probability distribution for x, we may construct the ideal behavior of F_N as

$$F(x,a) = \int^x P(a,y)\, dy \quad .$$

By comparing F and F_N we get analogues of the χ^2 test. This is commonly done in two ways.

Test 1. (Cramér-von Mises-Smirnov Test)

Define

$$C_N^2 = N \int (F_N(x) - F(x,a))^2 dF(x,a) \quad ,$$

which for well-behaved cases is

$$C_N^2 = N \int (F_N(x) - F(x,a))^2 P(a,x)\, dx \quad .$$

Test 2. (Kolmogorov-Smirnov Test)

Define

$$K_N = \sqrt{N}\, \mathrm{MAX} |F_N(x) - F(x,a)| \quad .$$

Since F_N and $F(x,a)$ are monotonic, the maximum can only occur at a data point. Thus K_N is particularly easy to calculate.

If the parameter, a, is known ahead of time, both of these tests give confidence levels which are independent of the form of $P(a,x)$, even for small N. The confidence level function is, however, a function of N. As the χ^2 test is not truly independent of $P(a,x)$ for small N, these new tests are

superior in principle to the χ^2 test. Note that Tests 1 and 2 could also be applied to the results of a least squares fit. A review of these Tests is given by Birnbaum[7] and in more detail, with many references, by Darling[8].

A confidence function, C, for the result, R, of some test of goodness of fit is defined by

$$C(r) = \text{probability that } R < r \quad .$$

For Test 1, the confidence function, C(r), is tabulated by Anderson and Darling[9] for the limit of large N. For Test 2, also in the limit of large N,

$$C(r) = \sum_{n=-\infty}^{\infty} (-1)^n \, e^{-2n^2r^2} \quad .$$

This result is also given in a table by Smirnov[10]. For finite N the confidence function is tabulated in References 11, 12, and 13.

The difficulty with these tests is that if the parameter a is inferred from the data itself, then the distribution free confidence functions no longer apply. The general result of fitting the parameters with the data is to lower the result of a goodness of fit test, be it χ^2, C_N^2, or K_N. One commonly compensates for this effect for the χ^2 test by using a confidence function for N equal to the number of degrees of freedom rather than the number of data points. This procedure is also not distribution free for small samples. However, for the C_N^2 and K_N tests there is no known simple procedure to correct for this reduction in the degrees of freedom. Darling[14] gives a complicated procedure for the C_N^2 test which could be carried out for any particular P(a,x) on a computer.

We can give a practical, though somewhat tedious, method for producing

confidence levels for any of these tests when the parameters are determined from a finite number of events. It is a Monte Carlo calculation. Choose a probability distribution and parameters of interest, and produce a large number of "experiments" on a computer with a random number generator. They all should have the same number of events. For each experiment use the maximum likelihood method to estimate the value of the parameter (which you fixed for purposes of calculation). Using this estimate, calculate the value of χ^2, C_N^2, or K_N. In this way you accumulate statistics on the frequency of appearance of various values of χ^2, etc. This is your confidence level table. If you use the fixed value of the parameter to calculate χ^2, etc., you should generate the distribution free confidence lebels mentioned above. This would provide a check on the calculation.

Experience with the distribution $1 + a \cos x$ (Bruce Winstein, private communication) shows that the confidence function for C_N^2 remains closer to the distribution free function than does that for K_N. when the data is used to fit the parameters.

* * *

I would like to thank Bruce Winstein and Steve Yellin for the many conversations during which this review was developed.

References

1. R.A. Fisher, Mess. of Math. 41, 155 (1912); see also R.A. Fisher, Contributions to Mathematical Statistics, Wiley, New York (1950).
2. Harald Cramér, Mathematical Methods of Statistics, Princeton U.P. (1945) p. 497 ff.
3. Bruce Winstein, thesis, Caltech (1970), p. 81 ff.
4. Steve Yellin, thesis, Caltech (1971), p. 134 ff.
5. W.G. Cochran, Ann. Math. Stat. 23, 315 (1952).
6. Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill, New York (1956).
7. Z.W. Birnbaum, Ann. Math. Stat. 24, 1 (1953).
8. D.A. Darling, Ann. Math. Stat. 28, 823 (1957).
9. T.W. Anderson and D.A. Darling, Ann. Math. Stat. 23, 193 (1952).
10. N.V. Smirnov, Ann. Math. Stat. 19, 279 (1948).
11. F.J. Massey, Jr., J. Amer. Stat. Assn. 46, 68 (1951).
12. Z.W. Birnbaum, J. Amer. Stat. Assn. 47, 425 (1952).
13. L.H. Miller, J. Amer. Stat. Assn. 51, 111 (1956).
14. D.A. Darling, Ann. Math. Stat. 26, 1 (1955).