**http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm**

**Lecture 8—Wednesday, September 22, 2010**

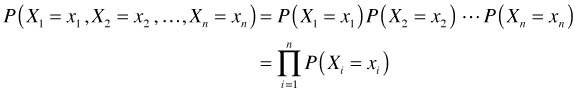
**Topics**

* [Review of likelihood theory](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#review)
* [The information matrix](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#information)
* [Properties of maximum likelihood estimators](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#properties)
  + [Some not so nice properties](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#notnice)
  + [A few of the nice properties](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#nice)
* [Interpreting the information](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#interpreting)
* [Likelihood ratio test](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#LRtest)
* [Wald test](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#waldtest)
* [Constructing confidence intervals for MLEs](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#constructing)
  + [Wald confidence intervals](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#waldci)
  + [Profile likelihood confidence intervals](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#profileci)
  + [Profile likelihood confidence intervals using the Bhat package](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#Bhat)
* [Books and articles on likelihood](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#books)
* [Some web references on likelihood](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#web)
* [R code](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#Rcode)

**Review of likelihood theory**

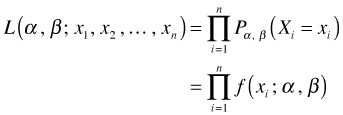
**The probability of obtaining a random sample**

Let random sampledenote a random sample of size *n*. We wish to compute the probability of obtaining this particular sample for different probability models in order to help us choose a model. Because the observations in a random sample are independent we can write the generic expression for the probability of obtaining this particular sample as follows.



**The likelihood**

The next step is to propose a particular probability model for our data. Letprob model denote this probability model where the notation is meant to indicate that the model requires the specification of two parameters α and β. We can replace the generic probability terms in the above expression with the proposed model.



The change in notation on the left hand side is deliberate. The new expression for the probability when treated as a function of α and β is referred to as the likelihood, *L*, of our data under the proposed probability model.

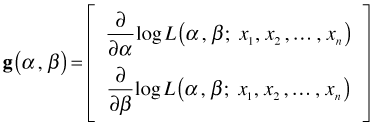
**The log-likelihood**

Typically we work with the log of this expression now called the log-likelihood of our data under the proposed probability model.

log-likelihood

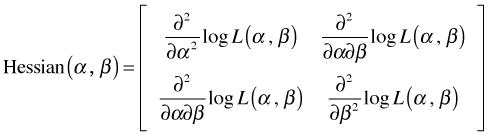
**The score (gradient) vector**

The maximum likelihood estimates (MLEs) of α and β are those values that make the log-likelihood (and hence the likelihood) as large as possible. The MLEs maximize the likelihood of obtaining the data we obtained. Calculus is used for finding MLEs. The analytical protocol would involve taking the derivative of the log-likelihood with respect to each of the parameters separately, setting the derivatives equal to zero, and solving for the parameters (algebraically when possible, numerically if not). The derivative of the log-likelihood is called the score or gradient vector.



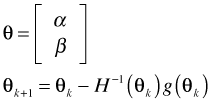
**The Hessian**

The matrix of second partial derivatives of the log-likelihood is called the Hessian matrix.



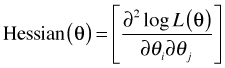
**Newton-Raphson method**

One method for obtaining maximum likelihood estimates is the [Newton-Raphson method](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture6.htm#Newton). It is an iterative method that makes use of both the gradient vector and the Hessian matrix to update the estimates of the parameters at each step of the algorithm.



**The information matrix**

We've already defined the score function as being the first derivative of the log-likelihood. If there is more than one parameter so that **θ** is a vector of parameters, then we speak of the score vector whose components are the first partial derivatives of the log-likelihood. The matrix of second partial derivatives of the log-likelihood is called the **Hessian matrix**.

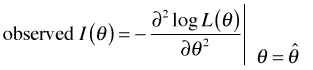


If there is only a single parameter *θ*, then the Hessian is a scalar function.

hessian

The **information matrix**, I(θ), is an important quantity in likelihood theory. It is defined in terms of the Hessian and comes in two versions: the **observed information** and the **expected information**.

