

## **Statistical Inference: Likelihood to Significance**

D.A.S. Fraser

York University, Mathematics Department, North York, Canada M3J 1P3

### **Summary**

The concepts of likelihood and significance were defined and initially developed by R.A. Fisher, but followed almost separate and distinct routes. We suggest that a central function of statistical inference is in fact the conversion of the first, likelihood, into the second, significance: a linking of the Fisher concepts. A first order asymptotic route for this is incorporated into most statistical packages. It uses the standardized maximum likelihood estimate, the standardized score, or the signed square root of the likelihood ratio statistic as arguments for the standard normal distribution function thus giving approximate tail probabilities or observed levels of significance. Recent third order asymptotic methods provide a substantial increase in accuracy but need the first derivative dependence of likelihood on the data value as an additional input. This can be envisaged as the effect on the likelihood function of dithering the data point. Extensions to the multivariate, multiparameter context are surveyed indicating major areas for continuing research.

**Keywords:** Likelihood; Lugannani and Rice formula;  $p$ -value; Significance; Tail probability; Third order asymptotics.

## 1. INTRODUCTION

The central area of Fisher's greatest influence in statistics covers the formulation and analysis of parametric statistical models, both the development of models for specific applications and the development of theory for general application. His isolation of the likelihood function as the relevant inferential component of these models laid the foundations for most of statistical theory. His parallel development of statistical analyses by means of observed levels of significance, assessing the consistency of data with model, had great impact in practical statistics. Recent work indicates that these concepts can be very closely linked and suggests that this linking is the substance of statistical inference.

In Section 2 a simple example from Fisher (1925) is used to show a close correspondence between likelihood as probability *at* a data point and significance as probability *left of* a data point. Some discussion is included concerning the use and interpretation of significance.

The standardized maximum likelihood estimate, the standardized score, and the signed square root of the likelihood ratio statistic lead to approximate  $p$ -values or observed levels of significance by calculating the corresponding value from the standard normal distribution function. Section 3 discusses these first order asymptotic approximations which are based directly or indirectly on the Central Limit Theorem. A discrete and a continuous example are used to show that the three approximations can give quite different conversions of likelihood to approximate significance.

Section 4 examines some recent third order asymptotic conversions of likelihood to significance, for a real parameter of interest. Lugannani and Rice (1980) used saddlepoint methods to obtain an  $O(n^{-\frac{3}{2}})$  approximation to a distribution function from a corresponding cumulant generating function or from a corresponding likelihood function in the exponential model context. The formula was made parameterization invariant in Fraser (1990) and shown to be  $O(n^{-\frac{3}{2}})$  for a general model context in Fraser and Reid (1990). The two examples from Section 3 are examined again and the approximation is seen to be

almost exact in a plot of the observed significance.

Section 5 describes a simple numerical program that takes an observed likelihood function  $\ell(\theta)$  as input and outputs the observed significance function  $p(\theta)$ . For the generalization from the exponential model context a reparameterization  $\phi(\theta)$  is also needed and is obtained as a sample space derivative of the likelihood function. The program thus converts  $\ell(\cdot)$  and  $\phi(\cdot)$  into  $p(\cdot)$ .

Section 6 surveys the theory available for the general context of a vector variable and a vector parameter. Let  $\theta = (\psi, \lambda)$  with  $\psi$  being the scalar parameter of interest; and let  $y$  be the variable after appropriate reduction by sufficiency or conditioning. To apply the preceding numerical conversion procedure two things are needed: a likelihood function  $\ell(\psi)$  for the component parameter and a reparameterization  $\phi(\psi)$  of that parameter: the likelihood is obtained as an adjusted profile likelihood, and the reparameterization as a sample space directed derivative of likelihood. This is discussed for the exponential model, the transformation model, and for curved components of these models.

Section 7 records some concluding remarks.

## 2. A SIMPLE EXAMPLE

Fisher (1925) examined data that had been collected by Cushney and Peebles and analyzed earlier by Student (1908). The 10 observations are a measure of improvement under a change in drug therapy: 1.2 2.4 1.3 1.3 0.0 1.0 1.8 0.8 4.6 1.4. For simple illustration here we follow Fisher and analyze the data using a normal model and in addition assume that the standard deviation  $\sigma = 1.5$  is known. Sufficiency simplifies the sample space to one dimension and we have

$$\text{Model : } \bar{y} \sim \text{Normal}(\theta, 0.4743)$$

$$\text{Data : } \bar{y}^0 = 1.58$$

where  $1.5/\sqrt{10} = 0.4743$  and the observed sample average is 1.58.

