J. Phys. A: Math. Gen. 25 (1992) 3813–3826. Printed in the UK

How large a sample is needed for the maximum likelihood estimator to be approximately Gaussian?

Samuel L Braunstein
Department of Physics, Technion–Israel Institute of Technology, 32000 Haifa, Israel

Received 10 January 1992

Abstract. This paper concerns the failure of the Gaussian approximation to the distribution of the maximum-likelihood estimator in one-parameter families for finite sample sizes. Fisher has shown that this approximation is valid when an asymptotically large sample of data points is used. He did this by treating the likelihood equation (i.e. the equation obtained by setting the derivative of the likelihood function with respect to the parameter to zero) statistically and finding its solution as the sample size n is taken to infinity. In this paper the statistical treatment of the likelihood equation is extended to include corrections for *finite* sample sizes. The $O(1/n)$ corrections to Fisher's asymptotic Gaussian result are calculated *with* corrections to the central limit theorem, and are used to derive *sufficient* conditions on the sample size for Fisher's result to break down. Such conditions are useful for the design of experiments. The procedure developed here can be extended to the maximum-likelihood estimation of several parameters in multivariate distributions.

1. Introduction

The maximum-likelihood estimator is a widely used method of parameter estimation. The asymptotic behaviour of this estimator is known; however, it is not generally known when this asymptotic regime is reached. There are many applications where a set of 100 data points is considered a 'large' sample, and yet figure 1 shows that, at least for some distributions, 100 data points are not nearly enough. In this paper two sufficient conditions on the number of data points are derived which ensure the failure of this asymptotic behaviour. This is done by deriving an expansion to $O(1/n)$ for the probability $P(\Delta\Phi)$ of the error $\Delta\Phi$ of this estimator for the parameter Φ. The $O(1/n)$ terms correct for the finite size n of the sample.

For an asymptotically large number of sample points, the variance of the distribution of $\Delta\Phi$ was shown by Fisher [4, 7] to go as $\mathrm{var}(\Delta\Phi) \sim 1/(n\mathcal{I})$, and this behaviour is associated with the distribution $P(\Delta\Phi)$ approaching a Gaussian. (Here $\mathcal{I}$ is called the 'Fisher information', or sometimes the 'expected Fisher information'.) A Gaussian likelihood function leads to great simplifications in data analysis; straightforward techniques can be used to find the maximum [8] and hence estimate the parameter. Also, calculation of the confidence intervals may be performed using the local curvature of the likelihood function at the maximum [2]. For these reasons it is important to know when the Gaussian approximation is invalid.

Further, knowing when the asymptotic regime begins means that one also knows the efficiency of maximum-likelihood estimation from that point on. Thus, the re-

0305-4470/92/133813+14$04.50

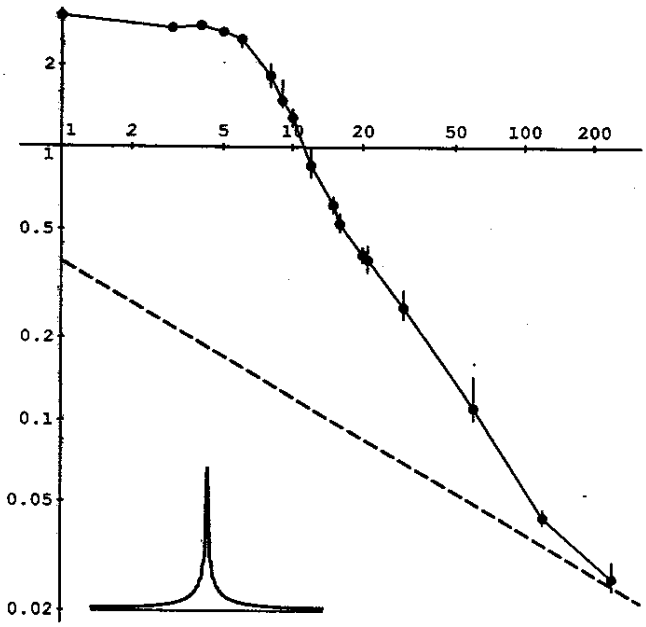

Figure 1. This figure shows the half-width 99% confidence interval for estimating the location of the peak of a distribution versus sample size. Even for 100 data points the Gaussian approximation given by the broken curve is shockingly bad. (The sampled distribution shown in the inset is taken from expression (3.5) with $M = 74$, and the error bars represent a 95% confidence.

sults presented here also play an important role in the design of experiments having 'optimal' efficiency [1, 9].

