Outline for today

- What is probability
- What is likelihood
- Maximum likelihood estimation
- Example: estimate a proportion
- Likelihood works backward from probability
- Likelihood-based confidence intervals
- Example: estimating speciation and extinction rates
- Log-likelihood ratio test
- Example: test a proportion

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.

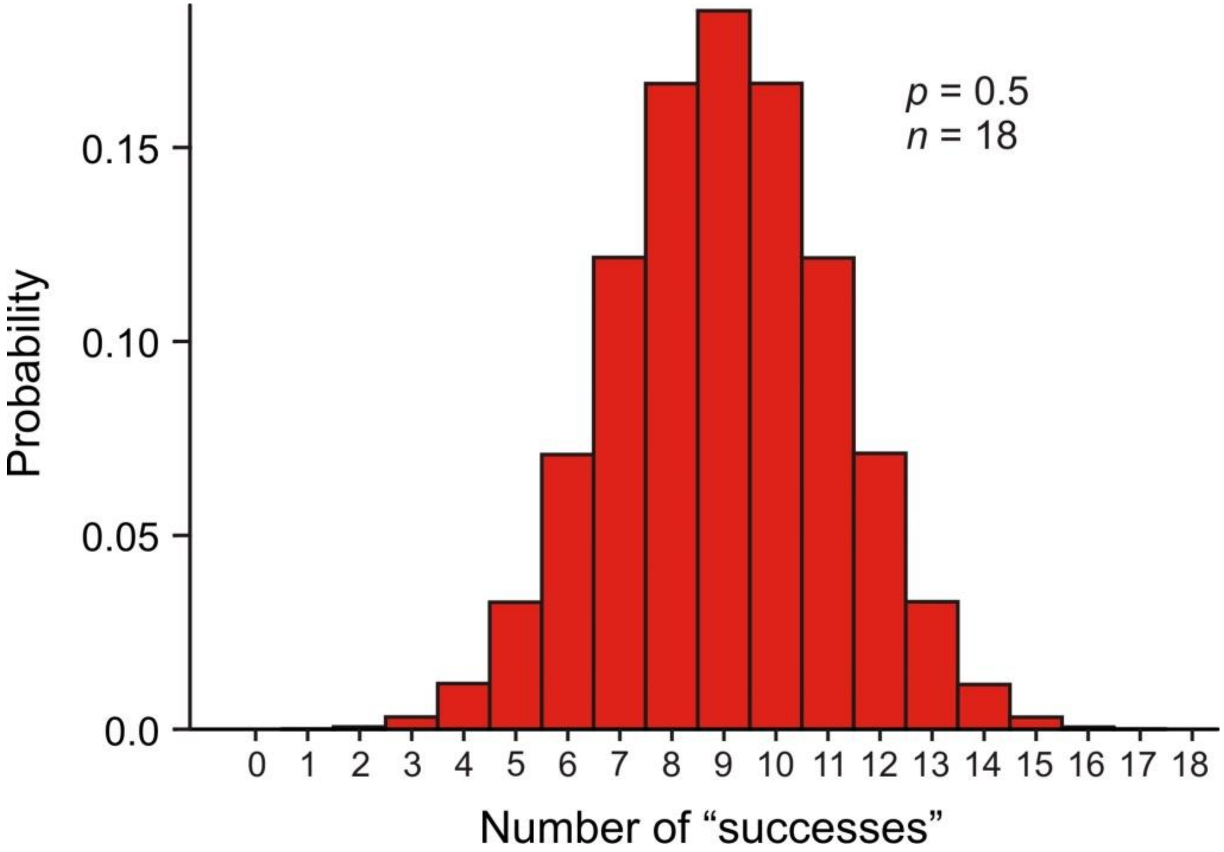
Example: binomial distribution

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.

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

$\binom{n}{Y}$ counts up the different ways of getting *Y* successes and *n* – *Y* failures (e.g., S-S-F; S-F-S; F-S-S)

Graph shows Pr[0], Pr[1] , Pr[2], … when *p* = 0.50 and *n* = 18



What is 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[second child is girl | first child is girl]

Other conditional probabilities:

Pr[we see an elephant today | we are in the Serengeti]

Pr[we see an elephant today | we are in Manhattan]

Pr[12 successes in *n* trials | *p* = 0.50]

Pr[12 successes in *n* trials | *p* = 0.10]

What is 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.