### 2.1 Likelihood

The likelihood function  $L(\theta) = cf(\bar{y}^0; \theta)$  provides an assessment of the data and is a function of possible values for the parameter: it records the amount of probability *at the data point*, and in accord with Fisher is usually left indeterminate to an arbitrary multiplicative constant.

Consider the  $\bar{y}$  axis with the designated observed value 1.58. If  $\theta$  is large, say 2.53, i.e., far to the right of the data, then density at 1.58 is 0.14 (relative to the maximum possible density); if  $\theta = 2.05$ , then density at 1.58 is 0.61; if  $\theta = 1.58$  (exactly at the data) then density at 1.58 is 1.00; if  $\theta = 1.11$ , then density at 1.58 is 0.61; if  $\theta = 0.63$  far to the left of the data, then density at 1.58 is 0.14. These different values for the density at the data are recorded in Figure 1 and extended to the full range of  $\theta$ ; this is the *observed likelihood function*:

$$L(\theta) = c \exp \left\{ -\frac{1}{2 \times 0.225} (\theta - 1.58)^2 \right\} .$$

For many purposes it is more convenient to work with log-likelihood  $\ell(\theta) = \log L(\theta)$ ; the observed log likelihood is recorded in Figure 2:

$$\ell(\theta) = a - \frac{1}{2 \times 0.225} (\theta - 1.58)^2 ;$$

we find it convenient to call  $\ell(\theta)$  just likelihood, letting the context provide the proper clarification.

### 2.3 Significance

The significance function  $p(\theta) = P(\bar{y} \leq \bar{y}^0; \theta)$  also provides an assessment of the data and is a function of possible values for the parameter: it records probability *left of the data*.

Again consider the  $\bar{y}$  axis with the designated observed value 1.58. If  $\theta$  is large, say 2.53, far right of the data, then probability left of 1.58 is 0.977; if  $\theta = 2.05$ , then probability left of 1.58 is 0.841; if  $\theta = 1.58$  then probability left of 1.58 is 0.500; if  $\theta = 1.11$  then probability left of 1.58 is 0.159; if  $\theta = 0.63$ , then probability left of 1.58

is 0.023. These different values for the left tail probability are recorded in Figure 3 and extended to the full range for  $\theta$ . We call this left tail probability function the *significance function*  $p(\theta)$ .

A 95% confidence interval for  $\theta$  is obtained by inverting the interval (0.025, 0.975):

$$(p^{-1}(0.975), p^{-1}(0.025)) = (0.65, 2.51)$$

which is of course just  $(\bar{y}^0 \pm 1.960 \times 0.4743)$ . In fact the full spectrum of confidence intervals is obtained by inverting  $p(\theta)$ , suggesting the alternate name *confidence distribution function* for  $p(\theta)$ .

### 2.3 About significance

If  $\bar{y}^0$  is far on the left tail then  $p(\theta)$  is close to zero indicating high significance. If  $\bar{y}^0$  is far on the right tail then  $p(\theta)$  is close to 1 and  $1 - p(\theta)$  is close to zero indicating high significance. We could perhaps use  $\min\{p(\theta), 1 - p(\theta)\}/(1/2)$  as a two tailed significance level or as a conditional level of departure from the centre given the direction of departure. For our purposes here we will call  $p(\theta)$  itself the significance function and leave the two sided or directional measure as a fine tuning for specific contexts. The function  $p(\theta)$  records where the data is on the  $\theta$  distribution of  $\bar{y}$  values.

We note in conclusion that  $F(\bar{y}; \theta)$  is a pivotal quantity with a uniform distribution (0, 1), and thus  $p(\theta)$  is an observed value from this uniform distribution when  $\theta$  is the true parameter value.

### 2.4 From likelihood to significance

We have seen for a simple example that likelihood and significance are probability *at* and *left of* the data value. Each is easily available in this example.

The present example is so simple that we would not ordinarily go from likelihood  $\ell(\theta)$  to significance  $p(\theta)$ ; we would just calculate  $p(\theta)$  directly. For statistical problems quite generally however we will see that it is relatively easy to go from an  $\ell(\theta)$  as in Figure 2 to a  $p(\theta)$  as in Figure 3 and do so with high accuracy. We feel that this process is at the

very core of statistical inference. For multiparameter problems a preliminary reduction to a likelihood for a component parameter is needed.

### 3. LIKELIHOOD TO SIGNIFICANCE: FIRST ORDER

Likelihood is generally recognized as the primary summary of information in the data, given a parametric or semi-parametric model. As noted by Fisher (1925) *its dependence on the sample space is the minimal sufficient statistic*.