For an unbiased maximum-likelihood estimation of a single parameter Φ from n data points $\{\phi_i : i = 1, \ldots, n\}$ taken independently from a distribution of the form $p(\phi \mid \Phi) = f(\phi - \Phi)$ (i.e. the translation family of single-parameter univariate distributions), two sufficient conditions are derived below for the failure of Fisher's asymptotic result. The first condition is that if

$$\text{number of data points} \equiv n \leqslant \left[\frac{2}{\mathcal{I}^2}\int \mathrm{d}\phi\left(\frac{p''^2}{p} - \frac{1}{3}\frac{p'^4}{p^3}\right) - 2\right]\left(\frac{100\%}{e\%}\right)$$

then the $O(1/n^2)$ correction to the variance $\mathrm{var}(\Delta\Phi)$ differs from Fisher's asymptotic result of $1/(n\mathcal{I})$ by more than $e\%$. Here $p = p(\phi \mid \Phi)$ and primes on p denote differentiation with respect to ϕ. Moreover, since Fisher's result corresponds to the Cramér–Rao lower bound it is an *underestimate* of $\mathrm{var}(\Delta\Phi)$. The second condition is that if

$$n \leqslant \left|\frac{1}{\mathcal{I}^2}\int \mathrm{d}\phi\left(\frac{4p''^2}{p} - \frac{5}{3}\frac{p'^4}{p^3}\right) - 3\right|\left(\frac{100\%}{f\%}\right)$$

then the lowest-order corrections to the kurtosis of $P(\Delta\Phi)$ will be more than $f\%$ of the square of its variance. For a Gaussian distribution the kurtosis would be identically zero.

It is shown in section 2 how the asymptotic expansion of Fisher can be extended so as to include corrections due to the breakdown of the central limit theorem. This

is used in section 3 to obtain the above sufficient conditions for the case of the translation family of distributions. Then in section 4 the validity of these conditions is demonstrated for two particular distributions by comparison with the independent method of Monte Carlo simulations. The appendix gives a summary of the results needed in section 2 for asymptotic corrections to the central limit theorem.

2. Statistical treatment of the likelihood equation

For n independent selections of data $\{\phi_i : i = 1, \ldots, n\}$, the log-likelihood function $\ell(\Phi) \equiv \sum \log p(\phi_i \mid \Phi)$, gives the logarithm of the likelihood (up to an additive constant) of this sample having been selected from the distribution $p(\phi \mid \Phi)$ with parameter Φ. The value of Φ for which $\ell(\Phi)$ is an absolute maximum corresponds to the best estimate for the parameter based on the maximum-likelihood estimation. Unconstrained maxima may be found using elementary calculus by solving the 'likelihood equation' $\ell'(\Phi) = 0$, where primes on $\ell(\Phi)$ denote differentiation with respect to Φ.

Fisher treated this equation statistically for asymptotically large sample sizes to solve for the distribution $P(\Delta\Phi)$ of the error $\Delta\Phi$ in the parameter estimate, i.e. $\Delta\Phi = \Phi - \Phi_0$, with Φ_0 the actual parameter used in obtaining the sample. He was able to show that for asymptotically large sample sizes there was only one real solution (with any probability) to the likelihood equation.

In this section the statistical solution of the likelihood equation will be extended to obtain $O(1/n)$ corrections to $P(\Delta\Phi)$. For finite samples the likelihood equation does not asymptote to a trivial equation with a single real root. Because of this there are two difficulties that arise with the use of the likelihood equation to obtain the maximum likelihood. All local maxima and minima of the likelihood function will be roots to the likelihood equation, but only one of them can be the absolute maximum. (This difficulty also exists in the method of uniformly accurate approximations to distributions [5], but it is not recognized there.) Further, any constraint on the parameter values may invalidate the use of the unconstrained likelihood equation. Restricting the parameter to be real does not invalidate the likelihood equation.

In what follows, these two difficulties will be neglected, since the aim here is to determine *sufficient* conditions for when Fisher's asymptotic results cannot be trusted without an independent check (such as numerical simulation). For this purpose, it suffices to know when the $O(1/n)$ corrections calculated here are non-negligible. It is important to note, as illustrated in section 4, that the results here cannot be expected to be used to calculate the behaviour of $P(\Delta\Phi)$ for arbitrary sample sizes.

Fisher studied the Taylor expansion of the likelihood equation about the actual parameter value Φ_0:

$$0 = \ell'(\Phi_0) + \Delta\Phi\ell''(\Phi_0) + \frac{(\Delta\Phi)^2}{2}\ell'''(\Phi_0) + \cdots . \tag{2.1}$$

For asymptotically large n, application of the central limit theorem to expression (2.1) shows that all but the first two terms may be ignored. To keep the $O(1/n)$ corrections to the solutions to this equation, corrections to this order in the central limit theorem must be included. The appendix gives the corrections to the central limit theorem

that are needed here. To apply these corrections it is useful to define the multivariate variables

$$x_i(\phi) \equiv \frac{\mathrm{d}^i}{\mathrm{d}\Phi_0^i} \log p(\phi \mid \Phi_0)$$

$$X_i \equiv \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [x_i(\phi_j) - \mu_i]$$

where the means μ_i are given by $\mu_i \equiv \int \mathrm{d}\phi\, p(\phi \mid \Phi_0) x_i(\phi)$. Condition (2.1) now simplifies to

$$0 = \sum_i A_i (X_i + \sqrt{n}\mu_i)$$

where $A_m \equiv (\Delta\Phi)^m / m!$. (The convention that repeated subscripts are summed over is followed throughout this paper.) The quantities c_{ij}, c_{ijk}, and c_{ijkl} are defined by expression (A2) as the expectations of powers of $[x_i(\phi) - \mu_i]$ with respect to the probability distribution $p(\phi \mid \Phi_0)$, i.e.

