A Child's Garden of Likelihood

Many approaches to statistical inference begin with an estimate motivated by intuition (e.g. the sample proportion) or by some essentially ad hoc principle, like least squares. It is then customary to use probability theory based on the assumption of random sampling to assess reliability. By contrast, likelihood methods make use of assumptions about random sampling to not only provide measures of reliability in the form of confidence and significance levels, but to provide the estimates in the first place.

One Parameter Problems

In the simplest situation one wishes to make inference regarding a single unknown parameter, θ, based on a sample of observations, $y_1, y_2, \ldots, y_n$.

e.g. Let θ represent the concentration of cells (say in cells/l) of some organism in some preparation. Counting cells directly is often impractical, so estimates are sometimes based on a "dilution series". In a dilution series, the initial preparation is split into smaller "aliquots" which are then cultured separately, to detect if **any** cells are present in the aliquot. The result for a particular aliquot is then either positive or negative, i.e. a binary outcome.

The sampling distribution of the data

If we assume "random mixing" in the dilution processes (analogous to sampling variability in a true sampling experiment) the distribution of the z_i=number of cells in aliquot i will be Poisson. (Recall the the probability function for a Poisson variate z, with $\mu=E(z)$, is $\Pr(z;\mu) = \frac{\mu^z e^{-\mu}}{z!}$). In the present experiment, the expected values for z_i depend on the dilution factor, and will satisfy $\mu_i=\theta \times x_i$, where x_i represents the fractional size of the i^{th} aliquot relative to the initial preparation.

Suppose that we record y_i=0 (negative) or 1 (positive), i=1,...,n and that the size of the aliquots cultured are x_i, i=1,...,n. The y_i's can be thought of as random variables, with $\pi_i=\Pr\{y_i=1\}$ depending on θ and x_i. The probability that y_i equals 0 is just the probability that z_i equals 0, which based on the Poisson formula is simply $e^{-\mu_i}$. Thus we have

$$\Pr\{y_i=0\} = e^{-\theta x_i} \text{ and } \pi_i=\Pr\{y_i=1\}=1-e^{-\theta x_i}.$$

For any given value of θ, probabilities for a particular set of outcomes y_i, i=1,..n can be worked out by calculating and multiplying the relevant probabilities from the above.

The likelihood for θ and the maximum likelihood estimate, $\hat{\theta}$

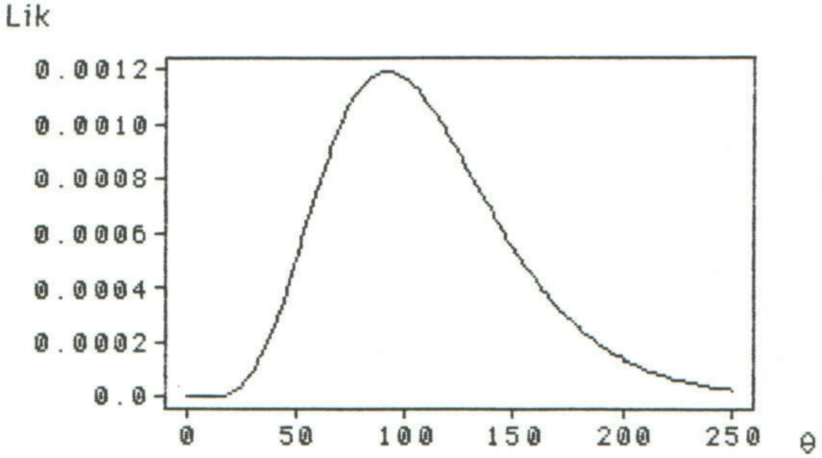
The above expression provides probabilities for arbitary outcomes as they depend on the unknown value of the parameter. Having observed particular outcomes, however, it can be viewed as providing a vehicle for inference concerning that parameter.

Suppose that ten 10 ml aliquots were obtained (i.e. x_i all = .01) and that 6 of the aliquots cultured positive. This gives rise to the following likelihood

$$\text{Lik}(\theta\ ;\ y_1,y_2,\ ...,\ y_n) = (e^{-\theta/100})^4 \times (1-e^{-\theta/100})^6$$

Note that in this example we don't need to know the individual y_i's, since the likelihood only depends on their total, T (in this case T=6). Here T is called a sufficient statistic, since it's all we need to form the likelihood.

This likelihood just tabulates the probabilities for the observed data, as they depend on hypothetical values of the unknown parameter. The graph on the top of the next page shows the likelihood for our example.



Based on the graph we see that the probability of our observed outcome is maximized at θ= 92, roughly. This is a plausible best estimate for θ. In general, the value of θ, denoted by $\hat{\theta}$, that maximizes the probability of the observed data (as determined by the sampling distribution postulated for the data) is called the maximum likelihood estimate. The exact value for the above data is $\hat{\theta}$=91.6.