1. The observed information is just the negative of the Hessian evaluated at the maximum likelihood estimate.



2. The expected information is the expected value of the negative of the Hessian, i.e., the mean of the sampling distribution of the negative Hessian,

expected information

which is then evaluated at the maximum likelihood estimate. Statistical software (for example, the **nlm** function in R) typically returns the Hessian evaluated at the MLE yielding what we're calling the observed information. Note: Because **nlm** does minimization rather than maximization we are forced to use the negative log-likelihood as the objective function. Since the negative sign has already been introduced as part of the log-likelihood, the Hessian produced by **nlm** already has the negative sign we require to obtain the information. So, in terms of likelihood theory, **nlm** actually returns minus hessian.

**Properties of maximum likelihood estimators (MLEs)**

The near universal popularity of maximum likelihood estimation derives from the fact that the estimates it produces have good properties. Nearly all of the properties of maximum likelihood estimators are **asymptotic**, i.e., they only kick in after sample size is sufficiently large. For small samples there are no guarantees that these properties hold and what constitutes "large" will vary on a case by case basis. In what follows I use the notation theta hat nas a shortcut for "the maximum likelihood estimate of θ based on a sample of size *n*."

**Some of the not so nice properties of MLEs**

1. Maximum likelihood estimators are often biased (see [below](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#asymptoticallyunbiased)).
2. Maximum likelihood estimators need not be unique.
3. Maximum likelihood estimators may not exist.
4. Maximum likelihood estimators can be difficult to derive analytically. In all but the simplest cases they need to be approximated numerically. Numerical methods can be very sensitive to initial guesses for the parameter estimates and may fail to converge.
5. Testing that the critical point corresponds to a maximum can be painful even for simple scenarios. The possibility of obtaining local maxima rather than global maxima is quite real.

**A few of the nice properties of MLEs**

This is an abbreviated list because many of the properties of MLEs will not make sense if you don't have the appropriate background in statistical theory. Even some of the ones I list here may seem puzzling to you. The most important properties for practitioners are numbers four and five that give the asymptotic variance and the asymptotic distribution of maximum likelihood estimators.

1. theta hat nis a **consistent estimator** of θ. This means

consistency

Thus the maximum likelihood estimate approaches the population value as sample size increases.

1. theta hat nis **asymptotically unbiased**, i.e., unbiased. So, for large samples the sampling distribution of an MLE is centered on the true population value. Maximum likelihood estimators can be biased, but the bias disappears as the sample size increases. As an example, if x1 xnis a random sample from a normal distribution with mean μ and variance sigma squared, the maximum likelihood estimator of sigma squaredis

sigma hat

This estimator is biased, which is why we typically use the sample variance as an estimator instead because it is unbiased.

s2

Clearly the difference between these two estimators becomes insignificant as *n* gets large.

1. theta hat nis **asymptotically efficient**, i.e., among all asymptotically unbiased estimators it has the minimum variance asymptotically. In other words, maximum likelihood estimators tend to be the most precise estimators possible.
2. The variance of theta hat nis known (at least asymptotically). For *n* large,

http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/varianceofthetathatn.gif

where inverse informationis the inverse of the information matrix (for a sample of size *n*). I discuss the information matrix further in the next section. This is an important result because it means that the standard error of a maximum likelihood estimator can be calculated.

1. theta hat nis asymptotically normally distributed. So we know what the sampling distribution of a maximum likelihood estimator looks like, at least for large *n*. This provides the basis for hypothesis testing and the construction of confidence intervals.

Properties 2, 4, and 5 together tell us that for large samples the maximum likelihood estimator theta hat nof a population parameter *θ* has an approximate normal distribution with mean *θ* and variance that can be estimated as asymptotic variance.

asymptotic distribution

In linear regression we assume normal distributionwhere mu regression. If we use maximum likelihood instead of least squares to obtain estimates of β0 and β1 we find that the estimates are identical to the ones we would have obtained using least squares. This happens because the function containing the regression parameters (the sum of squared errors) that is then mimimized in ordinary least squares also appears in exactly the same form in the log-likelihood function for this problem. The upstart is that if we use least squares to find the parameter estimates in a regression problem and then assume the errors are normally distributed, or we begin instead by assuming that the response is normally distributed and then use maximum likelihood to estimate the parameters in the regression equation for the mean, we obtain the same results.

On the other hand the maximum likelihood estimate of sigma squaredis not the same as the one that is typically used in ordinary regression. Just as with the MLE of the sample variance described [above](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#asymptoticallyunbiased), the maximum likelihood estimate of sigma squaredin regression is biased, but the bias does diminish with sample size.

