Uncertainty

env/bio 665 Bayesian inference for environmental models

Jim Clark

2020-02-24

Contents

#objectives

1. articulate a frequency versus a subjective probability interpretation of uncertainty
2. recognize the asymptotic Gaussian approximation and the role of the standard error in classical interval estimation
3. relate the likelihood function to Fisher Information and the classical standard error
4. relate the likelihood profile to classical hypothesis testing
5. implement a simple bootstrap
6. find the posterior distribution and Bayesian standard error for a conjugate likelihood-prior pair

#classical interval estimation

The **maximum likelihood estimate** (MLE) of a parameter is defined as the value finding most support from the data. But how much better is the MLE than some other value? As data accumulate confidence in the MLE should increase. The shape of the likelihood function contains additional information about the parameter estimate (Fisher 1959); increasing curvature in -log L with accumulation of data describes increasing confidence in the ML estimate.

Both classical and Bayesian approaches assume some underlying true value for the parameter that is fixed, and both produce estimates that are, in different senses, 'random'. A frequentist confidence interval is interpreted as the fraction of intervals estimated from a large number of idealized data sets generated by the same process that would include the true parameter value. The true (unknown) parameter value is viewed as being fixed and the data as being random. The term *frequentist confidence interval* is used to distinguish this interval, which is based on the idea of "coverage", from a Bayesian confidence interval or *credible interval*, which comes from integrating a likelihood (times prior) as though it were a parameter density. Because the data are viewed as random, the confidence interval is random—like the point estimate, the confidence interval is evaluated from random data.

##several interpretations

The frequency concept of randomness comes from the notion of repeating the same experiment, thereby repeatedly generating data sets. From each of these data sets we would obtain a different estimate. In the nonparametric bootstrap, this idea is simulated by resampling from the sample itself, generating random data sets, and building up a frequency distribution of estimates.

Although Bayesians speak of a parameter as random, the 'randomness' describes degree of belief in parameter values. The Bayesian approach uses a probability model for the parameter value to quantify uncertainty prior to data collection, which is updated by data. The posterior is conditioned on the data actually collected. In both cases, the confidence limits represent uncertainty as to the 'true' value. In this sense, Bayesian and classical approaches are not as different as they first appear.

In this unit I compare different types of "confidence envelopes"—a term I use to include both confidence and credible intervals—but focus first on a classical notion of a confidence interval.

##a confidence interval from introductory statistics class A brief example outlines the frequentist notion of a confidence interval. Consider a random variate y_i drawn from a normal distribution,

$$y_i \sim N(\mu, \sigma^2)$$

with parameters μ and σ^2. For the moment, suppose the variance parameter σ^2 is known, as would be the case if y_i were a measurement obtained on an instrument used previously. Based on this one observation, uncertainty as to the true value of μ could be represented by the density. This density is centered on the sample value y_i, because y_i is the best (ML) estimate of μ. To see this, consider the normal density centered on the a value μ,

$$N(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right]$$

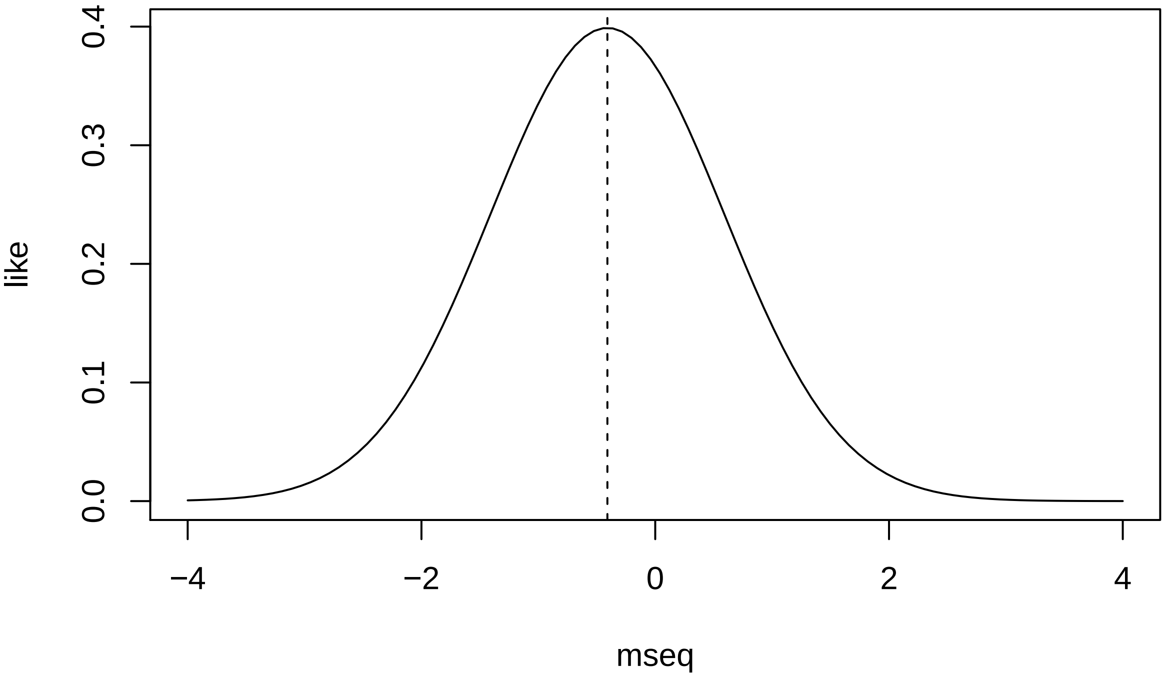
The "most likely" value of μ is that which yields the highest probability of the data, i.e., the ML estimate $\hat{\mu}$. To demonstrate that y_i is the ML estimate of μ, recall the log likelihood (including only terms involving μ),

$$\log L \propto -\frac{1}{2\sigma^2}(y_i - \mu)^2$$

Differentiating, setting the derivative equal to zero, and solving for μ gives gives the value $\hat{\mu}$, which is equal to y_i. Because y_i is the only observation, a density centered at y_i is more likely than is a density centered on any other value of y. Here is a plot:

```
mu   <- 0
y    <- rnorm(1,mu)
mseq <- seq(-4,4,length=100)
like <- dnorm(mseq,y)
```

```r
plot(mseq,like,type='l')
abline(v=y,lty=2)
```



=====

Example 1. Using count data, I want to estimate a mean value. Here is the Poisson likelihood:

$$L(\mathbf{y};\lambda) = \prod_i^n Poi(y_i|\lambda) = \prod_i^n \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \propto \lambda^{\sum_i y_i} e^{-n\lambda}$$

The log likelihood is

$$\log L \propto n\bar{y} \log \lambda - n\lambda$$

I find the MLE by differentiating and solving for $\hat{\lambda}$:

$$\frac{\partial \log L}{\partial \lambda} = \frac{n\bar{y}}{\lambda} - n$$

The solution is $\hat{\lambda} = \bar{y}$. Interpret this result.