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

Law of Likelihood:

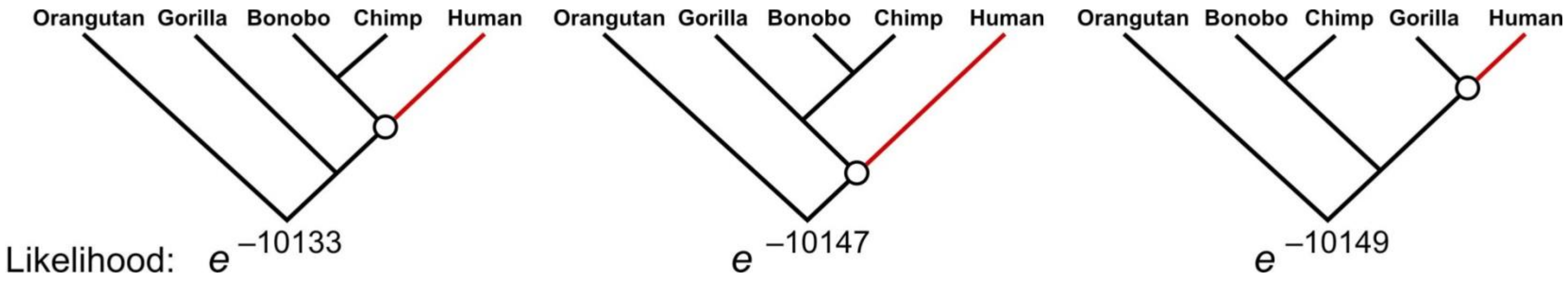
The extent to which data supports one parameter value or hypothesis against another is equal to the ratio of their likelihoods (difference in their log-likelihoods)

Method invented by R. A. Fisher when a 3rd-year undergraduate.

Likelihood is used a lot in phylogeny estimation

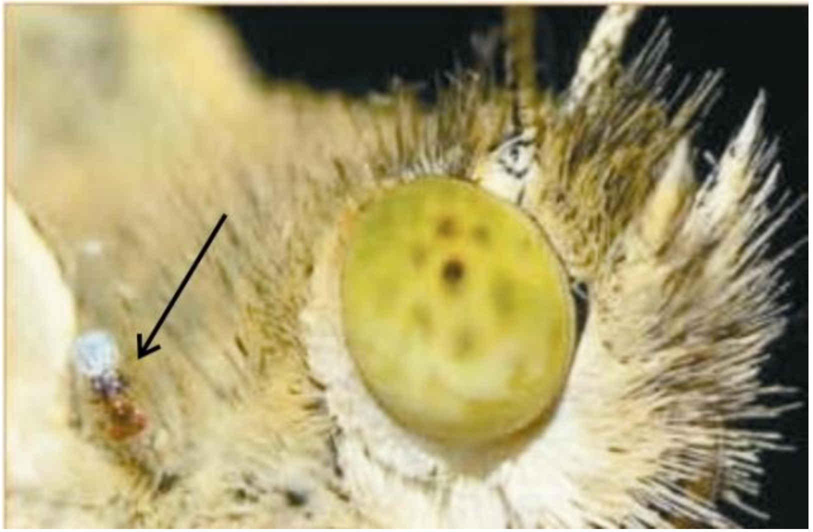
Three proposed trees of ancestor–descendant relationships between humans and the other great apes. The human branch and our shared ancestor with the other apes is highlighted. Numbers at the bottom are the likelihoods of each proposed tree based on gene sequence data (Rannala and Yang 1996). The likelihood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

L[tree = i | gene sequences] = Pr[gene sequences | tree = i]


Orangutan Gorilla Bonobo Chimp Human
Orangutan Gorilla Bonobo Chimp Human
Orangutan Bonobo Chimp Gorilla Human
Likelihood: e^{-10133}
e^{-10147}
e^{-10149}

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

Example 1: 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) 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 1: Estimate a binomial proportion *p*

Likelihood function for the binomial proportion *p*

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

$$L[p \mid Y \text{ chose mated female}] = \Pr[Y \text{ chose mated female} \mid p]$$