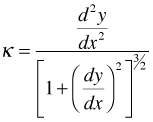
Likelihood theory is one of the few places in statistics where Bayesians and frequentists are in agreement. Both believe that the likelihood is fundamental to statistical inference. They diverge in that frequentists focus solely on the likelihood and use it obtain maximum likelihood estimates of the parameters, while a Bayesian uses it to construct something called the posterior distribution of the parameters.

**Interpreting the information**

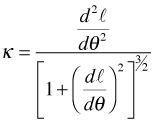
You may have been introduced to the notion of **curvature**, κ, in your calculus class. The formal definition of the curvature of a curve is the following.

curvature

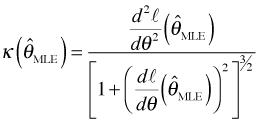
Here φ is the angle the tangent line makes with the curve and *s* is arc length. Thus curvature is the rate at which you turn (in radians per unit distance) as you walk along the curve. For a function written in the formy=f(x), its curvature can be calculated as follows.



What happens if we apply the curvature formula to the log-likelihood, log likelihood? If the log-likelihood is a function of a single scalar parameter θ, then we have



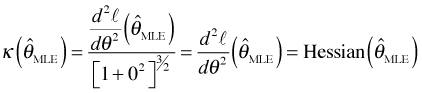
Now suppose we evaluate the curvature at the maximum likelihood estimate, mle theta.



Recall how the MLE is obtained. We differentiate the log-likelihood and set the derivative equal to zero. Thus, mle thetais the value of θ at which the score is zero, i.e.,

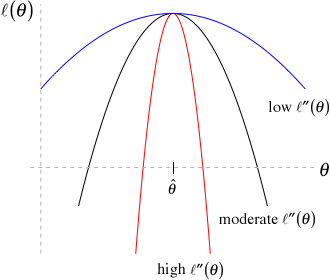
score equals zero

Using this result in the curvature equation above we obtain the following.



Recall that except for the sign, the Hessian (scalar version) at the MLE is what is called the observed information. At the global maximum of a function the second derivative is required to be negative, so taking the negative of the Hessian is just a way of ensuring that the observed information is non-negative. Equivalently we could have taken the absolute value. Thus the observed information is just the magnitude of the curvature of the log-likelihood when the curvature is evaluated at the MLE.

information and curvature



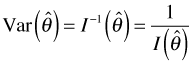
**Fig. 1** Curvature and information

A similar argument can be made for a multivariate log-likelihood except that we have multiple directions (corresponding to curves obtained by taking different vertical sections of the log-likelihood surface) to consider when discussing curvature.

So what do these ideas tell us about information in the sense used in likelihood theory? Consider the the graph in Fig. 1 in which three different log-likelihoods are shown. Observe that the three log-likelihoods are all functions of a single parameter θ, they are all maximized at the same place, mle theta, but they have very different curvatures. From red to black to blue we go from high curvature to moderate curvature to low curvature at the maximum likelihood estimate mle theta(the value of θ corresponding to the peak of the log-likelihoods). Using the diagram we can make the following observations.

1. Low curvature (**blue** curve) translates into a fairly flat log-likelihood. Thus in a neighborhood of mle thetamost values of θ yield roughly the same log-likelihood value and hence the log-likelihood is not useful in discriminating one θ from another. In other words we have low information about the true value of θ.
2. High curvature (**red** curve) translates into a rapidly changing log-likelihood. Thus in a neighborhood of mle thetaeven those values of θ that differ from mle thetaby a small amount have very different log-likelihoods and hence are readily distinguishable from one another. In this case we have a lot of information about the true value of θ.

A similar set of statements can be made about the variance of the estimator because for scalar θ the information and the variance are reciprocals of each other.



Using the relationship between information and the variance, we can draw the following conclusions.

* Low information translates into a high variance of our estimator. Hence confidence intervals for θ will be wide.
* High information translates into a low variance of our estimator. Hence confidence intervals for θ will be narrow.

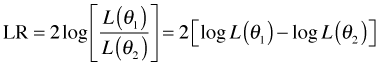
The table below summarizes these results more succinctly.