=====

To construct a confidence interval for, say, $\alpha = 0.05$ (i.e., a 95% CI) draw two other densities, one for the lower confidence limit μ_L and one for the upper μ_H. The lower confidence limit comes from a density that has an area to the right of the sample value y_i equal to $\alpha/2 = 0.025$. The mean of this density is the lower confidence limit μ_L, which is selected such that the density $N(\mu_L, \sigma^2)$ has area$\alpha/2 = 0.025$ to the right of the estimate $\hat{\mu} = y_i$,

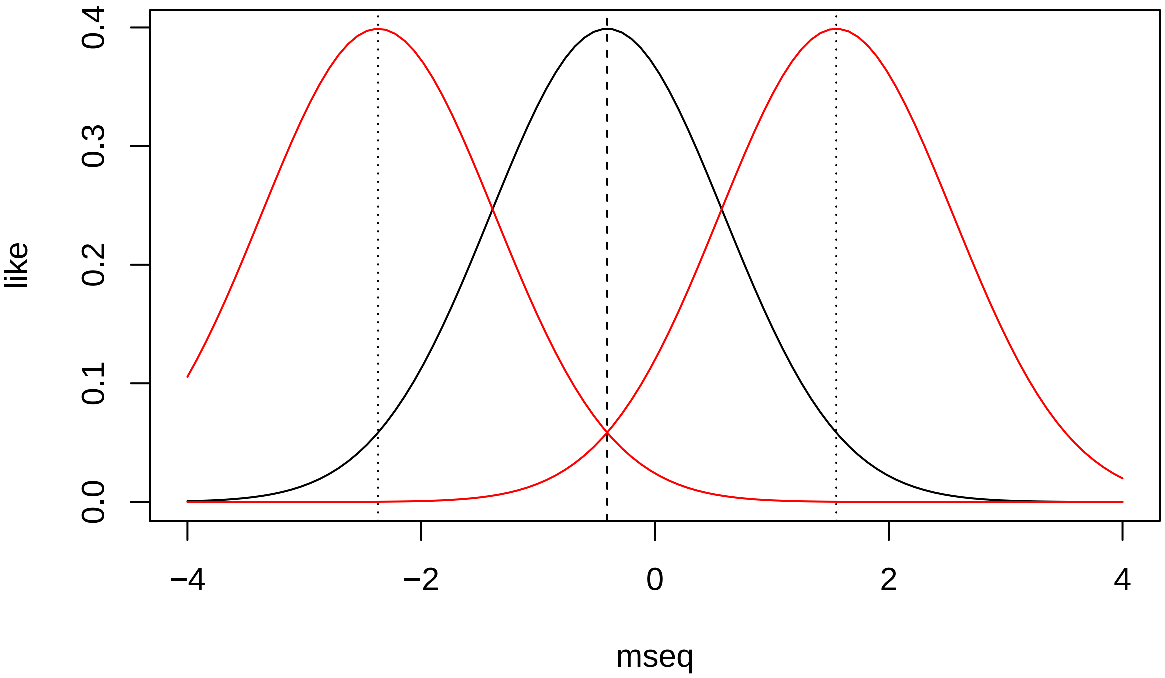
$$\int_{y_i}^{\infty} N(y|\mu_L, \sigma^2) dy = \alpha/2$$

The upper confidence limit satisfies,

$$\int_{-\infty}^{y_i} N(y|\mu_U, \sigma^2) dy = \alpha/2$$

Here is R code to evaluate the 95% confidence interval for a standard normal distribution. The function `qnorm` is supplied with the values of for the two tails (0.025 and 0.975 for $\alpha = 0.05$).

```r
plot(mseq,like,type='l')
abline(v=y,lty=2)
alpha <- .05
ci <- qnorm(c(alpha/2,1-alpha/2),y)
abline(v=ci,lty=3)
lines(mseq,dnorm(mseq,ci[1]),col=2)
lines(mseq,dnorm(mseq,ci[2]),col=2)
```



The interpretation reflects the view that the confidence interval is random, and the underlying true value of μ is fixed. With each new experiment there is a new estimate for the confidence interval. The concept of coverage means that $100(1˘\alpha)\%$ of confidence intervals calculated from identical experiments should contain μ.

The foregoing method follows a classical definition of a confidence interval. But it is not the way we usually construct confidence intervals, and it is not the way we usually think about them. Instead of two distributions, we tend to think of one distribution centered on the best estimate, with the confidence interval defined by the areas in the two tails. This method of drawing a distribution does not work for an example involving a Poisson likelihood.

The classical confidence interval can be calculated or approximated in several ways. I introduce the topic with a review of the standard error. Three methods for constructing confidence intervals that follow all involve the likelihood function. The first technique, the **likelihood profile**, focuses on likelihood shape. I follow with **Fisher Information**, which summarizes that shape in terms of its curvature at the MLE. Next is a numerical method, the **bootstrap**, which yields a frequency distribution of estimates that has the shape of a likelihood function. All of the methods approximate the frequentist idea of coverage (the true parameter value lies within a fraction $(1-\alpha)$ of the CIs that would be constructed from similar experiments). A fourth method, integrating the likelihood only makes sense in the context of Bayes' theorem, assuming that the prior is so flat as to be ignored. Last I describe a Bayesian credible interval, which comes from mixing the likelihood with prior.

##the standard error as basis for a confidence interval

The standard error expresses uncertainty in an estimate. This example concerns the standard error for a mean parameter of a normal distribution μ, which is estimated by taking the average over n samples, i.e., the sample mean

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

This is the maximum likelihood estimate of μ. Confidence in this estimate increases with sample size n. The residual variance is the estimate of σ^2,

$$\hat{\sigma}^2 = var[y] = \frac{1}{n-1}\sum_{i=1}(y_i - \hat{\mu})^2$$

The standard error summarizes support from the data for a particular value. It is the square root of the variance divided by the sample size

