BISC-869, Likelihood

March 4, 2020

Why Likelihood?

Likelihood lets you come up with your own statistical analyses.

Many of the built in methods we are learning here (e.g., `lmer`) use likelihood to find parameter estimates. So, its important to know what's going on behind the scenes.

What is probability?

Frequentist definition:

The probability of an event is the proportion of times that the event would occur if a random trial is repeated over and over again under the same conditions.

A probability distribution is a list of all mutually exclusive outcomes of a random trial and their probabilities of occurrence.

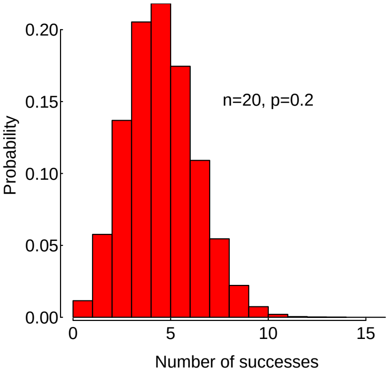
Example: binomial distribution

The binomial distribution is the probability distribution of the number of "successes" in n independent trials, when the probability of success p is the same in each trial.

$$\mathrm{Pr}[Y \mathrm{successes}] = \binom{n}{Y} p^Y (1-p)^{n-Y}$$

$\binom{n}{Y}$, which denotes "n choose Y", counts up the different ways of getting exactly Y successes out of n trials.

For $n = 3$ and $Y = 2$, we have *SSF*, *SFS*, *FSS*.



Conditional probability

The conditional probability of an event is the probability of that event occurring given that a condition is met. "|" symbol used to indicate "given".

The probability that the second child born to a couple is a girl, given that their first child was a girl,

$$\Pr[\text{second child is girl} \mid \text{first child is girl}]$$

Other conditional probabilities:

$$\Pr[\text{we see an elephant today} \mid \text{we are in the Serengeti}]$$

$$\Pr[\text{we see an elephant today} \mid \text{we are in Manhattan}]$$

$$\Pr[12 \text{ successes in } n \text{ trials} \mid p = 0.50]$$

$$\Pr[12 \text{ successes in } n \text{ trials} \mid p = 0.10]$$

Likelihood

Likelihood is a conditional probability.

The likelihood of a population parameter equaling a specific value, given the data, is the probability of obtaining the observed data given that the population parameter equals the specific value.

$$\mathcal{L}[\text{parameter} = \rho|\text{data}] = Pr[\text{data}|\text{parameter} = \rho]$$

Law of Likelihood:

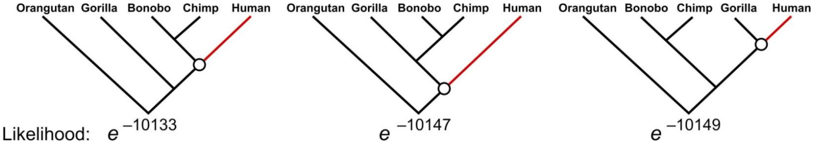
The extent to which data supports one parameter value or hypothesis against another is equal to the ratio of their likelihoods (difference in their log-likelihoods).

Method invented by R. A. Fisher when he was a 3^{rd}-year undergraduate.

Likelihood

Likelihood is used a lot in phylogeny estimation

Three proposed trees of ancestor–descendant relationships between humans and the other great apes. The human branch and our shared ancestor with the other apes is highlighted. Numbers at the bottom are the likelihoods of each proposed tree based on gene sequence data (Rannala and Yang 1996). The likelihood of the left-most tree is the highest.



What matters is not the likelihood of each tree, but the likelihood of each tree *relative to the others*.

Example: Estimate a binomial proportion p

Data: The tiny wasp, *Trichogramma brassicae*, rides on female cabbage white butterflies, *Pieris brassicae*. When a butterfly lays her eggs on a cabbage, the wasp climbs down and parasitizes the freshly laid eggs.

Fatouros et al. (2005, *Nature*), carried out trials to determine whether the wasps can distinguish mated female butterflies from unmated females. In each trial a single wasp was presented with two female cabbage white butterflies, one a virgin female, the other recently mated. $Y = 23$ of 32 wasps tested chose the mated female.