$$c_{ij} = E\{[x_i(\phi) - \mu_i][x_j(\phi) - \mu_j]\}$$

$$c_{ijk} = E\{[x_i(\phi) - \mu_i][x_j(\phi) - \mu_j][x_k(\phi) - \mu_k]\}$$

$$c_{ijkl} = E\{[x_i(\phi) - \mu_i][x_j(\phi) - \mu_j][x_k(\phi) - \mu_k][x_l(\phi) - \mu_l]\} - 3c_{ij}c_{kl}\,.$$

The probability distribution $P(\Delta\Phi)$ of the deviation $\Delta\Phi$ of the estimate from the true parameter value Φ_0 is given by

$$P(\Delta\Phi) \propto \int \prod_i \mathrm{d}X_i P(X_j) \delta[A_k(X_k + \sqrt{n}\mu_k)]$$

where $\delta(x)$ is the Dirac delta function which may be thought of as a shorthand for any Jacobian factors needed, and the distribution $P(X_i)$ of the variable X_i is given asymptotically by the Fourier transform of expression (A1). (Note the proportionality sign which is a reminder of the need to normalize the resulting distribution.) The goal of this section is to obtain an asymptotic expansion for this probability for a large number n of sample points. An asymptotic expansion of this expression as it stands would be very difficult, if for no other reason than the difficulty of obtaining $(c^{-1})_{ij}$, so first $P(\Delta\Phi)$ is rewritten using the Fourier integral theorem. This gives

$$P(\Delta\Phi) \propto \int \frac{\mathrm{d}k}{2\pi} \prod_i \left(\mathrm{d}X_i \frac{\mathrm{d}K_i}{2\pi}\right) \exp(\mathrm{i}K_j X_j) \chi_n(K_m) \exp[\mathrm{i}kA_l(X_l + \sqrt{n}\mu_l)]$$

$$= \int \frac{\mathrm{d}k}{2\pi} \prod_i \mathrm{d}K_i \exp(\mathrm{i}k\sqrt{n}A_l\mu_l) \delta(kA_j + K_j) \chi_n(K_m)$$

$$= \int \frac{\mathrm{d}k}{2\pi} \exp\,(\mathrm{i}k\sqrt{n}A_l\mu_l) \chi_n(-kA_j)$$

By using (A1), the asymptotic expansion up to $O(1/n)$ of $P(\Delta\Phi)$ is

$$P(\Delta\Phi) \propto \int \frac{dk}{2\pi} \exp(ik\sqrt{n}A_q\mu_q) \exp\left(-\frac{k^2}{2}A_mA_nc_{mn}\right)\left[1 - \frac{ik^3A_iA_jA_kc_{ijk}}{6\sqrt{n}} + \frac{3k^4A_iA_jA_kA_lc_{ijkl} - k^6(A_iA_jA_kc_{ijk})^2}{72n} + O\left(\frac{1}{n^{3/2}}\right)\right].$$

It is still necessary to expand the exponentials in powers of $1/\sqrt{n}$. It is useful to define a change of variables by $B \equiv (A_iA_jc_{ij})^{-1/2}$, and $y \equiv \sqrt{n}BA_i\mu_i$, which yields

$$P(\Delta\Phi) \propto \frac{B}{(2\pi)^{1/2}} \exp\left(-\frac{y^2}{2}\right)\left[1 - \frac{H_3(y)B^3A_iA_jA_kc_{ijk}}{6\sqrt{n}} + \frac{3H_4(y)B^4A_iA_jA_kA_lc_{ijkl} + H_6(y)B^6(A_iA_jA_kc_{ijk})^2}{72n} + O\left(\frac{1}{n^{3/2}}\right)\right] \tag{2.2}$$

where the Hermite polynomials $H_n(y)$ are given by

$$\int \frac{dK}{2\pi} e^{iKX} e^{-K^2/2} K^n = \frac{1}{i^n}\frac{d^n}{dX^n}\left(\frac{1}{(2\pi)^{1/2}}e^{-X^2/2}\right) = i^n H_n(X)\frac{1}{(2\pi)^{1/2}}e^{-X^2/2}.$$

A great simplification can now be made by noting that the expectation of a derivative of the likelihood function is zero; this implies that $\mu_1 = 0$. Using this result, the asymptotic form of $P(\Delta\Phi)$ can be obtained as an expansion in powers of $1/\sqrt{n}$ about Fisher's Gaussian result, which has the form

$$P_0(\Delta\Phi) = \left(\frac{n\mu_2^2}{2\pi c_{11}}\right)^{1/2} \exp\left(\frac{-n\mu_2^2(\Delta\Phi)^2}{2c_{11}}\right). \tag{2.3}$$

The expectations of $(\sqrt{n}\Delta\Phi)^m$ will be of $O(1)$, as can be verified by integrating them over the P_0 in (2.3). Thus, for the purposes of expanding about Fisher's result, $\Delta\Phi$ will be of $O(1/\sqrt{n})$. Taking this into account allows an expansion of the various terms making up (2.2)

$$B = \frac{1}{\sqrt{c_{11}}}\left[1 - \frac{c_{12}\Delta\Phi}{c_{11}} + \frac{(\Delta\Phi)^2}{2c_{11}^2}(3c_{12}^2 - c_{11}c_{13} - c_{11}c_{22}) + O\left(\frac{1}{n^{3/2}}\right)\right] \tag{2.4}$$