The typical way of obtaining probabilities from likelihood is based on the Central Limit Theorem in the context of a sample size going to infinity or more generally information going to infinity. The starting point is usually the direct limiting normality of the score function

$$S(\theta; y) = \frac{\partial}{\partial \theta} \log f(y; \theta)$$

where  $y$  here designates the full variable. The variance of the score can not be calculated from the likelihood function alone but an approximation is available as the observed information

$$\hat{j} = -\frac{\partial^2}{\partial \theta^2} \log f(y; \theta)|_{\hat{\theta}}$$

which describes the curvature of the observed likelihood function  $\ell(\theta)$  at its maximum. The Central Limit Theorem with the Law of Large Numbers then gives an asymptotic Normal(0,1) distribution for the *standardized score*  $q_2 = S(\theta)\hat{j}^{-1/2}$ . Other standard normal limiting variables are available from the likelihood itself. The most prominent and widely used is based on the maximum likelihood estimate; the standardized maximum likelihood estimate is  $q_1 = (\hat{\theta} - \theta)\hat{j}^{1/2}$ . A more recently promoted and seemingly more promising variable is the signed square root of the likelihood ratio statistic  $r = \text{sgn}(\hat{\theta} - \theta)[2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{1/2}$ .

If the likelihood function has come from a normal sample with known variance, then the three variables  $q_1$ ,  $q_2$ ,  $r$  are equal and have the familiar standardized form

$$z = (\bar{y} - \mu)/(\sigma_0^2/n)^{1/2}.$$

Otherwise the three represent different ways of seeking  $z$  from the likelihood function alone: the slope at the point  $\theta$ , the departure of  $\theta$  from the maximizing point, and the rise in likelihood from  $\theta$  to the maximizing point. In fact they seem to be the only obvious yet different ways of extracting an operational  $z$  value from likelihood.

The three standardized variables lead to three approximations for the significance function:

$$p_1(\theta) = \Phi(q_1), \quad p_2(\theta) = \Phi(q_2), \quad p_0(\theta) = \Phi(r) .$$

And we have  $p_i(\theta) = p(\theta) + O(n^{-\frac{1}{2}})$  where  $p(\theta)$  represents the exact probability in each case; the bound as given applies for a bounded range of the standardized variable.

### 3.2 An example from Fisher

Fisher (1956, p. 64, p72f) considers the simple case of 3 successes in 14 trials with parameter  $p$  as the probability of success. The model is the Binomial(14,  $p$ ), and the datum is  $y^0 = 3$ . Fisher uses the example to show the importance of likelihood and plots likelihood  $L(p)$  against  $p$  and also against the probit of  $p$ . For our purposes here we find it more suitable to work with a plot of the log likelihood  $\ell(p)$  against the logit of  $p$ , that is against  $\log\{p/(1-p)\}$ , which is the canonical parameter of the exponential model defined by the binomial. The curve obtained has somewhat the form of that in Figure 2 but with horizontal scale in the range from  $-4$  to  $+1$ .

The three approximations to the significance function  $p(\theta)$  are easily calculated and plotted in Figure 4. For central values of  $p$  they tend to agree but on the two tails they disagree sharply and would give quite different confidence intervals and quite different  $p$ -values for parameter values in the range of interest, near 0 or 1. Which, if any, should we trust?

The exact  $p(\theta)$  is also plotted in Figure 4. For the exact value we have used the mid- $p$  value obtained from the incomplete beta integral, with  $P(y < y^0)$  and  $P(y \leq y^0)$  falling on either side, an effect of the discreteness of that distribution. For this example,

the likelihood ratio approximation is significantly better than the others.

### 3.3 An example from reliability

The gamma distribution has wide application in environmetrics and reliability. We construct a simple example where the shape parameter is known and use a logarithmic scale. As model consider the location log gamma (3) model with density

$$f(y; \theta) = \frac{1}{2} \exp\{3(y - \theta) - e^{y-\theta}\} ;$$

as datum consider the value  $y^0 = 3.14$ . This could have arisen from a sample of 3 from the exponential with  $y$  taken as the logarithm of the minimal sufficient statistic.

The observed likelihood function is

$$\ell(\theta) = a - 3\theta - \exp(3.14 - \theta)$$

and gives a plot somewhat like Figure 2 but with horizontal scale in the range from 0 to 4.