What is the proportion p of wasps in the population choosing the mated female?

$Y = 23$ "successes", $n = 32$ trials. Use these data to estimate p.

Example: Estimate a binomial proportion p

Likelihood function for the binomial proportion p

Data: $Y = 23$, $n = 32$

$$\mathcal{L}[p|Y \text{ chose mated female}] = Pr[Y \text{ chose mated female}|p]$$
$$\mathcal{L}[p|23 \text{ chose mated female}] = \binom{32}{23} p^{23}(1-p)^9$$

For example, the likelihood of $p = 0.5$, given the data, is

$$\mathcal{L}[p = 0.5|23 \text{ chose mated}] = \binom{32}{23}(0.5)^{23}(1-0.5)^9 = 0.00653$$

In R: `choose(32,23)*0.5^23*(1-0.5)^9`

or `dbinom(x=23, size=32, prob=0.5)`

Example: Estimate a binomial proportion p

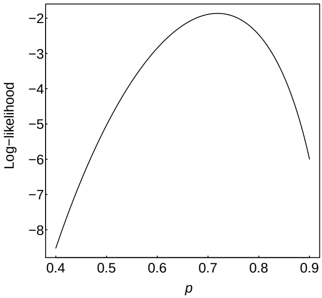
For most likelihood functions, the computed likelihood will be tiny. Thus, it is much easier to work with "log-likelihood"

$$\ln \mathcal{L}[0.5|23 \text{ chose mated female}] = \ln \left[\binom{32}{23} 0.5^{23} (1-0.5)^{9} \right] = -5.03$$

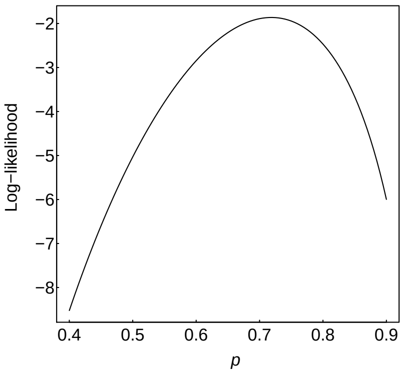
In R: `dbinom(23, 32, prob=0.5, log=TRUE)`

Plot for many values of p to get the log-likelihood curve:

```
curve(dbinom(23, 32, prob=x, log=TRUE), from=0.4, to=0.9)
```



Likelihood works backward from probability



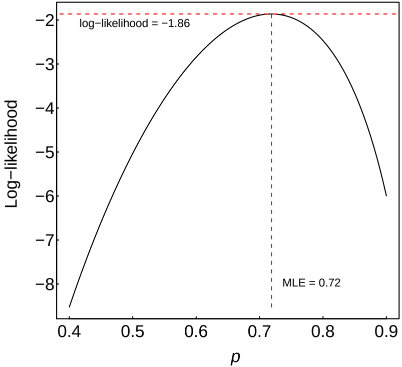
We use likelihood to estimate unknown parameters based on *known data*.

The parameters are treated as variables, the data are a constant, unvarying.

The likelihood function is not a probability distribution.

The population proportion, p, is the variable of the function, but it is not a random variable (its value is not determined by random trial).

Maximum likelihood estimate



The likelihood ratio (difference between the log-likelihoods) measures *relative support* for alternative parameter values.

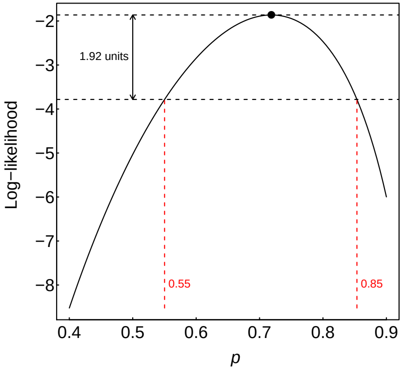
The **maximum likelihood estimate (MLE)** of a parameter is the parameter value having the highest likelihood (and log-likelihood), given the data. This is the parameter value *most strongly supported* by the data.