$$se_{\hat{\mu}} = \sqrt{\frac{\hat{\sigma}^2}{n}}$$

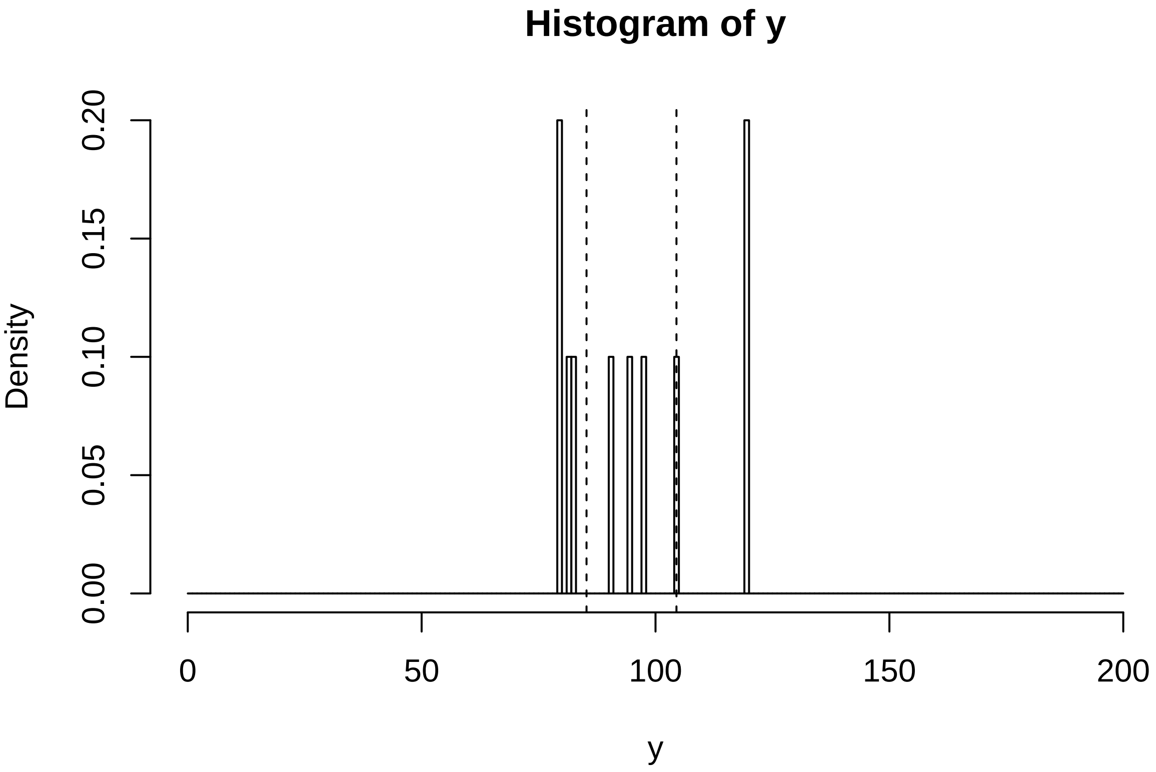
This is the standard error of the estimate of μ. From the formula for the standard error it is clear that the width of the confidence interval must decrease with increasing sample size.

Recall from introductory statistics that 68% of the intervals spanning 1 se on either side of the mean estimate are expected to contain the true value. Ninety five percent of the intervals spanning 1.96 standard errors of the mean estimate should contain the true value. This is equivalent to saying that the frequency distribution of parameter estimates we would obtain from repeating the experiment can be approximated by a normal distribution

$$\hat{\mu} \sim N\left(\bar{y}, se_{\hat{\mu}}^2\right)$$

Here is a random sample of n = 5 observations drawn from the normal density with mean $\mu = 100$ and variance $\sigma^2 = 10$

```
n    <- 10
mu   <- 100
sd   <- 20
yseq <- seq(0,2*mu,by=1)
y    <- rnorm(n,mu,sd)
hist(y,breaks=yseq,probability=T)
se   <- sqrt(var(y)/n)
mle  <- mean(y)
ci   <- c(mle - 1.96*se,mle + 1.96*se)
abline(v=ci,lty=2)
```



The sample mean $\hat{\mu}$ will be close to the true population mean μ. So with just a few observations I have a reasonable estimate of the mean. Ninety five percent of the confidence intervals obtained from samples of size $n = 10$, calculated as 1.96 standard errors of sample means, should contain the true μ.

The 95% confidence interval I obtain for this experiment at 1.96 standard errors from the mean is

```
ci
```

```
## [1]  85.25211 104.47851
```

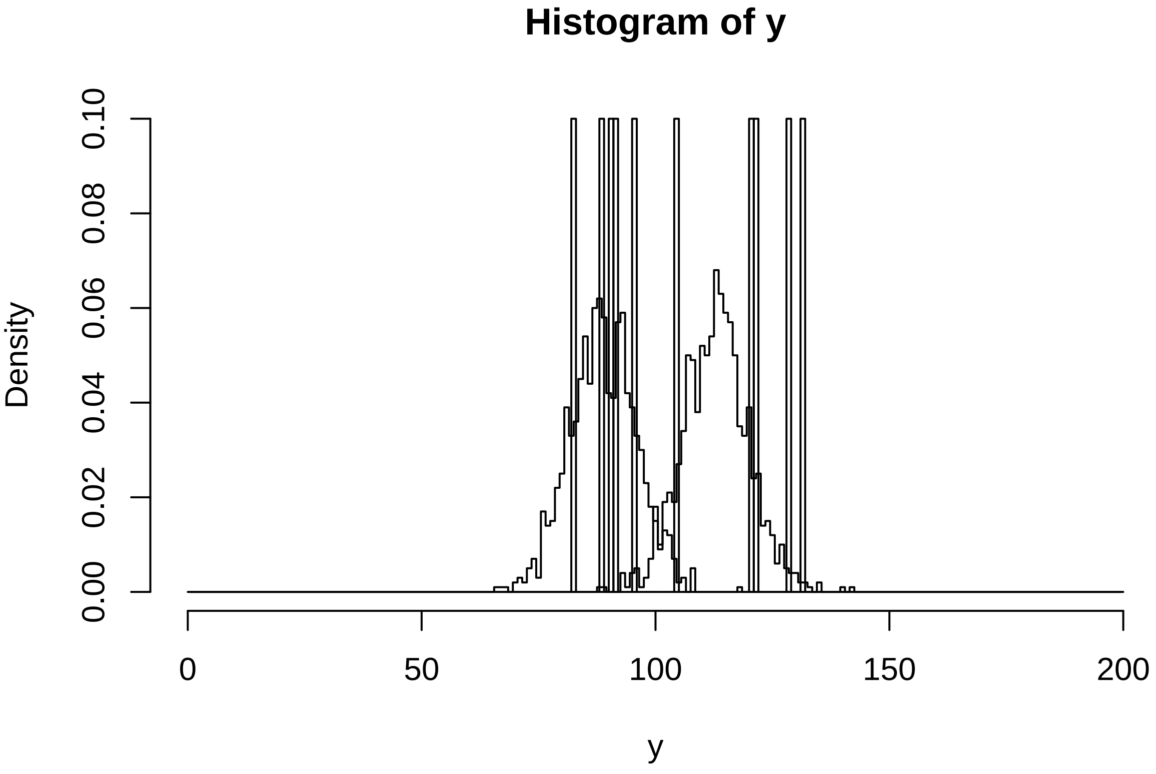
I could check this result using `qnorm` to find the values where the lower and upper tails of the distribution equal $\alpha/2$, where $\alpha = 0.05$,

```
alpha <- .05
qnorm(c(alpha/2,1 - alpha/2),mle,se)
```

```
## [1]  85.25229 104.47833
```

If I could repeat the experiment 1000 times, a histogram of CIs might be represented like this:

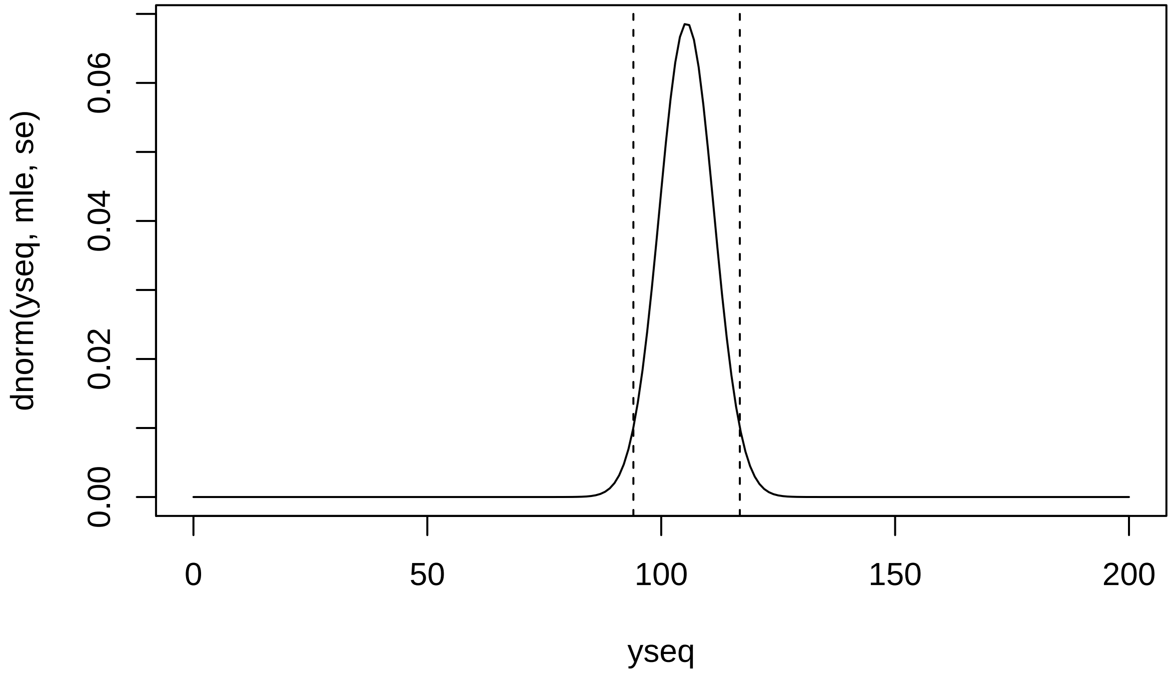
```
rep  <- 1000
ci   <- matrix(0,rep,2)
for(i in 1:rep){
  y   <- rnorm(n,mu,sd)
  se  <- sqrt(var(y)/n)
  mle <- mean(y)
  ci[i,] <- c(mle - 1.96*se,mle + 1.96*se)
}
clo <- hist(ci[,1],breaks=yseq,plot=F) #lower conf limit
chi <- hist(ci[,2],breaks=yseq,plot=F) #upper conf limit
hist(y,breaks=yseq,probability=T)
lines(clo$mids,clo$density,type='s')
lines(chi$mids,chi$density,type='s')
```



```
cover <- length(which(mu > ci[,1] & mu < ci[,2]))/rep
```

The value `cover` is the fraction of confidence intervals that include the true value of the parameter. Because the likelihood is Gaussian, these two densities yield an answer not too different from what I would obtain if I simply integrated the tails of the density $N\left(\hat{\mu}|\bar{y}, se^2_{\hat{\mu}}\right)$, one centered on the sample estimate $\hat{\mu} = \bar{y}$,

```
plot(yseq,dnorm(yseq,mle,se),type='l')
ci <- qnorm(c(.025,.975),mle,se)
abline(v = ci, lty = 2)
```



A sample of size $n = 50$ results in an estimate $\hat{\mu}$ that is much closer to the true value. The increased information available from a sample of size $n = 50$ over one of size $n = 5$ is reflected in a narrow 95% CI. This example illustrates the random character of a classical confidence interval. Any given sample from the same population would produce a different confidence interval, referenced by the estimate and its standard error. Thus, a confidence interval can be calculated from a standard error. When the likelihood is not normal, the confidence interval is approximate.

#likelihood profile

An absolute probability cannot be assigned to a particular parameter estimate or to a model. We can think of the probability of a particular value of a parameter only in the context of some other value or model. The likelihood function is used for such comparisons. Contrary to intuition, classical methods do not permit integrating the likelihood function. Instead confidence intervals can be derived using methods that involve ratios of likelihoods using a **likelihood ratio test** (LRT).

Let the vector $\mathbf{y} = [y_1, \ldots, y_n]$ represent n observations and μ be one or more (a vector of) parameters about which we wish to draw inference. Ideas about the value(s) of μ can take the form of hypotheses. Because a classical approach does not yield a direct probability statement about μ, hypothesis testing is 'evolutionary', through progressive competitions with alternative views about μ. By eliminating the alternatives, one at a time, the idea is that confidence will accumulate in support of a particular hypothesis. A probability statement is made in the context of two competing hypotheses, e.g., a "null" (say, $\mu = \mu_0$) and an alternative, such as $\mu \neq \mu_0$ or $\mu = \mu_1$. The standard F, t, and χ^2 tests all arise in the context of classical methods for relating likelihoods of two hypotheses based on the view that both values of μ are fixed, and the data are random. The probability involves two likelihoods, both of which assume the respective hypothesis to be true. I begin with a sample mean.

##likelihood ratio and deviance

Suppose we wish to make a probability statement about the value of μ finding most support from a data set, i.e., $\hat{\mu}$, when the likelihood is normal. The alternative hypothesis concerns a rival value of μ, call it μ_0. The test statistic for this comparison involves a likelihood ratio

$$R = \frac{L(\mathbf{y}; \mu_0)}{L(\mathbf{y}; \hat{\mu})}$$

This is the ratio of two likelihoods taken over the same data $\mathbf{y}$, but assuming different values of μ. The likelihood ratio can be evaluated for any two values of μ. When the denominator is taken at the MLE (as in this case), the LR is termed the **normed likelihood function** (Lindsey 1999) and has range $[0, 1]$. The test statistic is sometimes called a **deviance**,

$$D = -2 \log R$$

and is distributed as χ^2 with, in this case, 1 degree of freedom, because there is only one parameter at issue.

Consider again the likelihood for the mean of the normal distribution,

$$\begin{aligned} R &= \frac{L(\mathbf{y}; \mu_0)}{L(\mathbf{y}; \hat{\mu})} \\ &= \frac{\sigma_0^{-n} exp\left[-\frac{1}{2\sigma_0^2} \sum_i (y_i - \mu_0)^2\right]}{\hat{\sigma}^{-n} exp\left[-\frac{1}{2\hat{\sigma}^2} \sum_i (y_i - \hat{\mu})^2\right]} \\ &= \left(\frac{\hat{\sigma}}{\sigma_0}\right)^n \end{aligned}$$

with the deviance

$$D = -2 \log R = -2n \left(\log \hat{\sigma} - \log \sigma_0\right)$$

The likelihood ratio test provides a confidence interval that is identical to that found by the integral method. Because the sampling distribution is normal, the log likelihood is a parabola and symmetric about the minimum at $\hat{\mu}$. If $\hat{\mu}_0$ is taken to be the value of $\hat{\mu}$ where the log likelihood function is 1/2 unit below the value at the maximum, then $D = 1$, and the probability of D (χ^2 with 1 df) is 0.68. A value of $D = 3.84$ corresponds to a probability of 0.05.