The three approximations to the significance function  $p(\theta)$  again are easily calculated and are plotted in Figure 5. They agree for middle values of  $p$  but disagree sharply in the tails where they would be of interest for tests and confidence regions. The exact  $p(\theta)$  is also plotted in Figure 5.

### 3.4 Reliability of the first order approximations

The three approximations to a significance level can give even more discordant results when there is a greater distance from normality. As illustration we take four examples with a location parameter  $\theta$  and an observed  $y$  and record the significance level  $p(\theta)$  as a left tail probability in percent, based on the maximum likelihood, the score, and the likelihood ratio methods. For selected  $y$  and  $\theta$  values using the normal  $N$ , the extreme value  $EV$ , the gamma (3) $G$ , and the Cauchy  $C$ , we have in percent:

	$N$	$EV$	$G$	$C$
mle	0.173	0.014	1.69	$10^{-196}$
score	0.173	6.43	19.81	48.7
likratio	0.173	0.34	6.33	0.01

The approximations can be *very* different. Is there something more reliable, say for contexts where the underlying sampling or data collection process indicates some direction towards normality?

#### 4. LIKELIHOOD TO SIGNIFICANCE: THIRD ORDER

Saddlepoint methods have been used in applied mathematics since the last century but were introduced to statistics only recently by Daniels (1954). It was not until the paper by Barndorff-Nielsen and Cox (1979) that the importance of the method became well known. The method focuses on an approximate conversion of a cumulant generating function to a corresponding density function. For statistics this can be reformulated as a conversion from a likelihood function to a density function in the context of an exponential model.

Lugannani and Rice (1980) and Daniels (1987) used saddlepoint methods to convert a cumulant generating function to a corresponding distribution function; for statistics, a likelihood function in the exponential model context can replace the cumulant generating function. This type of saddlepoint approximation has more importance for statistics as it produces an actual probability rather than a density; and it forms the basis for the third order conversions discussed in this section. For a detailed review of saddlepoint methods in statistics, see Reid (1988).

##### 4.1 With exponential models

Lugannani and Rice (1980) developed a third order procedure for converting a cumulant generating function to a corresponding distribution function; see also Daniels (1987). For an exponential model in statistics the procedure provides a means for going from likelihood to significance with high accuracy. Let

$$f(y; \theta) = \exp\{t(y)\theta - c(\theta) + h(y)\}$$

be an exponential model with canonical parameter  $\theta$ , and let  $y^0$  be an observed value with corresponding likelihood function  $\ell(\theta)$ . The approximation uses two of the standardized

variables discussed in Section 3, the standardized maximum likelihood estimate and the signed likelihood ratio statistic

$$\begin{aligned} q_1 &= (\hat{\theta} - \theta) \hat{j}^{\frac{1}{2}} \\ r &= \text{sgn}(\hat{\theta} - \theta) [2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{\frac{1}{2}} \end{aligned}$$

together with

$$\begin{aligned} p(\theta) &= P\{t(y) \leq t(y^0); \theta\} = P(\hat{\theta} \leq \hat{\theta}^0; \theta) \\ &= \Phi(r) + \phi(r) \left\{ \frac{1}{r} - \frac{1}{q_1} \right\} + O(n^{-\frac{3}{2}}) \end{aligned}$$

where  $\phi$  and  $\Phi$  are the density and distribution function for the standard normal.

For an example, consider the Binomial  $(14, p)$  model with datum  $y^0 = 3$ , from Fisher as discussed in Section 3.2. To the accuracy of the diagram the approximation equals the exact value as recorded in Figure 4.

## 4.2 The parameterization invariant formula

The approximation formula in Section 4.1 uses the standardized maximum likelihood estimate  $q_1 = (\hat{\theta} - \theta) \hat{j}^{\frac{1}{2}}$  where  $\theta$  is assumed to be the canonical parameter of the exponential model. For cases where the canonical parameter of an exponential model has not been extracted it seems reasonable to have a version of the formula that is parameterization invariant. Indeed, if the invariant version is easily calculated, the resulting formula could easily have good performance more generally.

Fraser (1988, 1990) considered a statistical model  $f(y; \theta)$  on the real line and its approximation by an exponential model to first derivative at a data value  $y^0$ . In the case that the model is exponential, then the approximation and the original model coincide. The canonical parameter of the approximating exponential model is given as

$$\phi = \phi(\theta) = \frac{d}{dy} \ell(\theta; y)|_{y^0} = \dot{\ell}(\theta; y^0).$$

If the original model is exponential, this approximation gives the canonical parameter.