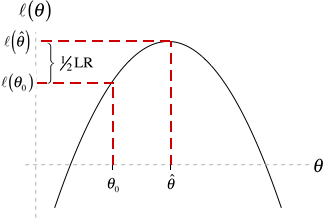
|  |  |  |  |
| --- | --- | --- | --- |
| Curvature | Information | var of theta hat | Confidence interval for θ |
| high | high | low | narrow |
| low | low | high | wide |

**Likelihood Ratio Test**

The **likelihood ratio (LR) test** is to likelihood analysis as ANOVA (more properly partial F-tests) is to ordinary linear regression. Partial F-tests are used to compare nested ordinary regression models; likelihood ratio tests are used to compare nested models that were fit using maximum likelihood estimation. Nested models are those that share the same probability generating mechanism, the same response, and all of the same parameters. The only difference is that in one of the models one or more of the parameters are set to specific values (usually zero), while in the other model those same parameters are estimated instead. So we have a restricted model (one in which some parameters are set to specific values) and a second less restricted model (one in which these same parameters are estimated).

Let θ1 denote the set of estimated parameters from the less restricted model and θ2 denote the set of estimated parameters for the restricted model. Because the models are nested the parameter set θ2 is a subset of the parameter set θ1. The likelihood ratio test takes the following form.

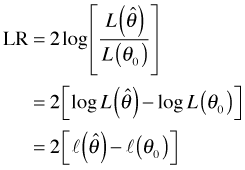




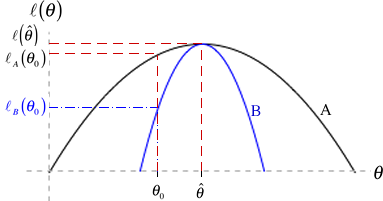
**Fig. 2** Geometry of the LR test

It turns out LR distributionwhere the degrees of freedom *p* is the difference in the number of estimated parameters in the two models. (The chi-squared distribution is a special type of gamma distribution in which gamma betagamma alpha.)

Let's consider the special case when there is only a single parameter θ and the restricted model specifies a value for this parameter that we'll denote θ0. (A common case would be θ0 = 0.) Then we have



where to simplify notation I use script l of thetato denote log(L(theta)). Fig. 2 illustrates the geometry of the LR test. The LR statistic maps closeness on the θ-axis into closeness on the log-likelihood axis (through the mapping given by the log-likelihood function). In the LR test two values of θ are deemed close only if their log-likelihoods are also found to be close. The chi-squared distribution provides the absolute scale for measuring closeness on the log-likelihood axis.

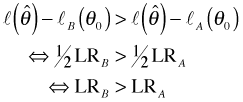


**Fig. 3** LR test when the information varies

The role the log-likelihood curve plays in this becomes even clearer when we compare two different log-likelihoods, perhaps arising from different data sets, that yield the same maximum likelihood estimate of θ, but have different curvatures. Fig. 3 illustrates two such log-likelihoods. If we are interested in testing

null hypothesis

then scenario B gives us far more information for rejecting the null hypothesis than does scenario A. Observe from Fig. 3 that



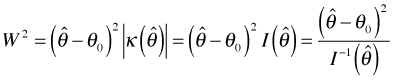
So even though the distance theta distanceis the same in both cases, the distances on the log-likelihood scale are different due to their different curvatures. Using the LR test we are far more likely to reject the null hypothesis under scenario B than under scenario A because scenario B provides more information about θ.

**Wald Test**

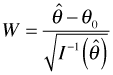
The **Wald test** is an alternative to the likelihood ratio test that attempts to incorporate the distance http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/thetadistance.gifmore directly into a test of the null hypothesis

null hypothesis

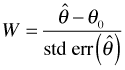
As we've seen in Fig. 3 distance on the θ-axis is not enough. We also need to take into account the curvature of the log-likelihood. In Fig. 3 the more informative scenario B is the one with the greater curvature. In the Wald test we weight the distance on the θ-axis by the curvature of the log-likelihood curves. Formally, the Wald statistic, *W*, is the following.



where in the second inequality I make use of the relationship between curvature and that was information described above. Taking square roots of both sides yields the Wald statistic.



Recall though that variance of thetahat. Thus the Wald statistic *W* can also be written as



Now under the null hypothesis theta naughtis the true value of θ. From likelihood theory we also know that asymptotically the MLE is unbiased for θ. So asymptotically, at least, if the null hypothesis is true then expected theta hat. Thus asymptotically the expression for *W* given above is a *z*-score: a statistic minus its mean divided by its standard error. Since we also know that the MLE of θ is asymptotically normally distributed, it follows that *W*, being a *z*-score, must have a standard normal distribution. This provides the basis for the Wald test as well as Wald confidence intervals. (Note: Another way of viewing the Wald test is that it locally approximates the log-likelihood surface with a quadratic function.)