In addition, the curve of probabilities below may also be view as a plausible basis for "ruling out" certain values of θ providing probabilities that are "too low". By formalizing this intuitive notion, we can derive likelihood based significant tests and confidence intervals.

Likelihood based inference procedures

The development above has only brought us to the point of determining a point estimate, namely the m.l.e. Tests and confidence intervals can be derived by considering the sampling distribution of the m.l.e. In most situations, the exact distribution of the m.l.e. will be intractable (too mathematically complicated), so we resort to approximate results. A particularly nice feature of likelihood inference, is that in general, when sample sizes are large, the sampling distribution of the m.l.e. will be approximately normal, with mean θ. This implies that m.l.e.'s are approximately unbiased estimates of their "target" parameters.

Likelihood theory also provides estimated standard errors based on the "peakedness" of the likelihood at the value of the m.l.e. A numerical measure of this peakedness is the negative of the 2nd derivative of the logarithm of the likelihood, which is referred to as the Fisher information, and sometimes denoted $I(\hat{\theta})$. An approximate standard error for $\hat{\theta}$ is $\frac{1}{\sqrt{I(\hat{\theta})}}$ (note that this is small when the information is large).

Significance tests can be based on the z-statistic,

$$z = \frac{\hat{\theta}-\theta_0}{se(\hat{\theta})} = \sqrt{I(\hat{\theta})}\left\{\hat{\theta}-\theta_0\right\},$$

and confidence intervals take the usual form,

$$\hat{\theta} \pm (\text{critical z-value})\ se(\hat{\theta}).$$

In our example, the Fisher information takes the form

$$I(\hat{\theta}) = 10^{-3} \times (n-T)/T,$$

which for the data at hand is .0007, yielding an estimated standard error for our estimate of 38.7

The log likelihood and the likelihood ratio test

Because the likelihood is typically formed as a product, it is often convenient to take is logarithm (base e), converting the product to a sum. In the example above, the log likelihood, $l(\theta\ ;\ y_1,y_2,\ldots,y_n)$ takes the general form (after omitting irrelevant terms)

$$l(\theta\ ;\ y_1,y_2,\ldots,y_n) = (n-T)\times(-\theta/100) + T\times\log_e(1-e^{-\theta/100})$$

An alternative method for approximate inference which has been shown to be more reliable than that based on the asymptotic normality of the m.l.e. arises from the fact that the quantity

$$\Delta = -2 \times \left\{l(\theta;y_1,y_2,\ldots,y_n) - l(\hat{\theta};y_1,y_2,\ldots,y_n)\right\}$$

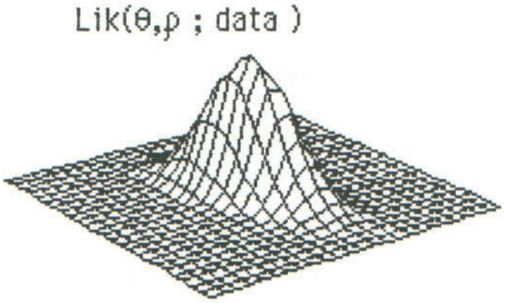
has a distribution which is approximately $\chi^2_{(1)}$ in large samples. Note that this quantity is not observable since it depends on the "true" value of θ. By replacing θ with some hypothetical value, θ_0, one arrives at a test statistic, D, which can be referred to $\chi^2_{(1)}$ tables. This test is sometimes referred to as the likelihood ratio test, because D is essentially the logarithm of the ratio of the likelihood evaluated θ_0 and $\hat{\theta}$.

The deviance that GLIM prints is essentially $- 2 \times \left\{ l(\hat{\theta};y_1,y_2,\ldots,y_n) \right\}$. Thus the change in deviance test is really a likelihood ratio test.

Multi-parameter Problems

Much of the discussion above generalizes directly to the case when there is more than one parameter of interest. In principle, there is no additional complication in allowing the sampling distribution of the data to depend on more than one quantity. For example, suppose we have solutions 1 and 2, with cell counts, $\theta_1=\theta$ and $\theta_2=\rho\theta$, to determine. (The "practical" interpretation of ρ is that it is the ratio of the counts). If we have results y_{1i}, $i=1,\ldots,n_1$ and y_{2i}, $i=1,\ldots,n_2$, from the two solutions respectively, with corresponding concentrations, x_{1i}, $i=1,\ldots,n_1$ and x_{2i}, $i=1,\ldots,n_2$ then the joint probability of the data can be formed as a product, leading to the joint likelihood, which we'll denote Lik(θ,ρ ; data).