The ML estimate could instead have been obtained more easily as

$$\frac{Y}{n} = \frac{23}{32} = 0.72$$

The conventional formula for estimating a proportion yields the MLE for the binomial.

Likelihood-based confidence intervals



When estimating a single parameter, an approximate 95% confidence interval is obtained with the values corresponding to 1.92 log-likelihood units below the maximum.

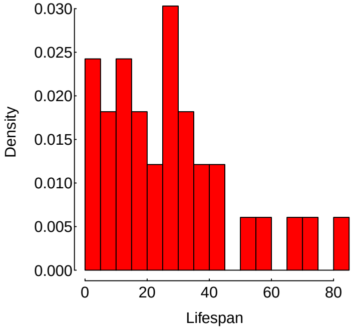
$1.92 = \frac{\chi^2_{0.05,1}}{2}$.
The connection to χ^2 will become apparent later.

So the 95% CI for p in the wasp example is $0.55 \leq p \leq 0.86$.

Note that the 95% likelihood-based confidence interval is not necessarily symmetric about the MLE (unlike the more familiar 95% Wald interval: $\pm 1.96 * se$). In fact, for the binomial case, Wald intervals can sometimes extend outside the interval $[0, 1]$.

Example: Life of bees

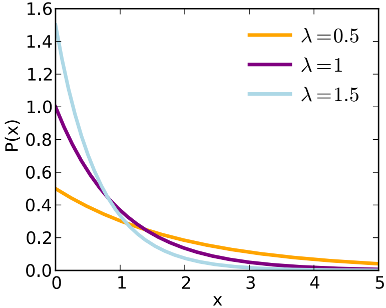
Life spans of individuals in a population are often approximated by an exponential distribution. To estimate the mortality rate of foraging honey bees, P. K. Visscher and R. Dukas (1997, *Insectes Sociaux*), recorded the entire foraging life span of 33 individual worker bees in a local bee population in a natural setting.



Example: Life of bees

Let's assume lifespan is exponentially distributed and use this data to find the mle for the shape of that distribution.

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$



This is a 1 parameter distribution. The mean is $1/\lambda$.

Example: Life of bees

```
head(bees, 10)

   id hours
1   1   7.1
2   2   2.3
3   3   9.6
4   4  25.8
5   5  14.6
6   6  12.8
7   7  20.9
8   8  30.0
9   9  71.1
10 10  36.9
```

Lets calculate the likelihood of $\lambda = 0.1$:

$$\mathcal{L}[\lambda = 0.1|\text{data}] = Pr[\text{data}|\lambda = 0.1]$$

For this data-set, $\mathcal{L}[\lambda = 0.1|\text{data}]$ would equal:

$$Pr[\text{bee}_1 = 7.1|\lambda = 0.1] \times Pr[\text{bee}_2 = 2.3|\lambda = 0.1] \times \cdots$$

which is

$$= 0.1e^{-0.1*7.1} \times 0.1e^{-0.1*2.3} \times \cdots$$

We can convert this to a log-likelihood (so the numbers are not tiny)

$$\ln\left[0.1e^{-0.1\cdot 7.1}\right] + \ln\left[0.1e^{-0.1\cdot 2.3}\right] + \cdots = -167.88$$

We could do the same thing with $\lambda = 0.05$, and we would find

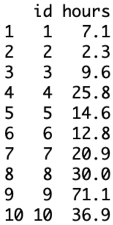
$$\mathcal{L}[\lambda = 0.05|\text{data}] = -144.8092$$

So... $\lambda = 0.05$ has a higher likelihood than $\lambda = 0.1$.

If we try a lot of candidate values, we can eventually find something close to the mle.