and

$$\exp\left(\frac{-y^2}{2}\right) = \exp\left(\frac{-n\mu_2^2(\Delta\Phi)^2}{2c_{11}}\right)\left[1 + \frac{n\mu_2(\Delta\Phi)^3}{2c_{11}^2}(2\mu_2c_{12} - \mu_3c_{11}) - \frac{n(\Delta\Phi)^4}{24c_{11}^3}(48\mu_2^2c_{12}^2 - 24\mu_2\mu_3c_{11}c_{12} - 12\mu_2^2c_{11}c_{13} - 12\mu_2^2c_{11}c_{22} + 4\mu_2\mu_4c_{11}^2 + 3\mu_3^2c_{11}^2) + \frac{n^2\mu_2^2(\Delta\Phi)^6}{8c_{11}^4}(2\mu_2c_{12} - \mu_3c_{11})^2 + O\left(\frac{1}{n^{3/2}}\right)\right]. \tag{2.5}$$

The final terms of (2.2) come from the explicit corrections to the Gaussian form of the central limit theorem. Again, since $\Delta\Phi$ is $O(1/\sqrt{n})$, some of the terms in (2.2) of $O(\Delta\Phi^2)$ with factors of $\sqrt{n}$ in the denominator, are actually negligible to $O(1/n)$. Keeping only terms that are genuinely $O(1/n)$ gives

$$-\frac{H_3(y)B^3A_iA_jA_kc_{ijk}}{6\sqrt{n}} = -\frac{c_{111}}{6(nc_{11})^{1/2}} \times H_3\left\{\left(\frac{n\mu_2^2}{c_{11}}\right)^{1/2}\Delta\Phi\left[1+\frac{\Delta\Phi}{2}\left(\frac{\mu_3}{\mu_2}-\frac{2c_{12}}{c_{11}}\right)\right]\right\} \times\left[1+\Delta\Phi\left(\frac{3c_{112}}{c_{111}}-\frac{c_{12}}{c_{11}}\right)\right]+O\left(\frac{1}{n^{3/2}}\right) \tag{2.6}$$

and

$$\frac{3H_4(y)B^4A_iA_jA_kA_lc_{ijkl}+H_6(y)B^6(A_iA_jA_kc_{ijk})^2}{72n} = \frac{1}{72nc_{11}^2}\left\{3H_4\left[\left(\frac{n\mu_2^2}{c_{11}}\right)^{1/2}\Delta\Phi\right]c_{1111} + H_6\left[\left(\frac{n\mu_2^2}{c_{11}}\right)^{1/2}\Delta\Phi\right]\frac{c_{111}^2}{c_{11}}\right\}+O\left(\frac{1}{n^{3/2}}\right) \tag{2.7}$$

where the arguments of the Hermite polynomials are shown explicitly.

So, equation (2.2), with (2.4) and (2.5) substituted into the prefactors and (2.6) and (2.7) in the square brackets, give the asymptotic expansion of $P(\Delta\Phi)$, that is, the asymptotic solution to the likelihood equation including first corrections for a finite sample. The $O(1/n)$ corrections to $P(\Delta\Phi)$ correspond to keeping terms up to cubic order in $\Delta\Phi$ in the expansion (2.1). The statistical approach used here includes both the correlations between terms in the expansion of the likelihood equation, and the corrections to the central limit theorem.

3. Translation families

In this section explicit conditions for the failure of the Gaussian likelihood approximation are obtained for the case of estimation of parameters of the so-called translation family. That is, the probability density takes the form $p(\phi \mid \Phi) = f(\phi - \Phi)$. It is further assumed that this density satisfies the 'sensible' boundary conditions

$$\int d\phi \frac{d}{d\phi} g[p(\phi \mid \Phi)] = 0$$

for any sufficiently 'sensible' function $g(p', p'', p''', p'''')$ which includes only up to the fourth derivatives of $p(\phi \mid \Phi)$ (recall that primes on p denote differentiation with

respect to ϕ). In this case the expectations of the derivatives of the log-likelihood function can be simplified by integration by parts:

$$E[x_2(\phi)] \equiv \mu_2 = -\int \mathrm{d}\phi \frac{p'^2}{p} \equiv -\mathcal{I}$$

$$E[x_3(\phi)] \equiv \mu_3 = -\frac{1}{2}\int \mathrm{d}\phi \frac{p'^3}{p^2}$$

$$E[x_4(\phi)] \equiv \mu_4 = \int \mathrm{d}\phi \left(\frac{p''^2}{p} - \frac{2}{3}\frac{p'^4}{p^3}\right)$$

(recall that $\mathcal{I}$ is the Fisher information). The higher-order moments can also be simplified:

$$\begin{aligned}
&E[x_1(\phi)^2] = c_{11} = \int \mathrm{d}\phi \frac{p'^2}{p} = \mathcal{I} \\
&E[x_1(\phi)x_2(\phi)] = c_{12} = \frac{1}{2}\int \mathrm{d}\phi \frac{p'^3}{p^2} = -\mu_3 \\
&E[x_1(\phi)x_3(\phi)] = c_{13} = \int \mathrm{d}\phi \left(\frac{2}{3}\frac{p'^4}{p^3} - \frac{p''^2}{p}\right) = -\mu_4 \\
&E[x_2(\phi)^2] = c_{22} + \mathcal{I}^2 = \int \mathrm{d}\phi \left(\frac{p''^2}{p} - \frac{1}{3}\frac{p'^4}{p^3}\right) \\
&E[x_1(\phi)^3] = c_{111} = -\int \mathrm{d}\phi \frac{p'^3}{p^2} = 2\mu_3 \\
&E[x_1(\phi)^2 x_2(\phi)] = c_{112} - \mathcal{I}^2 = -\frac{1}{3}\int \mathrm{d}\phi \frac{p'^4}{p^3} = \mu_4 - c_{22} - \mathcal{I}^2 \\
&E[x_1(\phi)^4] = c_{1111} + 3\mathcal{I}^2 = \int \mathrm{d}\phi \frac{p'^4}{p^3} = 3(c_{22} + \mathcal{I}^2 - \mu_4)\,.
\end{aligned} \tag{3.1}$$