##likelihood ratio test to profile

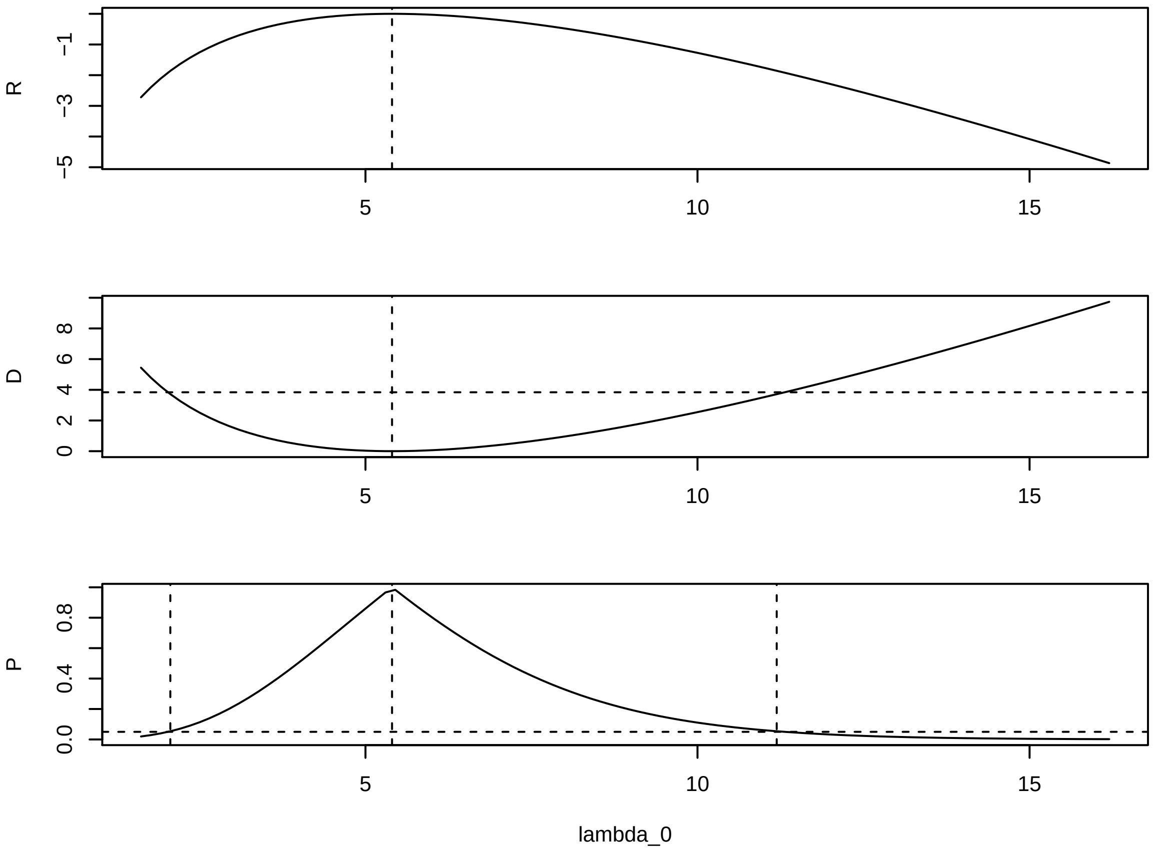
=====

Example 2. I want to use a likelihood ratio test to determine a confidence interval for the count data in Example 1. The MLE is $\hat{\lambda} = \bar{y}$. Substituting for the MLE, $\hat{\lambda} = \hat{y}$, the likelihood ratio is

$$\log R = \bar{y}(\log \lambda_0 - \log \bar{y} + 1) - \lambda_0$$

Here is a simulated data set:

```
n    <- 10
y    <- rpois(n,4.7)                                   #simulated counts
ybar <- mean(y)
rseq <- seq(.3,3,length=100)*ybar
R    <- ybar*( log(rseq) - log(ybar) + 1) - rseq    #normed likelihood

par(mfrow=c(3,1), mar=c(4,4,1,4))
plot(rseq, R, type='l', xlab='')
abline(v=ybar,lty=2)
D  <- -2*R                #Deviance
plot(rseq,D,type='l', xlab='')
abline(v=ybar,lty=2)
abline(h=3.84,lty=2)      #value of D at P = 0.05

P <- 1 - pchisq(D,1)      #P value
plot(rseq, P, type='l', xlab='lambda_0')
abline(v=ybar,lty=2)
abline(h=.05, lty=2)

#find the 95% CI:
CI <- range(rseq[P > .05])
abline(v=CI, lty=2)
```



I used the R function `pchisq` for the χ^2 distribution with one degree of freedom to find the area in the tail, beyond the value of D evaluated for this data set.

Repeat this experiment using a sample size of 100.

Exercise 1 Find the MLE, the likelihood profile, and the 95% CI for waiting times that are exponentially distributed:

$$L(\mathbf{y}; \lambda) = \prod_{i=1}^{n} \lambda e^{-y_i \lambda}$$

Unlike the Gaussian example, the CI for the Poisson and exponential distributions are not symmetric. The relationship between likelihood and deviance is apparent from a plot of the normed likelihood function R, where the likelihood in the denominator is taken to be the ML estimate. The normed likelihood thus has a maximum value of 1 at the MLE and declines on either side. At an α level of, say, 0.05 the CI can be constructed by finding the values of D that yield $(1 - \alpha) = 0.95$ probability from the χ^2 test. At the MLE, $R = 1$, $D = 0$, and $P = 1$, because the ratio is taken for two equivalent values finding equal support from the data. The profile itself is constructed by calculating R for a range of parameter values λ_0 in the neighborhood of the MLE. Deviances increase on either side of the MLE, whereas the associated P values decline.

The likelihood profile is based on the distance below the maximum likelihood, and not on integrating the likelihood. The likelihood ratio provides an economical means for comparing parameter values, because all coefficients that do not contribute drop out. The normal example collapsed to a ratio of standard deviations, because other coefficients were redundant between the two models being compared. We can ignore some coefficients provided the comparison involves the same functional form. Model comparisons that involve different functional forms can still be accomplished using likelihoods, but all coefficients must be retained.

#approximate CI's from Fisher information

The likelihood profile makes use of the likelihood function for the information it provides about a parameter. The decline in the normed likelihood function on either side of the MLE can be summarized by the width (curvature) of the function at the MLE (Fisher 1959). **Fisher Information** uses the curvature of the log likelihood function to estimate a variance for the parameter error distribution, which might be obtained as a frequency of estimates obtained by repeating the experiment. The curvature is actually a quadratic approximation (2nd derivative) of the log likelihood function in the neighborhood of the MLE. If this curvature is slight, then the data do not contain much information about the parameter, and the standard error is large, and vice versa. The quadratic is exact for a normal sampling distribution, because the log likelihood is quadratic, but it can be a poor approximation for asymmetric likelihoods when sample sizes are small.