To determine the m.l.e.'s, $\hat{\theta}$ and $\hat{\rho}$, we must maximize the value on the likelihood surface, above the plane of possible θ and ρ values. A typical likelihood surface looks like:

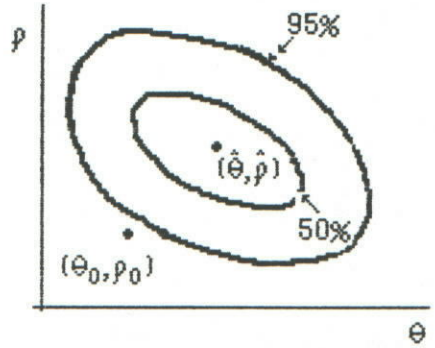


In this example, as in most practical example, $\hat{\theta}$ and $\hat{\rho}$ are asymptotically normal, with means corresponding to their "target" parameters. Approximate standard errors are obtainable by consideration of the shape of the likelihood surface at $(\hat{\theta},\hat{\rho})$, and give rise to tests and C.I.'s for θ and ρ considered separately.

An altogether new issue that arises is that of joint inference for the two parameters. Joint hypothesis tests pertain to hypotheses of the form

$$H_0\text{: } \theta=\theta_0 \ \& \ \rho=\rho_0 \quad \text{vs} \quad H_a\text{: } \theta\neq\theta_0 \ \& \ \rho\neq\rho_0$$

giving rise to significance levels for assessing the evidence against H_0. A joint confidence region is a region of plausible (θ,ρ) values in the (θ,ρ) plane, as determined from the data. Joint confidence regions are typically (ideally) ellipsoidal in shape.

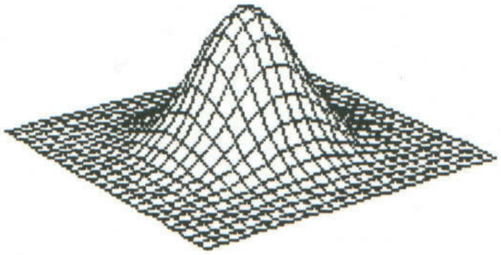


Joint inference procedures differ from separate inference procedures in some non-intuitive ways, arising out of certain features of sampling variation. One key feature that is sometimes present is correlation between the m.l.e.'s, $\hat{\theta}$ and $\hat{\rho}$. The practical importance of correlation arises from noting the the correlation of the errors of estimation, $\hat{\theta}-\theta$ and $\hat{\rho}-\rho$, is the same as that of $\hat{\theta}$ and $\hat{\rho}$. Thus if $\hat{\theta}$ and $\hat{\rho}$ are postively correlated, it is the case that values of $\hat{\theta}$ that overestimate θ will tend to be associated with values of $\hat{\rho}$ that overestimate ρ. The oblique nature of the confidence region in the plot above arises out of negative correlation.

Correlated estimates are also responsible for certain "paradoxes" in joint testing. For example, it is not uncommon to fail to have evidence against either of the separate null hypotheses, H_0:$\theta=\theta_0$ or

H_0:$\rho=\rho_0$, but to be able to reject the joint null, H_0: $\theta=\theta_0$ & $\rho=\rho_0$. This situation is best understood by reference to the corresponding confidence region for θ and ρ. Again the plot above is relevant. Once sees that the confidence region is not able to "rule out" all points with $\theta=\theta_0$ or $\rho=\rho_0$, but does rule out the point (θ_0,ρ_0).

The mathematical treatment of correlation is most easy to deal with in relation to the multivariate normal distribution, which is the multi-dimensional generalization of the normal distribution. For example, a bivariate normal distribution is plotted below. Large sample theory shows that the joint distribution of a set of m.l.e.'s will tend to be multivariate normal.



When a set of estimates has a multivariate normal distribution, their separate (marginal) behaviour is normal, with their respective means and standard errors, but their joint behavior is determined as well by their correlations, which can be specified in a correlation matrix. The likelihood surface provides an estimated correlation matrix just as it provides estimated standard errors, based on the geometry of the surface around the m.l.e. Joint hypothesis tests and confidence intervals can be deduced from this information.

In addition to inference based on asymptotic multivariate normality, the likelihood ratio test has analogues in the multiparameter situation. First we take logs, yielding

$$l(\theta,\rho\ ;\ \text{data}) = \log\left\{\text{Lik}(\theta,\rho;\text{data})\right\}.$$

We can then define the quantity

$$\Delta= -2\times \left\{ l(\theta,\rho;\text{data})- l(\hat{\theta},\hat{\rho};\text{data})\right\}$$

which has sampling distribution which is $\chi^2_{(2)}$. By substituting θ_0 and ρ_0 values, one obtains a test statistic for H_0: $\theta=\theta_0$ & $\rho=\rho_0$.