Substituting these simplified expressions for translation families into (2.2) to (2.7) gives corrections to $P(\Delta\Phi)$. For a completely general distribution of the translation family the expression is lengthy and therefore not written here. Only the expectations of the error and the square error are given here as

$$E(\Delta\Phi) = -\frac{\mu_3}{2n\mathcal{I}^2}\left[1 + \frac{1}{6n\mathcal{I}^3}(453\mu_3^2 + 86\mu_4\mathcal{I} - 159c_{22}\mathcal{I}) + \mathrm{O}\left(\frac{1}{n^{3/2}}\right)\right]$$

and

$$E[(\Delta\Phi)^2] = \frac{1}{n\mathcal{I}}\left[1 + \frac{1}{4n\mathcal{I}^3}(8c_{22}\mathcal{I} - 5\mu_3^2) + \mathrm{O}\left(\frac{1}{n^{3/2}}\right)\right].$$

Recall that (2.2) has a proportionality sign so that for calculating the above moments, a normalization constant is needed. This normalization for $P(\Delta\Phi)$ is given by

$$\int \mathrm{d}\Delta\Phi\, P(\Delta\Phi) \propto \frac{1}{\sqrt{n\mathcal{I}}}\left[1 + \frac{1}{2n\mathcal{I}^3}(2c_{22}\mathcal{I} - \mu_4\mathcal{I} - 3\mu_3^2) + \mathrm{O}\left(\frac{1}{n^{3/2}}\right)\right].$$

The moments c_{111}, c_{112} and c_{1111} which arise from corrections to the central limit theorem do not appear *explicitly* in the above expressions because for the translation family equations (3.1) give them in terms of c_{22}, μ_3 and μ_4. Nevertheless, the central limit theorem corrections have *implicitly* affected the $O(1/n)$ results, and were therefore an essential part of the expansion procedure. Different $O(1/n)$ corrections would have been erroneously obtained if the corrections to the central limit theorem were not included in the solution of the likelihood equation.

A case of special interest is when the maximum likelihood estimator is unbiased, that is, when $\mu_3 = 0$. In this case $E(\Delta\Phi) = 0$,

$$\mathrm{var}(\Delta\Phi) = \frac{1}{n\mathcal{I}}\left\{1 + \frac{2}{n}\left[\frac{1}{\mathcal{I}^2}\int \mathrm{d}\phi\left(\frac{p''^2}{p} - \frac{1}{3}\frac{p'^4}{p^3}\right) - 1\right] + O\left(\frac{1}{n^{3/2}}\right)\right\} \tag{3.2}$$

and by making the change of variable to $Y = (n\mathcal{I})^{1/2}\Delta\Phi$, the normalized probability distribution becomes

$$P(Y) = (2\pi)^{1/2}\exp(-Y^2/2)\left[1 + \frac{1}{8}\frac{c_{22}}{n\mathcal{I}^2}(Y^4 + 2Y^2 - 5)\right.$$
$$\left. + \frac{1}{24}\frac{\mu_4}{n\mathcal{I}^2}(Y^4 - 6Y^2 + 3) + O\left(\frac{1}{n^{3/2}}\right)\right]. \tag{3.3}$$

It is easy to turn this result into a condition for the failure of Fisher's asymptotic result. If

$$n \leqslant n_{\mathrm{var}}\frac{100\%}{e\%}$$

where n_{var} is given by

$$n_{\mathrm{var}} \equiv \left[\frac{2}{\mathcal{I}^2}\int \mathrm{d}\phi\left(\frac{p''^2}{p} - \frac{1}{3}\frac{p'^4}{p^3}\right) - 2\right]$$

then the $O(1/n^2)$ corrections to the variance $\mathrm{var}(\Delta\Phi)$ differ from Fisher's asymptotic result of $1/(n\mathcal{I})$ by more than $e\%$.

The second condition derived for the failure of Fisher's asymptotic result for the unbiased maximum likelihood estimator requires that the kurtosis

$$\kappa = \mathrm{var}(\Delta\Phi)^2\left[\frac{3c_{22} + \mu_4}{n\mathcal{I}^2} + O\left(\frac{1}{n^2}\right)\right]$$

of $P(\Delta\Phi)$ be non-negligible. (Recall that $\kappa = 0$ for a Gaussian distribution.) In particular, if

$$n \leqslant n_{\mathrm{kur}}\frac{100\%}{f\%}$$

where

$$n_{\mathrm{kur}} \equiv \left|\frac{1}{\mathcal{I}^2}\int \mathrm{d}\phi\left(\frac{4p''^2}{p} - \frac{5}{3}\frac{p'^4}{p^3}\right) - 3\right|$$

then the first corrections to the kurtosis of $P(\Delta\Phi)$ will be more than $f\%$ of the square of its variance.