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

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

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

in R:

```
dbinom(23, 32, prob=0.5)
[1] 0.00653062
```

Example 1: Estimate a binomial proportion *p*

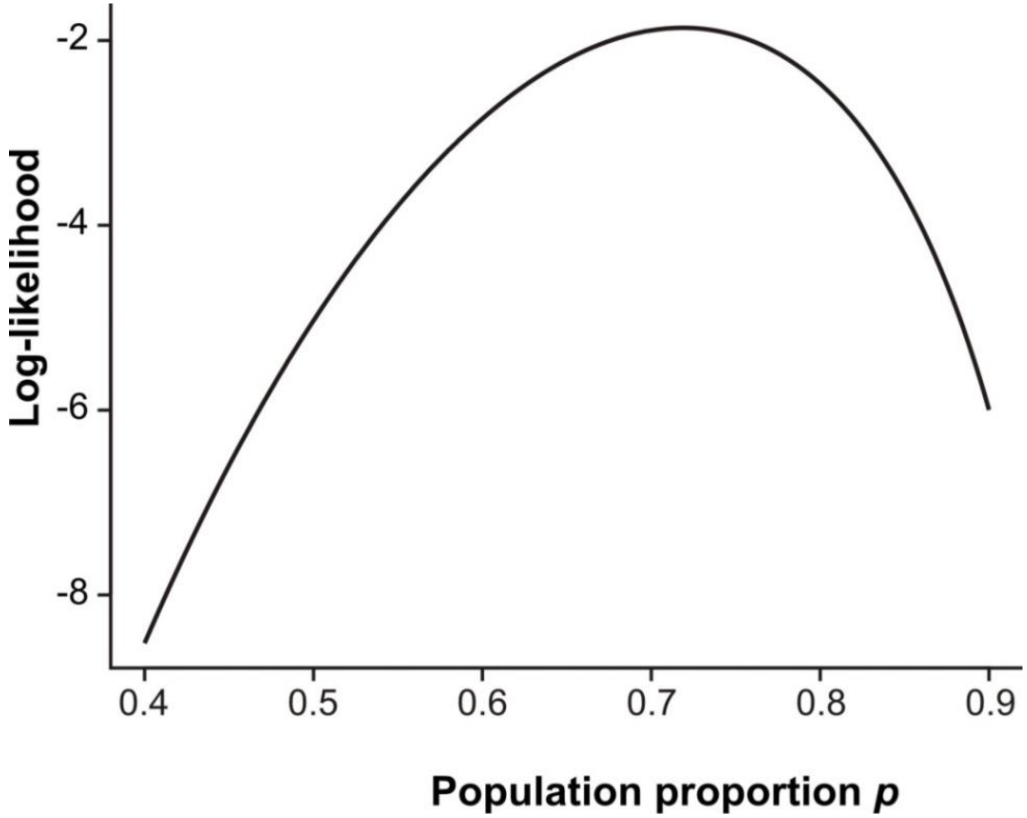
Easier to work with log-likelihoods

$$\ln L[0.5 \mid 23 \text{ chose mated}] = \ln\binom{32}{23} 23 \ln(0.5)\ 9 \ln(1-0.5) = -5.03125$$