##approximate likelihood

Fisher Information is the expected value of the width of the likelihood function, which is inversely related to the curvature. This curvature is found by approximating the log likelihood function with a polynomial, called a **Taylor expansion**,

$$\log L(\theta) = \sum_{k=0}^{\infty} \frac{(\theta - \hat{\theta})^k}{k!} \times \left. \frac{d^k \log L}{d\theta^k} \right|_{\hat{\theta}}$$

Including up through quadratic terms this series is

$$\log L(\theta) = \log L(\hat{\theta}) + (\theta - \hat{\theta}) \frac{d \log L(\hat{\theta})}{d\theta} + \frac{(\theta - \hat{\theta})^2}{2} \frac{d^2 \log L(\hat{\theta})}{d\theta^2} + \dots$$

The terms contribute increasingly less to the estimate with increasing order. The second term disappears, because the derivative of $\log L$ at the MLE is zero. To simplify notation, let

$$I = - \left. \frac{d^2 \log L(\theta)}{d\theta^2} \right|_{\hat{\theta}}$$

and write the series as

$$\log L(\theta) = \log L(\hat{\theta}) - \frac{I(\theta - \hat{\theta})^2}{2}$$

or, equivalently,

$$\log L(\theta) - \log L(\hat{\theta}) = -\frac{I(\theta - \hat{\theta})^2}{2}$$

The LHS of this expression is just the log of the normed likelihood, so

$$R(\theta) = exp\left[-\frac{I(\theta - \hat{\theta})^2}{2}\right]$$

This has the same shape as (is proportional to) a normal density with a mean parameter of $\hat{\theta}$ and a variance parameter of I^{-1}, i.e.,

$$R(\theta) \propto N(\theta|\hat{\theta}, I^{-1})$$

This approximate proportionality is the basis for the standard error, calculated in two steps: 1. Determine the curvature of $-\log L$ near the MLE. For one parameter, Fisher Information is

$$I = -\frac{d^2 \log L(\theta)}{d\theta^2}\bigg|_{\hat{\theta}}$$

2. Estimate the standard error as $se_{\hat{\theta}} = 1/\sqrt{I}$. I is known as the 'observed information', because it is obtained by plugging in the MLE obtained from the data.

=====

Exercise 2. Use Fisher Information to find the standard error of the mean of a normal sampling distribution. Hint: you will need this:

$$I = -\frac{d^2 \log L(\mu)}{d\mu^2}\bigg|_{\hat{\mu}}$$

=====

=====

Exercise 3. Estimate the standard error for the exponential model using Fisher Information.

=====

##Fisher information in multiple dimensions

=====

Example 3. This works for multiple dimensions. Take a regression model, where the likelihood is

$$\log L \propto -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Here are derivatives:

$$\frac{\partial \log L}{\partial \beta} = -\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{y} - \beta'\mathbf{X}'\mathbf{X}), \frac{\partial^2 \log L}{\partial \beta^2} = -\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}$$

The MLE is the vector $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, with parameter covariance $cov(\hat{\beta}) = \mathbf{I}^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The standard errors are the square roots of the diagonal of this matrix.

#Bootstrap

The bootstrap is a numerical method that can be used to obtain confidence intervals (Efron and Tibshirani 1993). The basic methodology can be extended for propagation of error for purposes of prediction, sometimes termed Monte Carlo simulation. Unlike other methods, bootstrapping does not require much math. If a point estimate is available, then so too is the bootstrap. There are two types of bootstraps to consider, non-parametric and parametric.

Standard errors and confidence intervals by nonparametric bootstrap—The nonparametric bootstrap produces estimates based on resampling the data. The resampling procedure involves sampling from the data 'with replacement'. Suppose the sample consists of n observations. The method involves drawing a sample of size n from the data set. This means that once an observation has been included in the sample, it can still be sampled again. This procedure simulates the frequentist concept of obtaining estimates from repeated similar experiments. It substitutes resampling of one data set for repeated experiments.

##simple recipe

The following five steps can be used to simultaneously estimate standard errors and confidence intervals. The extension to multiple parameters is straightforward. This algorithm assumes a sample of size n for which I can estimate the parameter θ using an estimator function (e.g., a likelihood function).

1. Draw with replacement a 'resample' of size n from the original sample.
2. Estimate θ from this resample. Let θ_b represent this b^{th} estimate of θ.
3. Repeat this procedure B times. For a standard error estimate, B might be as low as 50. For a 95% confidence interval, B might be more like 2000. (Smaller α values require larger samples.) There are now B estimates of the parameter, one for each bootstrap resample.
4. Estimate the parameter standard error as the standard deviation of the B replicates.
5. Estimate the confidence interval as the interior $100(1 - \alpha)\%$ of the bootstrapped estimates.

##application

————————————— ===== —————————————

Example 4. Estimate how cone production increases with tree size.
For counts, I used a Poisson likelihood, assuming that cone production increases with the

square of diameter x_i,

$$L = \prod_{i=1}^{n} Poi(y_i|\beta x_i^2) \propto \beta^{n\bar{y}} exp[-\beta n \bar{x^2}]$$

I get the MLE from

$$\frac{\partial \log L}{\partial \beta} = \frac{n\bar{y}}{\beta} - n\bar{x^2}$$

having the solution $\hat{\beta} = \bar{y}/\bar{x^2}$. To find an approximate standard error I differentiate again

$$\frac{\partial^2 \log L}{\partial \beta^2} = -\frac{n\bar{y}}{\beta^2} = -\frac{n}{\bar{y}}\left(\bar{x^2}\right)^2$$

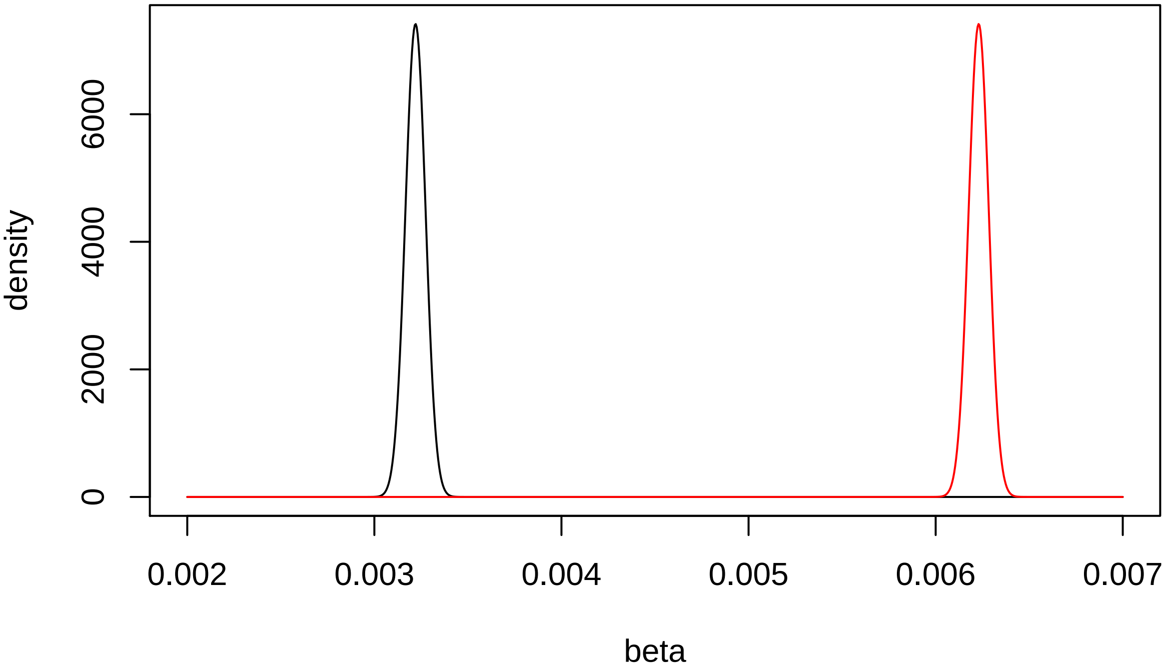
to obtain $se_{\hat{\beta}} = \frac{\sqrt{\bar{y}}}{\bar{x^2}\sqrt{n}}$.