Example: Life of bees

`head(bees, 10)`



```
   id hours
1   1   7.1
2   2   2.3
3   3   9.6
4   4  25.8
5   5  14.6
6   6  12.8
7   7  20.9
8   8  30.0
9   9  71.1
10 10  36.9
```

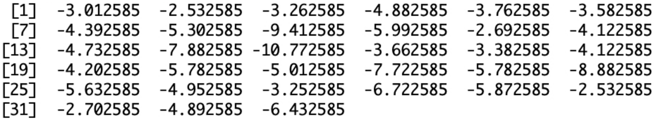
Now, in R:

The function `dexp` calculates the exponential density. For example, `dexp(5, rate=0.1)` $= 0.1e^{-0.1*5}$

and `dexp(5, rate=0.1, log=TRUE)` $= \ln\left[0.1e^{-0.1*5}\right]$

Thus `dexp(bees$hours, rate=0.1, log=TRUE)` calculates each term in our sum for our data-set, for $\lambda = 0.1$.

This yields:



```
 [1]  -3.012585  -2.532585  -3.262585  -4.882585  -3.762585  -3.582585
 [7]  -4.392585  -5.302585  -9.412585  -5.992585  -2.692585  -4.122585
[13]  -4.732585  -7.882585 -10.772585  -3.662585  -3.382585  -4.122585
[19]  -4.202585  -5.782585  -5.012585  -7.722585  -5.782585  -8.882585
[25]  -5.632585  -4.952585  -3.252585  -6.722585  -5.872585  -2.532585
[31]  -2.702585  -4.892585  -6.432585
```

And, finally, `sum` will let us sum these up.
`sum(dexp(bees$hours, rate=0.1, log=TRUE))`
gives us -167.8853.

Now we just do this for a bunch of values of the `rate` argument.

Example: Life of bees

To do this, let's first wrap our previous code into a function. This function will calculate the log-likelihood for a given rate:

```
f.loglik <- function(x) sum(dexp(bees$hours, rate=x, log=TRUE))
```

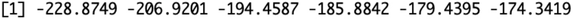
`f.loglik(0.1)` gives us −167.8853

Next, we can make a vector of candidate rates:

```
rates <- seq(from=0.001, to=0.1, length=100)
```

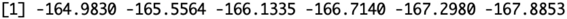
and then we can apply our function to this vector:

```
likelihoods <- sapply(rates, f.loglik)
head(likelihoods)
```



```
[1] -228.8749 -206.9201 -194.4587 -185.8842 -179.4395 -174.3419
```

```
tail(likelihoods)
```



```
[1] -164.9830 -165.5564 -166.1335 -166.7140 -167.2980 -167.8853
```

To get the position of the maximum value, we can use

```
which.max(likelihoods)
```

Thus, we can get our mle with

```
rates[which.max(likelihoods)]
```

which yields 0.036.

A more compact and precise way to do it is to use the `optimize` function.

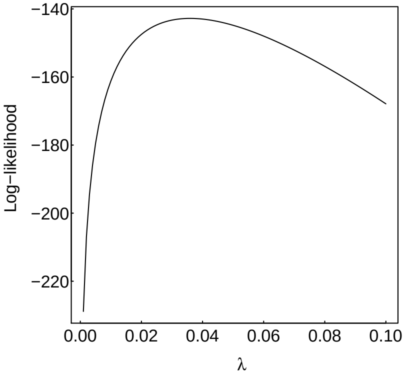
```
mle <- optimize(f.loglik, interval=c(0, 0.1), maximum=TRUE)$maximum
```

which yields 0.03589016.

Example: Life of bees

We can plot the `likelihoods` vector

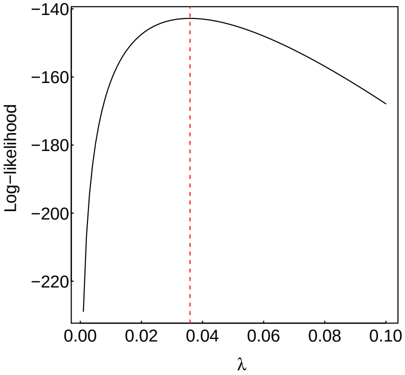
```
plot(x=rates, y=likelihoods, type='l', las=1,
     xlab=expression(lambda), ylab='Log-likelihood')
```



Example: Life of bees

We can plot the `likelihoods` vector

```
plot(x=rates, y=likelihoods, type='l', las=1,
     xlab=expression(lambda), ylab='Log-likelihood')
```