in R:

```
dbinom(23, 32, prob = 0.5, log=TRUE)
[1] -5.031253
```

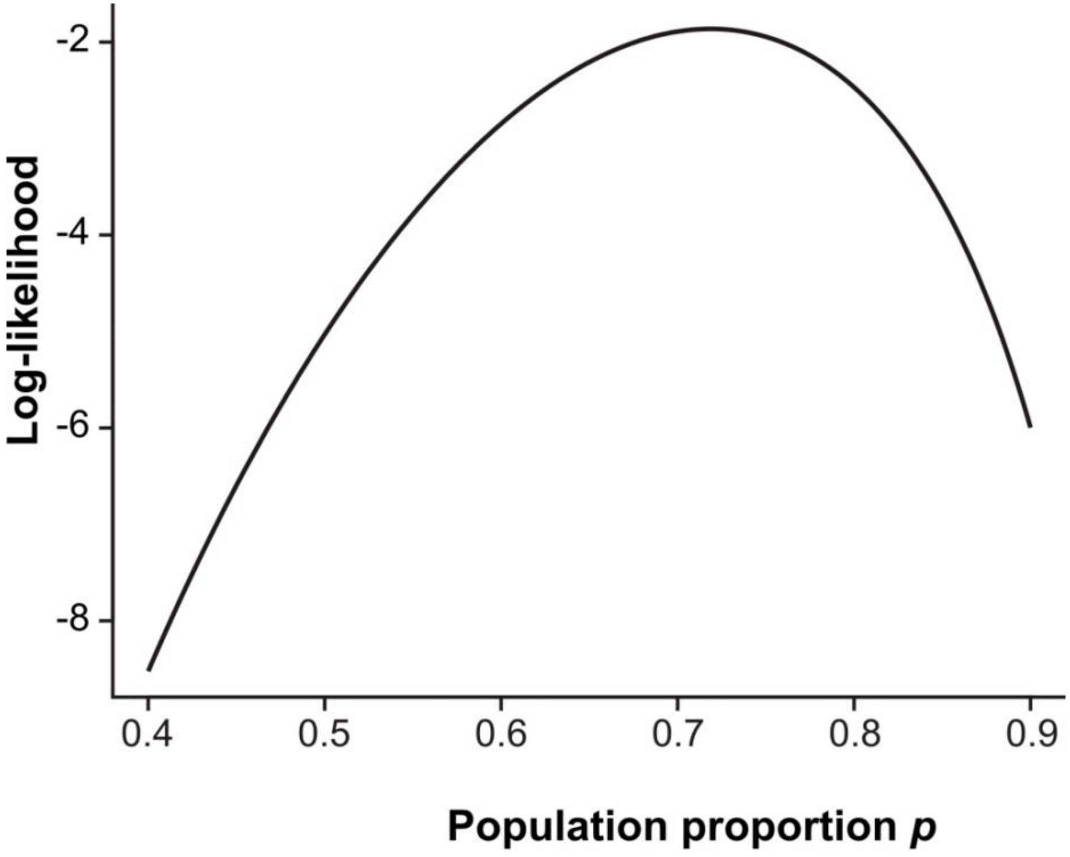
Repeat for many values of *p*
to get the log-likelihood curve:



Likelihood works backward from probability.

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

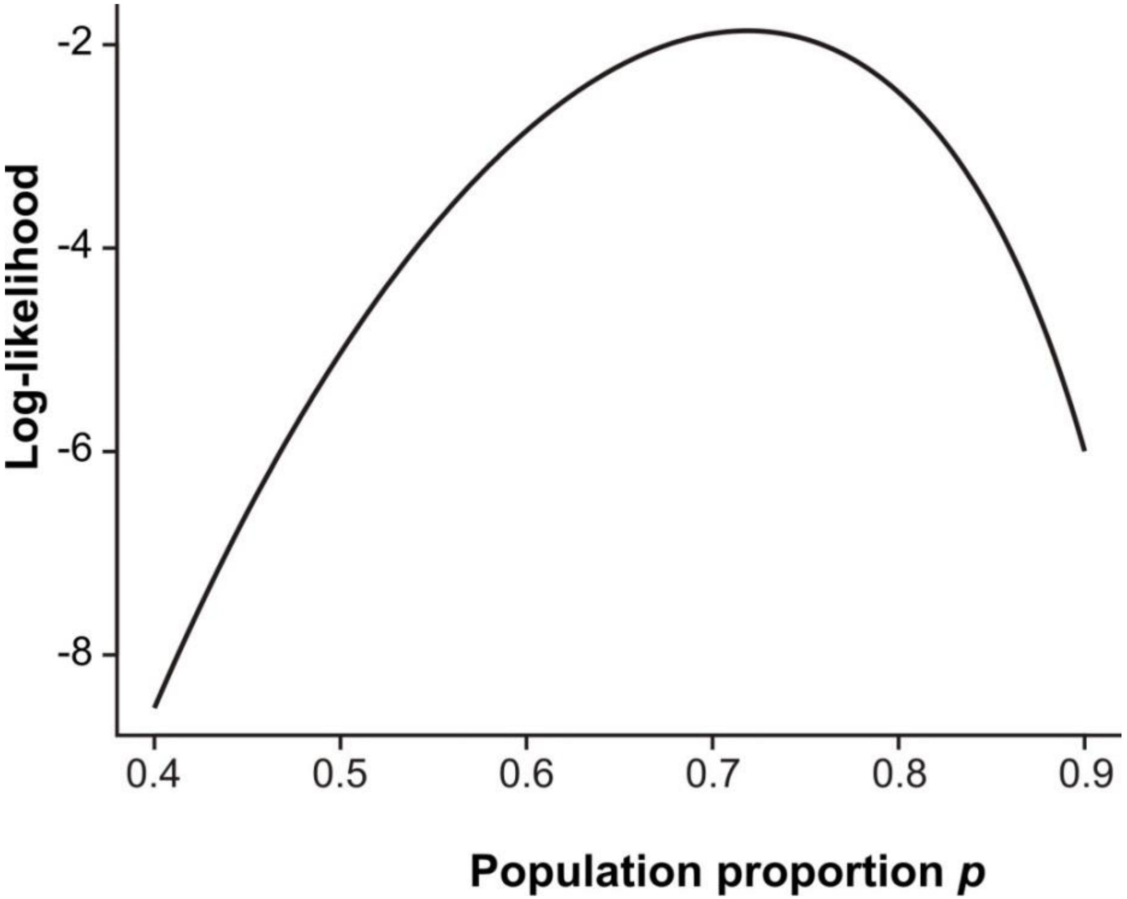
The parameters are treated as variables, the data are a constant, unvarying.