**Constructing confidence intervals for MLEs**

I consider two different ways to construct confidence intervals for maximum likelihood estimates: Wald confidence intervals and profile likelihood (likelihood ratio) confidence intervals.

**Wald confidence intervals**

Wald confidence intervals are the easiest to construct. They depend on the asymptotic normality of the maximum likelihood estimator. A 95% confidence interval for any normally distributed parameter theta hattakes the following form.

http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/waldconf.gif

Here z .975is the 0.975 quantile of a standard normal distribution.

As was explained above, the standard error for a (scalar) maximum likelihood estimator can be obtained by taking the square root of the reciprocal of the negative of the Hessian evaluated at the MLE. (Note: In R we don't take minus the Hessian because we have already introduced the minus sign into this problem because we're working with the negative log-likelihood.) Both **optim** and **nlm** will return minus the Hessian evaluated at the MLE if requested. In what follows I use **nlm**.

With **nlm** we need to add the argument **hessian=TRUE**. I assign the output from **nlm** to an object I call **out** so that I can access the various components of the output directly. I return to the data we examined in [lecture 7](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture7.htm#aphid) to illustrate these ideas.

num.stems <- c(6,8,9,6,6,2,5,3,1,4)

**#generate raw data from tabulated values**

aphid.data <- rep(0:9,num.stems)

**#ML estimation for Poisson model**

poisson.LL <- function(lambda) sum(log(dpois(aphid.data, lambda)))

poisson.negloglik <- function(lambda) -poisson.LL(lambda)

out <- nlm(poisson.negloglik, 3, hessian=TRUE)

out

$minimum  
[1] 124.1764

$estimate  
[1] 3.459998

$gradient  
[1] 6.571497e-08

$hessian  
[,1]  
[1,] 14.44799

$code  
[1] 1

$iterations  
[1] 4

To access the elements of **out**, a list, I use list $ notation.

out$estimate

[1] 3.459998

out$hessian

[,1]  
[1,] 14.44799

The .975 quantile of a standard normal distribution is obtained with the **qnorm** function.

qnorm(.975)

[1] 1.959964

Finally I put all the pieces together.

**#lower bound**

out$estimate-qnorm(.975)\*sqrt(1/out$hessian)

[,1]  
[1,] 2.944361

**#upper bound**

out$estimate+qnorm(.975)\*sqrt(1/out$hessian)

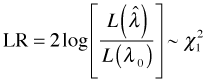
[,1]  
[1,] 3.975636

So our 95% Wald confidence interval rounded to two decimal places is (2.94, 3.98). We are 95% confident that the true value of the population parameter λ lies in this interval.

**Profile likelihood confidence intervals**

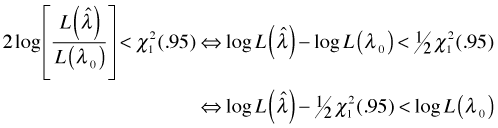
The profile likelihood confidence interval (also called the likelihood ratio confidence interval) derives from the asymptotic chi-squared distribution of the likelihood ratio statistic. It is generally considered to be more accurate than the Wald confidence interval when the sample size is small. The basic rationale for it is as follows.

Suppose *λ* is a scalar parameter and we wish to test whether lambda equal lambda naughtwhere http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/lambdanaught.gifis some specific value of interest. As we've seen the likelihood ratio statistic for this test is the following.



where http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/lambdahat.gifis the MLE of λ. We reject lambda equal lambda naughtat α = .05 if http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/LR%20reject.gifwhere chi square quantileis the .95 quantile of a chi-squared distribution with one degree of freedom. Similarly we would fail to reject lambda equal lambda naughtif it turns out http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/LRaccept.gif. This last inequality can be viewed as defining all those values of http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/lambdanaught.giffor which we would fail to reject the null hypothesis. By failing to reject the null hypothesis for these values it logically follows that these values are reasonable candidates for λ and hence should be included in a 95% confidence interval for λ.

I use the inequality http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/LRaccept.gifto solve for log lambda 0.