For the last step I substituted the MLE for β. Here is the estimate and standard error for the cone example at ambient CO2 in R:

```r
filename <- ('../dataFiles/FACEtrees.txt')
data <- read.table(filename, header=T)

y <- data[,'cones']
x <- data[,'diam']
w <- is.finite(x) & is.finite(y)
ambient  <- which(data[,'trt'] == 0 & w)
elevated <- which(data[,'trt'] == 1 & w)

Y    <- mean(y[ambient])
X2   <- mean(x[ambient]^2)
nlo  <- length(ambient)
bmuA <- Y/X2
bseA  <- sqrt(Y/nlo)/X2

Y    <- mean(y[elevated])
X2   <- mean(x[elevated]^2)
nhi  <- length(elevated)
bmuE <- Y/X2
bseE  <- sqrt(Y/nhi)/X2

estimates <- signif( matrix( cbind(c(bmuA, bseA),c(bmuE, bseE)), 2, 2), 3 )
rownames(estimates) <- c('mean','se')
colnames(estimates) <- c('ambient','elevated')
```

I could drawn the normal distribution approximated by this mean and standard deviation like this:

```
bseq <- seq(.002,.007,length=1000)
plot(bseq,dnorm(bseq,estimates['mean',1],estimates['se',1]),type='l',
     xlab='beta',ylab='density')
lines(bseq,dnorm(bseq,estimates['mean',2],estimates['se',1]),col=2)
```



For comparison here is a bootstrap. I begin by defining the number of estimates that will be obtained and a vector to store them:

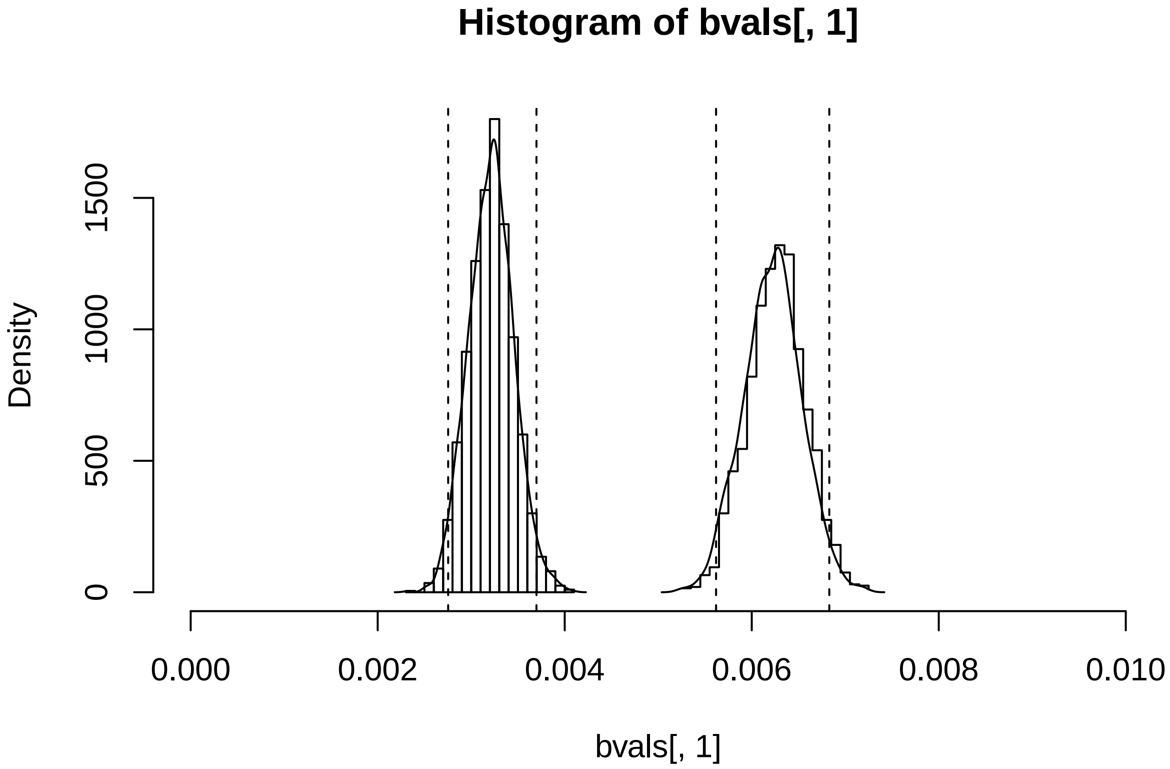
```
nboot <- 2000          #no. bootstrap estimates
bvals <- matrix(0, nboot, 2) #matrix to hold estimates
colnames(bvals) <- c('ambient','elevated')
```

The bootstrap is done in a loop. At each iteration, a new sample is obtained determined by the index for b, called bindex. The data are then extracted from ylo and xlo and used to calculate the MLE for the parameter.

```
for(b in 1:nboot){

  bindex <- sample(ambient, nlo, replace=T) #sample with replacement
  Y    <- mean(y[bindex])
  X2   <- mean(x[bindex]^2)
  bvals[b, 1] <- Y/X2

  bindex <- sample(elevated, nhi, replace=T) #sample with replacement
  Y    <- mean(y[bindex])
  X2   <- mean(x[bindex]^2)
  bvals[b, 2] <- Y/X2
}
```

From this sample I determine the standard deviation, which is an estimate of the standard error, and quantiles, as an estimate of the confidence interval,

```
se <- apply(bvals, 2, sd)

hist(bvals[,1],freq=F,nclass=20,xlim=c(0,.01))
  lines(density(bvals[,1]))
  ci1 <- quantile(bvals[,1],c(0.025,.975))
  abline(v=ci1,lty=2)

elev <- hist(bvals[,2],nclass=20, plot=F)
lines(elev$mids, elev$density, type='s')
  lines(density(bvals[,2]))
  ci2 <- quantile(bvals[,2],c(0.025,.975))
  abline(v=ci2,lty=2)
```



The two methods yield quite different confidence intervals, most likely because of the large number of zeros in the data. In the next example, I consider a classical analysis of these data that might be done with standard software, followed by one where we explicitly model the zeros as a separate process.