A nice feature of these conditions is that they did not require any assumptions beyond an unbiased translation family distribution. The use of these conditions will now be illustrated for two classes of distributions in the unbiased translation family.

These two classes were chosen with very different characteristics. The first is a power of a Gaussian distribution, which has a flat peak, is box-like, and has an insignificant tail. Their probability densities are given by

$$p_m^{(1)}(\phi \mid \Phi_0) = \frac{m}{\Gamma(1/2m)} \exp\left[-(\phi - \Phi_0)^{2m}\right] \tag{3.4}$$

with m a positive integer. For $m = 1$ this reduces to a Gaussian. As m increases the distribution becomes more box-like.

In contrast, the second class of distributions have thin peaks, and broad tails. Their probability densities are given by

$$p_M^{(2)}(\phi \mid \Phi_0) = \frac{1}{2\pi}\left(\sum_{j=1}^{M+1} \frac{1}{j^2}\right)^{-1} \left| \sum_{k=1}^{M+1} \frac{\exp[ik(\phi - \Phi_0)]}{k} \right|^2 \tag{3.5}$$

over the interval $\phi \in [-\pi, \pi)$, with M a positive integer. For $M = 1$, this reduces to a uniform distribution which is of no interest for maximum likelihood estimation. However, as M increases the distribution develops a sharp peak (approximated by a logarithmic divergence) and keeps its wide tail.

Tables 1 and 2 show the type of predictions that can be easily generated for many distributions, and that could prove useful in the design of experiments. These tables show, for various members of the classes (3.4) and (3.5), the number of data points n_{var} and n_{kur} which [to $O(1/n)$ in $P(\Delta\Phi)$] correspond, respectively, to a 100% deviation from Fisher's variance, and to a kurtosis which is 100% of the square of the variance of the estimate.

Table 1. Example of the type of table that can be generated from the results in this paper to be used in the design of experiments. For the class of distributions (3.4), $5n_{\text{var}}$ and $5n_{\text{kur}}$ estimate the number of data points at which the Gaussian approximation breaks down (coresponding to an error of 20% for $e\%$ and $f\%$ respectively). The distribution is Gaussian for $m = 1$; increasing m makes it more 'boxy'.

m	n_{var}	n_{kur}
1	0.0	0.0
2	2.377	2.106
3	5.743	5.517
4	9.378	9.191
5	13.14	12.98
10	32.61	32.52

As a trivial example, in table 1, $m = 1$ corresponds to a Gaussian distribution. For any non-zero choices of the confidence parameters e and f, both of the above conditions require that $n \leqslant 0$, but this is in exact agreement with an analytic calculation of the maximum likelihood behaviour, since this distribution has a Gaussian

Table 2. Example of the type of table that can be generated from the results in this paper to be used in the design of experiments. For the class of distributions (3.5), $5n_{\mathrm{var}}$ and $5n_{\mathrm{kur}}$ estimate the number of data points at which the Gaussian approximation breaks down (corresponding to an error of 20% for $e\%$ and $f\%$ respectively). The distribution has a long tail and a peak that logarithmically diverges as M increases.

M	n_{var}	n_{kur}
2	4.814	9.854
4	6.065	12.36
12	11.79	23.68
33	26.19	52.15
50	37.41	74.36
74	52.92	105.1

likelihood function for *all* sample sizes, with a variance identical to that predicted by Fisher's asymptotic results.

As a less trivial example, consider $p_4^{(1)}$ which is tabulated as $m = 4$ in table 1. Both n_{var} and n_{kur} are only a little larger than nine data points. Thus, if one wanted to perform an experiment which would be analysed assuming an approximately Gaussian likelihood function with the confidence parameters $e\%$ and $f\%$ approximately 5%, then one would need to sample around 180 to 190 data points. Further, it would be inappropriate to analyse smaller samples than this using techniques which assumed an approximately Gaussian likelihood function.

As a final example, table 2 predicts that $n_{\mathrm{kur}} \sim 100$ for the distribution corresponding to the one shown in figure 1. Choosing the kurtosis confidence parameter $f\%$ as 10% leads to the prediction that one should not assume that the likelihood function is sufficiently close to being Gaussian when the sample size is less than, say, 1000 data points!

4. Numerical simulations

The results of section 3 are quite convenient for the design of experiments because they are simple and applicable to many distributions. Nevertheless, the user might want some evidence of their validity. For this purpose, some comparison with numerical simulation is presented in this section.

Expression (3.3) gives the $O(1/n)$ correction to Fisher's asymptotic Gaussian result since these distributions are members of the unbiased translation family. For comparison, the actual statistics for $P(\Delta\Phi)$ were calculated using Monte Carlo simulation. This involves making repeated simulations of an 'experiment' in order to build up the distribution of the estimator. Each 'experiment' consists of selecting n points independently from a distribution, calculating the likelihood function, and finding the location of the absolute maximum. Since the translation family distributions depend only on the difference between the random variable ϕ and the actual parameter Φ_0 then for the purposes of simulation the actual parameter value may be chosen as zero, i.e. $\Phi_0 = 0$. The details of the computer algorithm and errors analysis are discussed in some detail elsewhere [6].