For general models with a real parameter  $\theta$  we use the standardized statistics

$$\begin{aligned} q_1 &= (\hat{\phi} - \phi) \hat{j}^{\frac{1}{2}} \\ r &= \text{sgn}(\hat{\phi} - \phi) \cdot [2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{\frac{1}{2}} \end{aligned}$$

where  $q_1$  is calculated in terms of the constructed parameter

$$\phi = \frac{d}{dy} \ell(\theta; y) |_{y^0}$$

and substitute them in the third order formula

$$p(\theta) = P(\hat{\theta} \leq \hat{\theta}^0; \theta) = \Phi(r) + \phi(r) \left\{ \frac{1}{r} - \frac{1}{q_1} \right\} .$$

Note that  $\phi = \phi(\theta)$  is data dependent, and is defined relative to the observed value  $y^0$ ; correspondingly it gives probability left of the data value as indicated by  $P(\hat{\theta} \leq \hat{\theta}^0; \theta)$ .

If there is a reduction to a real variable as with exponential and transformation models, the approximation is accurate to  $O(n^{-\frac{3}{2}})$  (Lugannani and Rice, 1980, DiCiccio, Field, Fraser, 1990, Fraser and Reid, 1990). And if an asymptotic condition holds for some version reduced to a real variable, then the approximation remains  $O(n^{-\frac{3}{2}})$  (Fraser, 1990, Fraser and Reid, 1990).

In general models, with  $\theta$  a real parameter, it is necessary to choose the direction  $d/dy$  in which the likelihood function  $\ell(\theta; y)$  is to be differentiated at the data point. Different directions may give different reparameterizations  $\phi = \phi(\theta)$ , but current work suggests that asymptotic validity remains, provided the direction chosen does not align with too few coordinate axes. The significance function  $p(\theta)$  would, for each  $\theta$ , correspond to inference from some selected conditional model (Fraser, 1990).

For an example consider the location log gamma(3) model with datum  $y^0 = 3.14$ , as discussed in Section 3.2. To apply the present approximation we need the derivative of the likelihood at the data point or as described in Section 2 the change in likelihood when the data point is moved. To apply this we examine the change in likelihood when the data point is moved from 3.14 to  $3.14 + 0.10 = 3.24$ ; see Figure 6a; the difference

$$\phi = \ell(\theta; 3.24) - \ell(\theta; 3.14)$$

appropriately scaled is recorded in Figure 6b.

To the accuracy of the diagram the approximation based on the reparameterization equals the exact value as plotted in Figure 5.

### 4.3 Accuracy of the third order approximation

In Section 3.4 we considered some simple examples with a real variable and a real parameter and saw how discordant the three first order approximations can be. For the same selected values of  $y$  and  $\theta$  with the normal  $N$ , extreme value  $EV$ , gamma (3)  $G$  and Cauchy  $C$  we record those approximations again together with the present third order approximation and the exact value:

	$N$	$EV$	$G$	$C$
mle	0.173	0.014	1.69	$10^{-196}$
score	0.173	6.43	19.81	48.7
likratio	0.173	0.34	6.33	0.01
3rd order	0.173	0.63	12.83	0.94
exact	0.173	0.63	12.47	1.06

We note that in each case the third order approximation is quite acceptable as a record of observed level of significance, certainly by comparison with the first order approximations.

This is not to suggest that for any real variable real parameter model  $f(y; \theta)$  the formula with  $\phi$  calculated at a nominal data point will give significance levels corresponding to that data value. An underlying assumption of the derivation is that the log density function, the log likelihood function, or the cumulant generating function is ‘headed’ towards the form found with normality. Thus with say a bathtub shaped density and a location parameter we would expect bad, indeed bizarre results. This targeting on normality can be expressed intuitively as log-density or cumulant generating function should be just a sum over the sample, and gives good indications of appropriate contexts for the approximation.

## 5. NUMERICALLY: FROM LIKELIHOOD TO SIGNIFICANCE

For exponential models  $f(y; \theta)$  the procedure described in Section 4.1 carries a like-

likelihood function  $\ell(\theta; y^0)$  into a significance function  $p(\theta) = P(\hat{\theta} \leq \hat{\theta}^0; \theta)$ . This can be extended to produce the distribution function  $F(\hat{\theta}; \theta)$  for arbitrary values of  $\hat{\theta}$  and  $\theta$  and surprising accuracy is found even in the  $n = 1$  case with quite nonnormal exponential families (Fraser, Reid, Wong, 1991). The computations can be handled numerically by a simple computer program that twice differences  $\ell(\theta; y^0)$  on a fine grid of input values; reasonable accuracy is needed for the tabulated likelihood values.