——————————————— ===== ———————————————

#Bayesian credible intervals for conjugate prior-likelihood pairs

The Bayesian **credible interval** can be obtained directly by integrating the posterior. This posterior is like the density I constructed with a bootstrap, and so returns us to the way we typically think about confidence intervals. Thus, the way we tend to think about confidence intervals corresponds to a Bayesian credible interval with a "uniform" prior. So far, I have talked about posterior distributions that have a parametric form. In other words, there is a standard distribution that can be described by parameters. For a 95% credible interval, I seek the upper and lower interval limits such that

$$0.95 = \int_{\theta_l}^{\theta_u} [\theta|\mathbf{y}] d\theta$$

or, if the parameter θ assumes discrete values,

$$0.95 \leq \sum_{\theta_l \leq \theta \leq \theta_u} [\theta|\mathbf{y}]$$

This intuitive definition represents the subjective probability that the true value of θ lies between θ_l and θ_u. This differs from the frequentist concept of coverage, which invokes the hypothetical idea of repeating the experiment many times and determining the fraction of confidence intervals that contain θ. Of course, the values ofθ_l and θ_u that satisfy these equalities are not unique, and I would like the interval to be as narrow as possible. The highest posterior density (HPD) is defined by a horizontal line drawn through the density. This line intersects the posterior density in two places, thus defining two tails. A different line drawn such that the fraction of the density in the tails is $1 - \alpha$ (a 95% credible interval would have $\alpha = 0.05$) defines the HPD. The areas in the two tails can differ by this method. Much easier to compute is the equal-tail interval, where each tail has the fraction $\alpha/2$ of the density, and is hereafter used throughout this book. The two are equivalent for symmetric, unimodal densities. Otherwise, the equal-tail interval is slightly wider.

##a Poisson-gamma pair

Example 5. For a Poisson example, the likelihood is

$$[\mathbf{y}|\lambda] = \prod_i^n Poi(y_i|\lambda) = \prod_i^n \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{n\bar{y}} e^{-n\lambda}}{\prod_i^n y_i!}$$

With a gamma prior distribution for λ this becomes

$$\prod_i^n Poi(y_i|\lambda) gam(\lambda|a,b) \propto \lambda^{n\bar{y}} e^{-n\lambda} \times b^a \lambda^{a-1} e^{-b\lambda}$$

I simplify this to

$$\lambda^{n\bar{y}+a-1} e^{-\lambda(n+b)}$$

Note that this is a new gamma density, $gam(\lambda|n\bar{y} + a, n + b)$.

Together, the Poisson likelihood with Gamma prior distribution yield a Gamma posterior having parameters that sum the contributions of prior and likelihood. For conjugate likelihood-prior pairs I can skip the integration of the denominator, because the normalizer is already known. To find the posterior, I follow these simple steps:

1. Ignore the coefficients that do not include the parameter of interest, writing the posterior as the product of those coefficients that do contain the parameter.
2. Collect coefficients.
3. Identify the parameters for the posterior based on comparison with the prior.

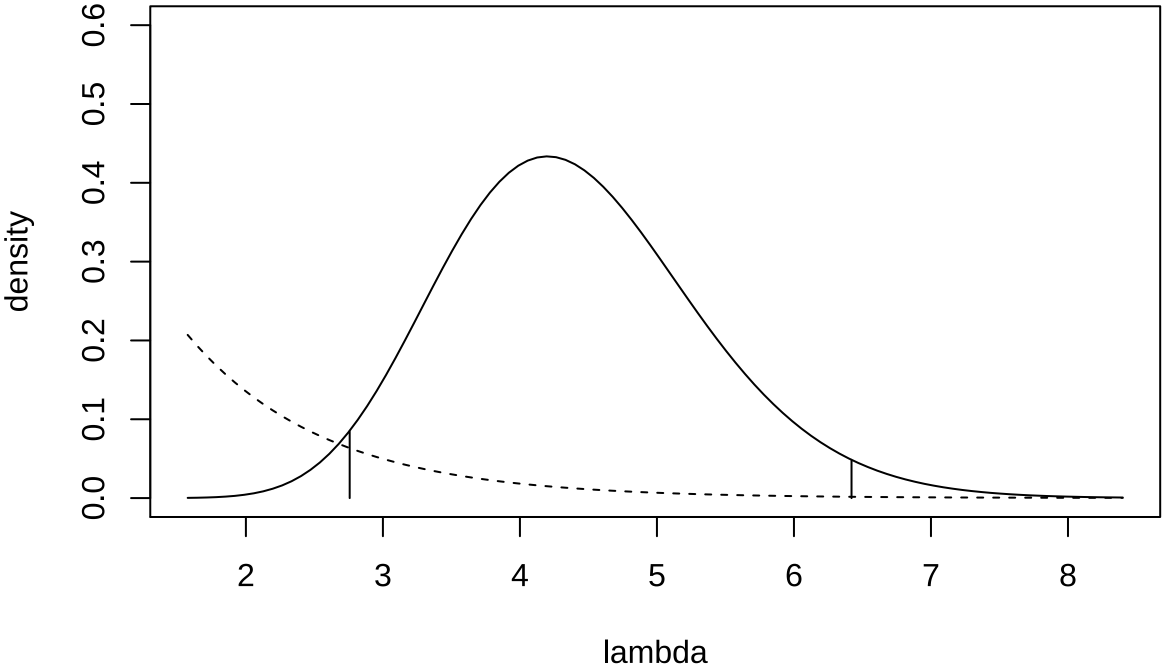
##standard error and credible interval

Because the gamma distribution has a parametric form I can easily find a standard error as the square root of the variance. The variance of the distribution $gam(\lambda|a, b)$ is a/b^2. Thus, the Bayesian standard error for the previous example is

$$se_{\hat{\lambda}} = \sqrt{\frac{n\hat{y} + a}{(n + b)^2}}$$

The credible interval can be found using the `qgamma` function in R.

```r
n    <- 4
y    <- rpois(n,4.7)                                   #simulated counts
ybar <- mean(y)
rseq <- seq(.3,1.6,length=100)*ybar
a <- b <- 1
plot(rseq,dgamma(rseq, a, b), type='l', lty=2, ylim=c(0,.6),
     xlab = 'lambda', ylab = 'density')
A <- sum(y) + a
B <- n + b
lines(rseq,dgamma(rseq, A, B), type='l')
ci <- qgamma(c(.025,.975), A, B)
segments(ci[1], 0, ci[1], dgamma(ci[1],A,B))
segments(ci[2], 0, ci[2], dgamma(ci[2],A,B))
```



=====

Exercise 4. For the cone example, I used the likelihood $Poi(y_i|\beta x_i)$. Combine this likelihood with the prior $gamma(\beta|a, b)$ and answer the following:

a) What is the posterior density for β?

b) For simulated data sets of $n = 5$, how do the standard errors and credible intervals for this model compare with Fisher information?

c) The form of the Bayesian standard error and the standard error from Fisher Information look different. Can you explain why numerically they are similar? [Hint: think about sample size n].

=====

#Group exercise

I want to get a feel for the uncertainty on the estimate of the variance for a Gaussian model for continuous observations y‘. I would like to compare different methods.

1. Find the MLE for the variance. If you have time to kill, derive the Fisher information and the SE.
2. Generate bootstrapped estimates for the uncertainty.
3. Complete a Bayesian analysis, choose a prior. Derive the result, then use MCMC.

Compare the estimates for different sample sizes.

#recap

Classical methods and Bayes often give similar estimates for small problems, but this is not the case for large models. The classical summaries reviewed here highlight some of the conceptual differences. Some advocates of classical philosophy combined with modern computation implement large models without prior distributions, apparently to maintain some purity about data uncorrupted by prior. Bayesians would respond that the underlying lack of an axiomatic foundation is the larger conceptual issue. Pragmatically, the prior distribution also stabilizes large models, while bringing in information outside what is contained in the data. As models increase in size, not only are the conceptual differences large, but results can differ substantially.