Tables 3 and 4 compare the results of section 3 with numerical simulations for the choices $m = 4$, and $M = 12$ in the distributions (3.4) and (3.5) respectively. These tables show the predicted deviation (based on (3.3) for $P(\Delta\Phi)$) from Fisher's

asymptotic prediction of $1/(n\mathcal{I})$ next to the actual percentage deviation as calculated by Monte Carlo simulation. For both cases $n_{\text{var}} \sim 10$ data points. For samples consisting of only 10 data points the predicted and actual deviations qualitatively agree, in that they both say there are large deviations from Fisher's results. However, they do not agree quantitatively. The quantitative agreement becomes better as successively larger sample sizes are studied. For these two cases, the tabulated results show that when the number of data points is large enough that $e\% \sim 20\%$, then there is even good quantitative agreement between the predicted and actual per cent deviations from Fisher's result. Thus, the conditions of the previous section are validated.

Table 3. The predicted percentage deviation of the variance in the estimate compared to the actual percentage deviation from Fisher's result for distribution (3.4) when $m = 4$. (The uncertainties represent one standard error for the variance.)

Number of data points	Predicted percentage deviation from Fisher's result	Actual percentage deviation from Fisher's result
10	93.8%	59.5% ± 2.1%
20	46.9%	22.3% ± 1.6%
30	31.3%	20.5% ± 3.0%
40	23.4%	15.8% ± 4.0%
60	15.6%	21.2% ± 10.1%

Table 4. The predicted percentage deviation of the variance in the estimate compared to the actual percentage deviation from Fisher's result for distribution (3.5) when $M = 12$. (The uncertainties represent one standard error for the variance.)

Number of data points	Predicted percentage deviation from Fisher's result	Actual percentage deviation from Fisher's result
10	117.9%	644.2% ± 38.9%
20	59.0%	132.7% ± 18.7%
30	39.3%	57.5% ± 4.6%
40	29.5%	30.5% ± 2.5%
60	19.7%	17.2% ± 2.3%

Why is there only poor quantitative agreement between the predicted and actual deviations for a smaller number of data points? It is because the predicted deviations are based only on the first corrections to the solution to the likelihood equation. Thus, the corrections to $P(\Delta\Phi)$ should *not* be used as a better approximation to the behaviour of the distribution of the estimate. The problem with doing this is illustrated in figures 2 and figure 3. In figure 2 the $O(1/n^2)$ correction for the estimate variance is shown by the dotted curve, and it overestimates the actual values (except for samples with 60 data points), whereas in figure 3 this correction underestimates the actual values (except for samples with a single data point or with 60 data points). These corrections do not predict well the small sample-size behaviour. In both cases, however, the Fisher prediction given by the broken curves is *always* an underestimate of the estimate variance, a consequence of the Cramér–Rao lower bound.

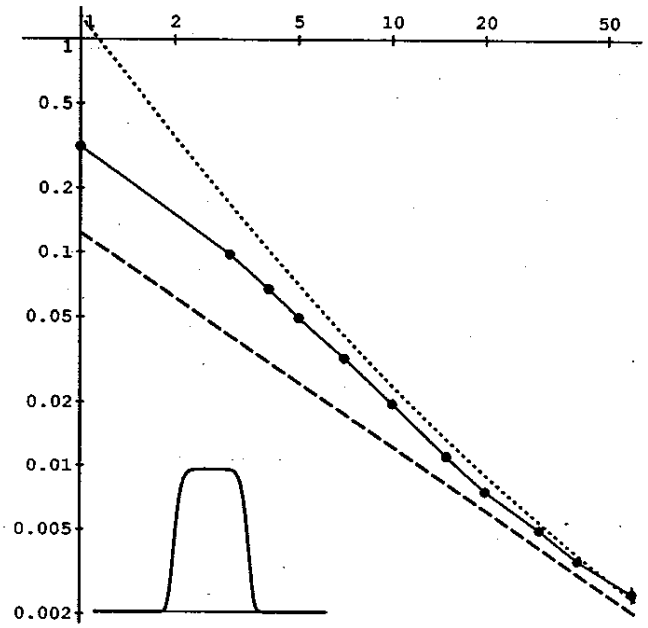

Figure 2. The variance var($\Delta\Phi$) of the estimate $\Delta\Phi$, for distribution (3.4) (shown in the inset) with $m = 4$, as computed by Monte Carlo simulations, as a function of n, the number of data points (full curve). For comparison, the broken curve represents the asymptotic Fisher result of $1/(n\mathcal{I})$, and the dotted curve represents the corrected variance given by expression (3.2). Fisher's result is an underestimate of the estimate variance, and the predicted correction from this paper gives an overestimate (except for samples with 60 data points).