For general models  $f(y; \theta)$  with real  $y$  and  $\theta$  the procedure described in Section 4.2 requires in addition a reparameterization  $\phi = \phi(\theta)$  which is specific to an observed data point. Thus a tabulation of  $\{\theta, \phi(\theta), \ell(\theta; y^0)\}$  produces a tabulation of  $\{\theta, p(\theta)\}$  which records the significance function  $p(\theta) = P(\hat{\theta} \leq \theta^0; \theta)$  corresponding to the data value  $y^0$ . The computation is particularly simple if the information  $\hat{j}$  for the reparameterization is available; otherwise twice differencing and perhaps smoothing produces the needed value.

For exponential and general models the procedure is organized as a generic computer program that inputs likelihood and outputs significance.

For multiparameter and multivariate problems the procedure is still applicable for a real component parameter say  $\psi = \psi(\theta)$ . The needed ingredients for this are (i) a likelihood function appropriate to the component parameter, and (ii) a reparameterization of  $\psi$ . These issues are discussed briefly in Section 6.

## 6. MULTIPARAMETER MODELS

In the preceding sections we have been discussing significance for scalar parameters. We continue with this restriction but now examine models that have additional parameters, nuisance parameters.

As a simple example consider a sample from the Normal  $(\mu; \sigma^2)$  where  $\mu$  is the interest parameter and  $\sigma^2$  is the nuisance parameter. As an alternative we could have a sample from the gamma distribution with mean  $\mu$  and shape  $\beta$ , and be interested in  $\mu$  with  $\beta$  as the nuisance parameter.

For general notation let  $\theta = (\psi, \lambda)$  with real interest parameter  $\psi$  and with nuisance parameter  $\lambda$ . Thus if we are interested in the highest regression coefficient of the model  $y = X\beta + \sigma e$  with  $e$  as a sample from a known distribution then  $\psi = \beta_r$  and  $\lambda = (\beta_1, \dots, \beta_{r-1}, \sigma)$ . We consider a statistical model  $f(y; \theta)$  and let  $\ell(\theta) = \ell(\theta; y^0)$  be the observed likelihood for the full parameter  $\theta$ .

The procedures and directions surveyed in this section build on those described in the preceding section. For various model situations we seek an appropriate likelihood  $\ell(\psi)$  for the component parameter and an appropriate reparameterization  $\phi = \phi(\psi)$  of that component parameter. The observed significance function  $p(\psi)$  is then obtained by the formulas in Section 4.2 or by the numerical program mentioned in Section 5.

### 6.1 Exponential model case

Consider an exponential model with canonical parameter  $\theta = (\psi, \lambda)$  and with canonical interest parameter  $\psi$ :

$$\exp\{\psi t_1 + \lambda' t_2 - c(\psi, \lambda) + h(y)\} dy .$$

The minimal sufficient statistic is  $t = (t_1, t_2)$ .

For inference concerning  $\psi$  the conditional distribution  $t_1|t_2$  is the largest range conditional distribution that is free of  $\lambda$ , based on the sufficient statistic. An  $O(n^{-\frac{3}{2}})$  approximation (Cox and Reid, 1987, Fraser and Reid, 1990) to the likelihood from this conditional distribution is given by

$$\ell(\psi) = \ell(\psi, \hat{\lambda}_\psi) + \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|$$

where  $\hat{\lambda}_\psi$  is the maximum likelihood estimate of  $\lambda$  for given  $\psi$  and  $j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$  is the observed information concerning  $\lambda$  for fixed  $\psi$ ; note that  $\ell(\psi, \hat{\lambda}_\psi)$  is the profile likelihood.

The theory in Fraser and Reid (1990) then shows that using  $\ell(\psi)$  with the identity reparameterization  $\phi(\psi) = \psi$  in the formula of Section 4.2 gives an  $O(n^{-\frac{3}{2}})$  approximation to the significance function  $p(\psi)$  based on the conditional distribution of  $t_1|t_2$ .

## 6.2 Transformation model case

A simple example of a transformation model is obtained with a sample from the location model  $f(y_1 - \psi, y_2 - \lambda)$  where  $y_2$  and  $\lambda$  can be vector valued. Let  $\ell(\psi, \lambda)$  be the likelihood from an observed sample.

Basic theory for such models leads to a consideration of the conditional model of say  $(\bar{y}_1, \bar{y}_2)$  given the residuals  $\{(y_{1i} - \bar{y}_1, y_{2i} - \bar{y}_2) : i = 1, \dots, n\}$  for the full data. The appropriate distribution for inference concerning  $\psi$  is then obtained by marginalization to the variable  $\bar{y}_1$ ; it has a distribution that is free of  $\lambda$ .