Essentially the inequality defines the lower limit for the likelihood confidence interval for λ but on a log-likelihood scale. I calculate this lower limit using the R function **qchisq** to return the .95 quantile of a chi-squared distribution with one degree of freedom.

lower.limit <- -out$minimum-.5\*qchisq(.95,1)

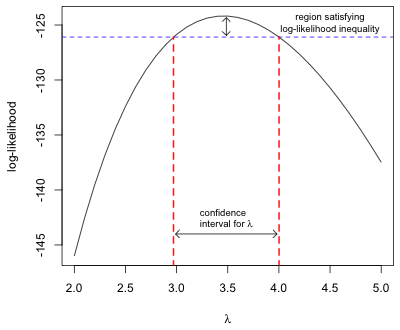
lower.limit

[1] -126.0971

We can understand this inequality better with a graph. I plot the log-likelihood and add the lower limit from the inequality as a horizontal line (Fig. 4).

plot(seq(2,5,.1), sapply(seq(2,5,.1), poisson.func), type='l', xlab=expression(lambda), ylab=log-likelihood.)

abline(h=lower.limit,col=4,lty=2)



**Fig. 4** Constructing a likelihood ratio-based confidence interval

As Fig. 4 shows the set of log-likelihood values that satisfy our inequality (indicated by the vertical arrow at top of graph) in turn define a corresponding set of values for the parameter λ (horizontal arrow at the bottom of the graph). These values for λ comprise what is called the profile likelihood confidence interval for λ. The boundaries of this confidence interval are defined by the places where the blue horizontal lower limit line intersects the graph of the log-likelihood then projected down to the λ-axis.

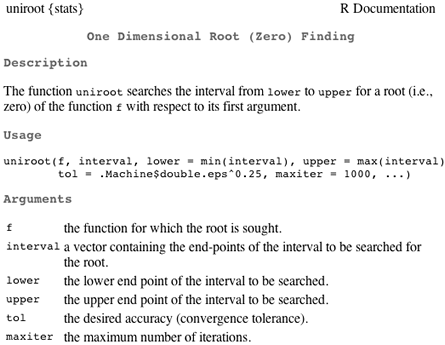
We could estimate the confidence limits graphically, but it is far simpler to use numerical methods. Let *H* represent the horizontal line in Fig. 4. We seek the values of λ such that

solve for lambda

where http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture8/rooteqn.gif. Thus we need to find the roots (zeros) of this function *f* .

Finding roots is a standard mathematical operation and there are many good methods for carrying this out. In R there is the function **uniroot**.

?uniroot



**Fig 5** Help screen for uniroot

The required arguments of **uniroot** are the function whose roots are desired and an interval over which to search. Since there are two roots (corresponding to the two endpoints of the confidence interval) we'll have to run **uniroot** twice with different search intervals each time. First I create the function, f of lambdadescribed above. Then I run **uniroot** twice each time using intervals that bracket the locations of the roots as determined from the graph in Fig. 4.

root.function<-function(lambda) poisson.func(lambda)-lower.limit

uniroot(root.function,c(2.5,3.5) )

$root  
[1] 2.96967

$f.root  
[1] -0.0002399496

$iter  
[1] 6

$estim.prec  
[1] 6.103516e-05

uniroot(root.function,c(3.5,4.5))

$root  
[1] 4.00152

$f.root  
[1] -8.254986e-05

$iter  
[1] 6

$estim.prec  
[1] 6.103516e-05

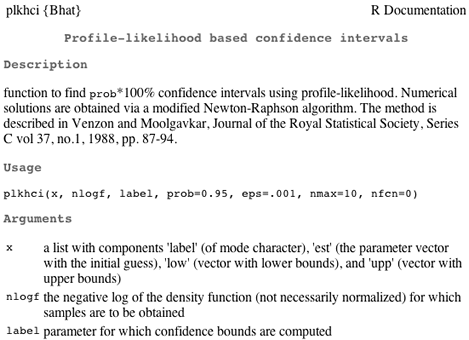
So to two decimal places a 95% profile likelihood confidence interval for λ is (2.97, 4.00). This compares favorably with the Wald confidence interval we found earlier: (2.94, 3.98).

**Profile likelihood confidence intervals using the Bhat package**

It turns out there is a function in the **Bhat** package that can be used to calculate profile likelihood confidence intervals directly. The function is called **plkhci**.

library(Bhat)

?plkhci



**Fig 6** Help documentation for plkhci

The usage instructions are fairly complicated and it is probably best to imitate one of the examples that appears at the end of the documentation. There are three required arguments.