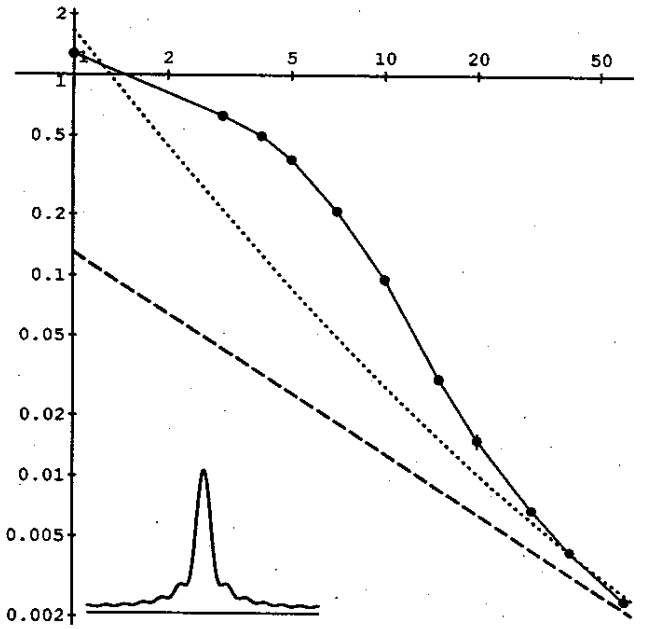

Figure 3. The variance var($\Delta\Phi$) of the estimate $\Delta\Phi$, for distribution (3.5) (shown in the inset) with $M = 12$, as computed by Monte Carlo simulations as a function of n, the number of data points (full curve). For comparison, the broken curve represents the asymptotic Fisher result of $1/(n\mathcal{I})$, and the dotted curve represents the corrected variance given by (3.2). Fisher's result is an underestimate of the estimate variance, and the predicted correction from this paper also gives an underestimate (except for samples with a single data point or with 60 data points).

5. Concluding remarks

The main result of this paper is a technique for obtaining explicit conditions for the failure of the Gaussian approximation to the likelihood function. While in this paper such conditions were actually only derived for the unbiased maximum likelihood estimation of single-parameter distributions in the translation family, enough details have been given in the derivation, so that conditions for more general classes of distributions, including multiparameter and multivariate distributions, may be derived.

These results are useful for the design of experiments if one is interested in applying simple methods of data analysis in maximum-likelihood estimation. These simplified methods are widely used, but their validity is seldom checked. The procedure given here can help assure that the error bar for a parameter estimate is not significantly *underestimated.*

Finally, the results presented here give the efficiency of maximum-likelihood estimation inside the asymptotic regime and also when this regime is reached. There has been a growing interest in the physics community in the design of 'optimally' efficient experiments. As these designs usually assume maximum-likelihood estimation is to be used as the method of data analysis, there is a role for the results of this paper in this endeavor.

Acknowledgments

The author is indebted to Dr Ardith W El-Kareh for many fruitful discussions, and to the hospitality of Research House, Tucson, where much of this work was done.

Appendix. Central limit theorem corrections

This appendix describes corrections to the multivariate central limit theorem for independently selected identical samples [3]. It will be sufficient to give the corrections in terms of the characteristic function. Consider the multivariate distribution $p(x_i)$. The goal is to obtain an approximation, for large n, of the distribution of the variable

$$X_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [(x_i)_j - \mu_i]$$

with μ_i the mean of x_i over $p(x_i)$, and $(x_i)_j$ the ith component of the jth piece of data. The characteristic function $\chi_n(K_i)$ for this distribution is then given by

$$\chi_n(K_i) = \int \prod_j \mathrm{d}X_j \exp(-\mathrm{i}K_l X_l) P(X_m) = \exp\left\{ n \log\left[\chi_1\left(\frac{K_i}{\sqrt{n}}\right)\right]\right\}$$

$$= \exp(-K_m K_n c_{mn}/2)\left[1 + \frac{iK_i K_j K_k c_{ijk}}{6\sqrt{n}} + \frac{3K_i K_j K_k K_l c_{ijkl} - (K_i K_j K_k c_{ijk})^2}{72n} + \mathrm{O}\left(\frac{1}{n^{3/2}}\right)\right] \qquad \text{(A1)}$$

where $\chi_1(k_i)$ is the characteristic function for $p(x_i)$, and defining $\Delta x_i = x_i - \mu_i$ the multivariate moments are given by

$$\begin{aligned} c_{ij} &\equiv E(\Delta x_i \Delta x_j) \\ c_{ijk} &\equiv E(\Delta x_i \Delta x_j \Delta x_k) \\ c_{ijkl} &\equiv E(\Delta x_i \Delta x_j \Delta x_k \Delta x_l) - 3c_{ij}c_{kl}\,. \end{aligned} \tag{A2}$$

References

[1] Braunstein S L 1992 *Phys. Rev. Lett.* submitted
[2] Efron B and Hinkely D V 1978 *Biometrika* 65 457
[3] Feller W 1971 *An Introduction to Probability Theory and its Applications* vol 2 (New York: Wiley) p 525
[4] Fisher R A 1925 *Proc. Camb. Phil. Soc.* 22 700
[5] Hinkley D V, Reid N and Snell E J 1991 *Statistical Theory and Modelling* (London: Chapman and Hall) p 287
[6] Lane A S, Braunstein S L and Caves C M 1992 *Phys. Rev.* A submitted
[7] Lehmann E L 1983 *Theory of Point Estimation* (New York: Wiley) p 415
[8] Rao C R 1973 *Linear Statistical Inference and its Applications* (New York: Wiley)
[9] Shapiro J H, Shepard S R and Wong N W 1990 *Phys. Rev. Lett.* 62 2377