The likelihood from the conditional distribution can be approximated (Fraser and Reid, 1990) to accuracy  $O(n^{-\frac{3}{2}})$  by

$$\ell(\psi) = \ell(\psi, \hat{\lambda}_\psi) - \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|.$$

The appropriate reparameterization to use for the conversion program is  $\phi(\psi) = -S(\psi)$  where  $S(\psi)$  is the score calculated from the approximate likelihood  $\ell(\psi)$ . This is easily handled by the general conversion program.

The sign here for the information adjustment in  $\ell(\psi)$  is opposite to that in the preceding Section 6.1. This is not a major difference as the parameterization for the two models is in some sense opposite and can be shown to account fully for the sign change.

The general transformation model requires some additional considerations relating to the choice of support measure and will not be discussed here.

## 6.3 More general models

The development of component likelihood  $\ell(\psi)$  and component reparameterization  $\phi(\psi)$  for more general models is we believe at the core of statistical inference. We cite here the results for the case where the variable and the parameter have the same dimension; the details and some extensions may be found in Fraser and Reid (1989, 1990). For this let  $d\hat{\eta} = j^{\frac{1}{2}} d\hat{\theta}$  be locally defined constant information coordinates.

An approximate  $O(n^{-1})$  conditional likelihood for  $\psi$  (Fraser and Reid, 1990) is

$$\ell(\psi) = \ell(\psi, \hat{\lambda}_\psi) - \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| - \log |\partial \hat{\lambda}_\psi / \partial \hat{\eta}|$$

where the Jacobian ‘determinant’ is the volume of the row vector array in the matrix. The reparameterization is

$$\phi(\psi) = k_1^{-1}(y^0) \dot{\ell}(\psi, \hat{\lambda}_\psi)$$

where  $\dot{\ell}(\theta) = \partial \ell(\theta; y) / \partial y$  and  $k_1^{-1}(y)$  is the first row of the inverse matrix to

$$k(y) = \frac{\partial^2}{\partial y \partial \theta'} \ell(\theta; y) |_{\hat{\theta}}.$$

This with the numerical conversion procedure leads to an  $O(n^{-1})$  significance function  $p(\psi)$ .

#### 6.4 Gamma distribution example

Gross and Clark (1975) examine the survival time  $y$  of 20 radiated mice; a gamma model with mean  $\mu$  and shape  $\beta$  is indicated and the interest parameter is  $\mu$ . The data are

152, 152, 115, 109, 137, 88, 94, 77, 160, 165, 125, 40, 128, 123, 136, 101, 62, 153, 83, 69

A 95% confidence interval for  $\mu$  was obtained by a first order analysis of the maximum likelihood estimate

$$(96.7, 130.8). \tag{1}$$

Grice and Bain (1980) noted that the model using logarithms was a location model for  $\log \mu$ . The corresponding transformation analysis with an estimated nuisance parameter including a Monte Carlo adjustment for the estimation gave the 95% confidence interval for  $\mu$ ,

$$(107.7, 134.7). \tag{2}$$

Jensen (1986) noted that the model was exponential and that the mean was in correspondence with the ratio of two canonical parameters. The indicated analysis is an iterative conditional analysis and requires testing each value for  $\mu$  and determining acceptance or rejection. The analysis based on the approximation in Section 6.1 is  $O(n^{-\frac{3}{2}})$  but is iterative, that is each value for  $p(\psi)$  requires a conversion calculation as described in Sections 4.1 and 6. The 95% confidence interval for  $\mu$  is

$$(96.8, 133.5) . \tag{3}$$

The procedure described in this section is  $O(n^{-1})$  but requires just the conversion of a single  $\ell(\psi)$  using a reparameterization  $\phi(\psi)$  that is data dependent. The 95% confidence interval for  $\mu$  is

$$(97.2, 134.0) . \tag{4}$$

From our present viewpoint we take (3) to be the exact value, and (4) to be an  $O(n^{-1})$  approximation using the simple numerical conversion program described in Section 5.

Figure 7 records the significance function  $p(\psi)$  as obtained by the first order approximations using the maximum likelihood estimate, the score, and the likelihood ratio. The iterative third order approximation and the second order single pass procedure are almost identical and coincide to the accuracy of the drawing. Among the first order methods we note again that the likelihood ratio seems better. Confidence intervals at level 95% are recorded beneath the plot in Figure 7.