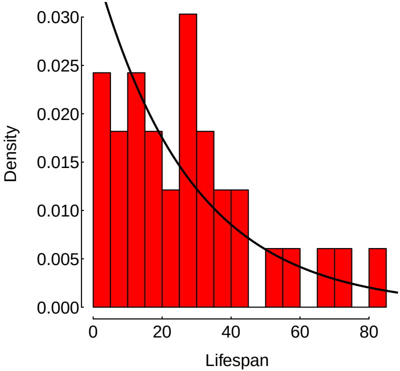
and add a vertical line at the mle

```
abline(v=0.036, lty=2, col='red')
```

Example: Life of bees

Let's next overlay the exponential curve with our estimated mle for the `rate` argument onto our original density plot.

```
hist(bees$hours, prob=TRUE, col='red', las=1, breaks=20,
     xlab='Lifespan', ylab='Density', main='')
hours <- seq(from=0,to=100, length=200)
lines(x=hours, y=dexp(hours, rate=0.036), lwd=2)
```



Example: Life of bees

Next, we will calculate a 95% likelihood-based confidence interval.

Recall that this is the range of values of λ for which the log-likelihood is within 1.92 of the maximum log-likelihood (that is, the log-likelihood for the mle).

`max(likelihoods)` yields our maximum log-likelihood (-142.78).

`ind <- which(max(likelihoods)-likelihoods < 1.92)`
yields a vector containing the positions of the likelihoods that are within 1.92 the max.

We can then look at the rates that gave us these likelihoods:

```
rates[ind]
 [1] 0.026 0.027 0.028 0.029 0.030 0.031 0.032 0.033 0.034 0.035 0.036 0.037
[13] 0.038 0.039 0.040 0.041 0.042 0.043 0.044 0.045 0.046 0.047 0.048 0.049
```

Thus, the 95% confidence interval is given by:

```
range(rates[ind])
[1] 0.026 0.049
```

An alternative that is more precise is to use the `uniroot` function.

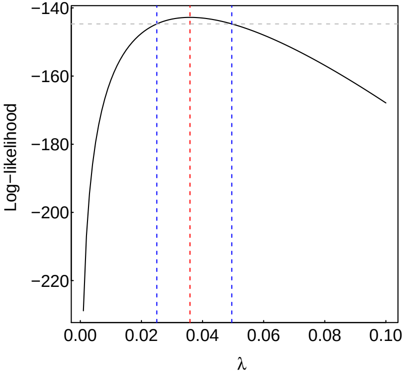
```
f.loglik.shifted <- function(x) f.loglik(x)-(-142.78)+1.92
```

We can then get lower and upper CI limits, respectively, with:

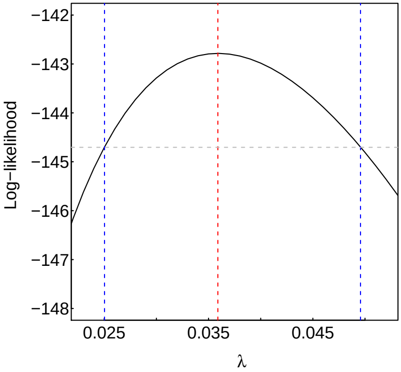
```
uniroot(f.loglik.shifted, lower=0, upper=0.036)$root
uniroot(f.loglik.shifted, lower=0.036, upper=0.1)$root
```

which yields $(0.025, 0.050)$.

We can add these to our log-likelihood plot



We can add these to our log-likelihood plot (zoomed in)



Example: Life of bees

Last step: convert these values into life-span (rather than λ).

Recall that, for the exponential distribution, the mean is $1/\lambda$. Therefore, the mle for age is:

```
1/0.036
```

which equals 27.7.

The 95% confidence interval for age is:

```
c(1/0.05,1/0.025)
```

which equals $(20, 40)$.

Final note: The 95% confidence interval is an approximation based on χ^2. It assumes that sample size is large.

Example: Life of bees

So, why did we do all this?

The beauty of this analysis was that we didn't need any pre-existing statistical models or packages! We just came up with our estimate and confidence interval directly from first principles.