1. The first argument is a list consisting of four components: a label containing the names of the parameters as used in the log-likelihood function, initial estimates for the MLEs, an estimated lower bound for the confidence interval, and an estimated upper bound for the confidence interval.
2. The second argument is the name of the negative log-likelihood function. (Note: Like **nlm** the **plkhci** function requires a negative log-likelihood!)
3. The third argument is the name of the parameter for which to calculate the confidence interval and must match a name that appears in the label component of the first argument to the **plkhci** function.

I create the list that is the first argument to **plkhci** and then evaluate the function on the required arguments.

x.in<-list(label='lambda',est=3.4, low=2, up=5)

plkhci(x.in, poisson.negloglik, 'lambda')

At the end of quite a bit of output we find the calculated confidence interval.

[1] 2.966524 4.005384

This matches our hand calculations above.

**Books and articles on likelihood**

* Azzalini, Adelchi. 1996. *Statistical inference: based on the likelihood.* New York: Chapman & Hall.
* Bolker, Benjamin M. 2008. *Ecological Models and Data in R*. Princeton University Press. Chapter 6 covers maximum likelihood.
* Devore, Jay L. 1995. *Probability and Statistics for Engineering and the Sciences*. Pacific Grove, CA: Duxbury Press. General discussion of maximum likelihood estimation with examples, pp 265–271.
* Edwards, A. W. F. 1992. *Likelihood*. Baltimore: John Hopkins University Press.
* Eliason, Scott R. 1993. *Maximum likelihood estimation: logic and practice*. Newbury Park, CA: Sage Publications.
* Hilborn, Ray and Marc Mangel. 1997. *The ecological detective: confronting models with data*. Princeton, NJ: Princeton University Press.
* Krebs, Charles J. 1999. *Ecological Methodology*. Menlo Park, CA: Addison-Wesley. A number of fairly advanced applications of maximum likelihood estimation appear on pp 84–88, 91, 128–131, 525–529.
* Larsen, Richard J. and Marx, Morris L. 1981. *An Introduction to Mathematical Statistics and Its Applications*. Englewood Cliffs, New Jersey: Prentice-Hall. General discussion of maximum likelihood estimation with examples, pp 212–218.
* McCallum, Hamish. 2000. *Population Parameters: Estimation of Ecological Models*. Oxford, England: Blackwell Science. General discussion of maximum likelihood estimation with examples, pp 29–44.
* Myung, In Jae. 2001. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* **47**: 90–100.
* Pawitan, Yudi. 2001. *In all likelihood: statistical modelling and inference using likelihood*. New York: Oxford University Press.
* Roff, Derek A. 2006. *Introduction to Computer-Intensive Methods of Data Analysis in Biology*. New York: Cambridge University Press. Chapter 2 covers maximum likelihood estimation. Uses S-Plus (code also works in R).
* Royall, Richard M. 1997. *Statistical evidence: a likelihood paradigm*. New York: Chapman & Hall.
* Severini, Thomas A. 2000. *Likelihood methods in statistics*. New York: Oxford University Press.
* Sorensen, Daniel. 2002. *Likelihood, Bayesian and MCMC methods in quantitative genetics*. New York: Springer-Verlag.

**Some web references on likelihood**

* [Maximum Likelihood Estimation](http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_1.html) by S. Purcell
* [Review of Likelihood Theory](http://data.princeton.edu/wws509/notes/a1.pdf) by German Rodriguez
* [Maximum Likelihood Estimation](http://cnx.rice.edu/content/m11446/latest/?format=pdf) by Clayton Scott and Rob Nowak

**R Code**

A compact collection of all the R code displayed in this document appears [here](http://www.unc.edu/courses/2010fall/ecol/563/001/notes/lecture8%20Rcode.html).

[Course Home Page](http://www.unc.edu/courses/2010fall/ecol/563/001/index.html)

|  |
| --- |
| Jack Weiss *Phone:* (919) 962-5930 *E-Mail:* jack\_weiss@unc.edu *Address:* Curriculum in Ecology, Box 3275, University of North Carolina, Chapel Hill, 27516 Copyright © 2010 Last Revised--October 2, 2010 URL: [http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture8.htm#lecture8) |