## 7. CONCLUDING REMARKS

For a real parameter  $\theta$  we have argued that the conversion of an observed likelihood function  $\ell(\theta)$  to the observed significance function  $p(\theta)$  is central to statistical inference and can be accomplished by a generic computer program to order  $O(n^{-\frac{3}{2}})$ . With nuisance parameters separating in a canonical manner for exponential or transformation models a

similar result holds in terms of an easily calculated reparameterization  $\phi(\psi)$  of the interest parameter. Some extensions to more general models are surveyed in Chapter 6.

## ACKNOWLEDGMENTS

This paper is based on the R.A. Fisher lecture given in Anaheim, California August 8, 1990. I would like to thank N. Reid and A. Wong for invaluable contributions to the development of the material. The lecture was presented using a Sony graphics projector driven by a Mac IIci computer, and assistance from the University of Toronto Media Centre is gratefully acknowledged.

## REFERENCES

- Barndorff-Nielsen, O.E., Cox, D.R. (1979). Edgeworth and saddlepoint approximations with statistical applications. *J. Royal Statist. Soc. B* **41**, 279–312.
- Cox, D.R., Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Royal Statist. Soc. B* **49**, 1–39.
- Daniels, H.E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25**, 631–650.
- Daniels, H.E. (1987). Tail probability approximation. *Int. Statist. Rev.* **55**, 37–48.
- DiCiccio, T.J., Field, C.A., Fraser, D.A.S. (1990). Approximation of marginal tail probabilities and inference for scalar parameter, *Biometrika* **77**, 77–96.
- Fisher, R.A. (1925). *Statistical Methods for Research Workers*. London: Oliver and Boyd.
- Fisher, R.A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22**, 700–725.
- Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. London: Oliver and Boyd.
- Fraser, D.A.S. (1988). Normed likelihood as saddlepoint approximation. *J. Mult. Anal.* **27**, 181–193.
- Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77**, 65–76.
- Fraser, D.A.S., Lee, H.S., Reid, N. (1990). Nonnormal linear regression: an example of significance levels in high dimensions. *Biometrika* **77**, 333–342.

- Fraser, D.A.S., Reid, N. (1989). Adjustments to profile likelihood. *Biometrika* **76**, 479–488.
- Fraser, D.A.S., Reid, N. (1990). Simple asymptotic connections between density and cumulant functions leading to accurate approximations for distribution functions, submitted for publication.
- Fraser, D.A.S., Reid, N., Wong, A. (1991). Exponential linear models: a two pass procedure in saddlepoint approximation. *J. Royal Statist. Soc. B* **52**, to appear.
- Grice, J.V., Bain, L.J. (1980). Inferences concerning the mean of the gamma distribution. *J. Amer. Statist. Assoc.* **75**, 929–933.
- Gross, A.J., Clark, V.A. (1975). *Survival Distributions: Reliability Applications in the Biomedical Services*. New York: John Wiley.
- Jensen, J.L. (1986). Inference for the mean of a gamma distribution with unknown shape parameter. *Scand. J. Statist.* **13**, 135–151.
- Lugannani, R., Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Prob.* **12**, 475–490.
- Reid, N. (1988). Saddlepoint methods and statistical inference. *Statist. Sci.* **3**, 213–237.
- Student (1908). The probable error of a mean. *Biometrika* **6**, 1–25.

**Figure 1.** Observed likelihood function  $L(\theta)$  for the Cushney Peebles data.

**Figure 2.** Observed log likelihood function  $\ell(\theta)$  for the Cushney Peebles data.

**Figure 3.** Observed significance function  $p(\theta)$  for the Cushney Peebles data.

**Figure 4.** The standardized maximum likelihood estimate, standardized score, and signed likelihood ratio produce three approximations for the significance function; the exact  $p(\theta)$ . Model: Binomial  $(14, p)$ ; Data:  $y^0 = 3$ .

**Figure 5.** The standardized maximum likelihood estimate, standardized score, and signed likelihood ratio produce three approximations for the significance function. Model: location log gamma  $(3)$ ; Data:  $y^0 = 3.14$ .

**Figure 6.** (a) The likelihood function for data  $y = 3.14$  and for  $y = 3.24$ . (b) The likelihood difference under data point change from 3.14 to 3.24.

**Figure 8.** For the Gross and Clark data, the significance function  $p(\mu)$  based on the maximum likelihood estimate, the score, the signed likelihood ratio, and the iterative third order and single pass second order which coincide to the accuracy of the diagram.