Fundamental Principle(s) of Statistics

For a multiparameter model write $L(\theta; y)$ for the likelihood function, and $l(\theta; y)$ for its (natural) logarithm. Consider a hypothesis that can be thought of as constraining the model to satisfy $g_j(\theta) = 0,\ \ j = 1, \cdots, q$. Let us write $\theta \in H$ to represent this hypothesis, and define the loglikelihood drop as

$$d(y) = \max_{\theta} l(\theta; y) - \max_{\theta \in H} l(\theta; y)$$

Under regularity conditions most importantly involving that these constraint functions are smooth, the approximate distribution of $W(y) = 2d(y)$ is chi-square on q d.f. when the hypothesis H is true. That is, in a cumulative manner each "unnecessary" parameter introduced into the valid model H increases the maximum loglikelihood by one-half a chi-square variate on one d.f. This is called Wilks' Theorem, and the approximation involved is usually excellent. I believe it should be referred to as the *Fundamental Principle of Inference.*

There are many corollaries to this, and related matters of inference. Many of these involve thinking of the setting slightly differently, in terms of a q-dimensional *interest parameter* $\psi = \psi(\theta)$ defined in the full model (without restriction to H). Then the *profile likelihood* function for this parameter is defined as

$$L_P(\psi; y) = \max_{\theta:\, \psi(\theta) = \psi} L(\theta; y)$$

Then Wilks Theorem becomes that when the true value of the interest parameter takes the value ψ, the quantity $W(y) = 2\{L_P(\hat{\psi}; y) - L_P(\psi; y)\}$, where $\hat{\psi}$ is the MLE, has approximately a chi-square distribution on $\dim(\psi)$ d.f. This generalizes to profile likelihood a fundamental result for ordinary likelihood, both of which follow from Wilk's Theorem. The implication of this is that the parameter region

$$\{\psi :\ L_P(\psi; y) / L_P(\hat{\psi}; y) > \exp(-\text{chi-square}_{\text{q,1-}\alpha})$$

is an approximate $1 - \alpha$ level confidence region.for ψ.

These matters are especially useful when $\psi = \psi(\theta)$ is one-dimensional, i.e. $q = 1$. Then *W(y)* defined above is approximately chi-squared on 1 df when the hypothesis is true, and $r(y) = sign(\hat{\psi} - \psi) W(y)^{1/2}$ has approximately a standard normal distribution.

This provides a *directional* test of the hypothesis that ψ is the true value of $\psi(\theta)$. Further, the profile likelihood is for the scalar parameter ψ, and the interval (usually) where $L_P(\psi; y)$ is at least any given multiple of the maximum value is an approximate confidence interval at the associated level of confidence.

This should be compared to the Wald confidence interval given based on the MLE and its approximate standard error. The primary limitation of this method is that the result is not invariant under monotonic transformations of the parameter ψ. Thus the Wald method can be good or poor, in terms of coverage probability, depending on the rather arbitrary representation of the parameter. The likelihood ratio method does not have this defect, and usually gives results that are about as good as when the scale for ψ is chosen carefully to make the normal approximation to the distribution of the MLE adequate.