If you have a question for which no existing tools exist, you may be able to answer it by writing your own likelihood function.

Log-likelihood ratio test

Likelihood method to compare the fit of two models to data.

Models must be nested, i.e., one of the models (reduced model) must have a subset of the terms present in the other model (full model).

Tests whether the “full model” fits the data statistically significantly better than a “reduced model”.

Very general method - applies to any type of data, not necessarily normally distributed.

P-value is approximate, but approximation improves with sample size.

Log-likelihood ratio test

$$G = 2 \ln \left[\frac{\mathcal{L}[\text{full model}|\text{data}]}{\mathcal{L}[\text{reduced model}|\text{data}]} \right]$$

G is the log-likelihood ratio test statistic.

Under H_0 (which is typically the reduced model), G is approximately χ^2 distributed.

Degrees of freedom are equal to the difference between the full model and the reduced model in the number of parameters estimated from the data.

Very general method - applies to any data, regardless of distribution from which they came.

The approximation to the χ^2 distribution improves with increasing sample size.

Log-likelihood ratio test

Fatouros et al. (2005, *Nature*), carried out trials to determine whether the wasps can distinguish mated female butterflies from unmated females. In each trial a single wasp was presented with two female cabbage white butterflies, one a virgin female, the other recently mated. **Result:** 23 of 32 wasps tested chose the mated female.

“Reduced” model:
H_0: Wasps choose mated and unmated females with equal probability ($p = 0.5$)

“Full” model: H_A: Wasps prefer one type of female over the other ($p \neq 0.5$)

To fit the full model, p is estimated from the data. In this sense, the full model has 1 more term than the reduced model.

Log-likelihood ratio test

$$G = 2\ln\left[\frac{\mathcal{L}[\text{full model}|\text{data}]}{\mathcal{L}[\text{reduced model}|\text{data}]}\right]$$

Applied to the wasp example:

$$G = 2\ln\left[\frac{\mathcal{L}[p = \hat{p} = 0.72|\text{23 of 32 chose mated female}]}{\mathcal{L}[p = p_0 = 0.50|\text{23 of 32 chose mated female}]}\right]$$

A parameter estimated from the data uses the maximum likelihood estimate (e.g., $\hat{p} = 0.72$ in the full model here).

Log-likelihood ratio test

From calculations using formulae shown earlier:

$$\mathcal{L}[0.72|23 \text{ of } 32 \text{ chose mated female}] = 0.1553$$

$$\mathcal{L}[0.50|23 \text{ of } 32 \text{ chose mated female}] = 0.00653$$

So we can calculate G as:

$$G = 2\ln\left[\frac{0.1553}{0.00653}\right] = 6.336$$

`df=1`, so critical value $\chi^2 = 3.841$.

$6.336 > 3.841$, and so we reject H_0.

The χ^2 value of 1.92 that we used for likelihood-based confidence intervals is half of 3.84.

Log-likelihood ratio test

What is a likelihood-based confidence interval?

A likelihood interval is defined as the set of parameter values with high enough likelihood:

$$\left\{\theta, \frac{\mathcal{L}(\theta)}{\mathcal{L}(\hat{\theta})} > c\right\}$$

for some cutoff point c, where $\mathcal{L}(\theta)/\mathcal{L}(\hat{\theta})$ is the normalized likelihood.

The problem is that c is not meaningful in 'probability' space.

It turns out that, for large samples,

$$2 \ln \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\theta)} \sim \chi^2_1$$

With some algebra, this can all be combined to show that

$$\left\{\theta, \frac{\mathcal{L}(\theta)}{\mathcal{L}(\hat{\theta})} > c\right\}$$

is a $100(1-\alpha)\%$ confidence interval, when we set c to $e^{-\frac{\chi^2_{1,(1-\alpha)}}{2}}$

The chemical that the wasps use to distinguish mated from unmated females is benzyl cyanide, which the male butterfly passes to the female during mating. The compound is an “anti-aphrodisiac”, rendering the mated female less attractive to other male butterflies (Fatouros et al. 2005, *Nature*).