Likelihood works backward from probability.

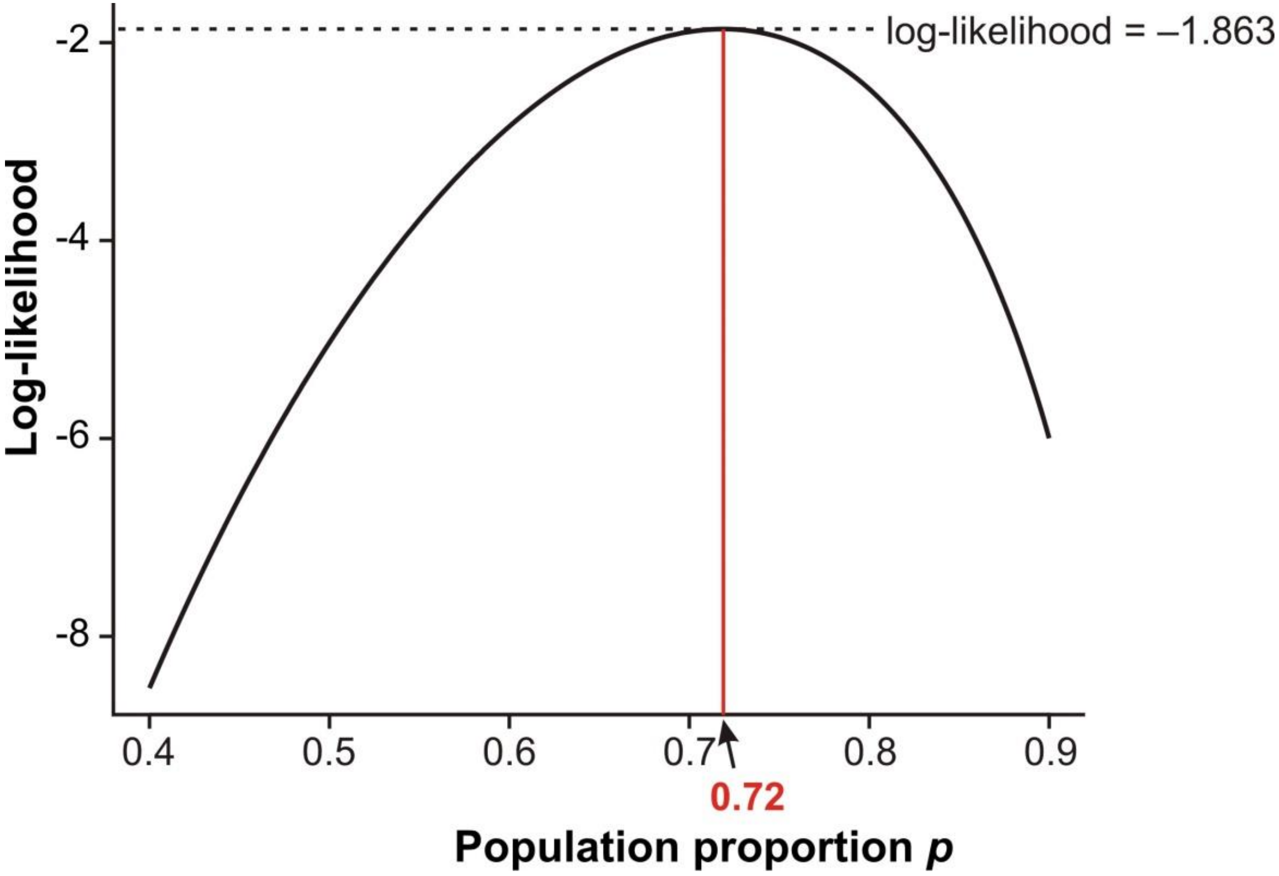
But the likelihood function is not a probability distribution.

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 of 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.



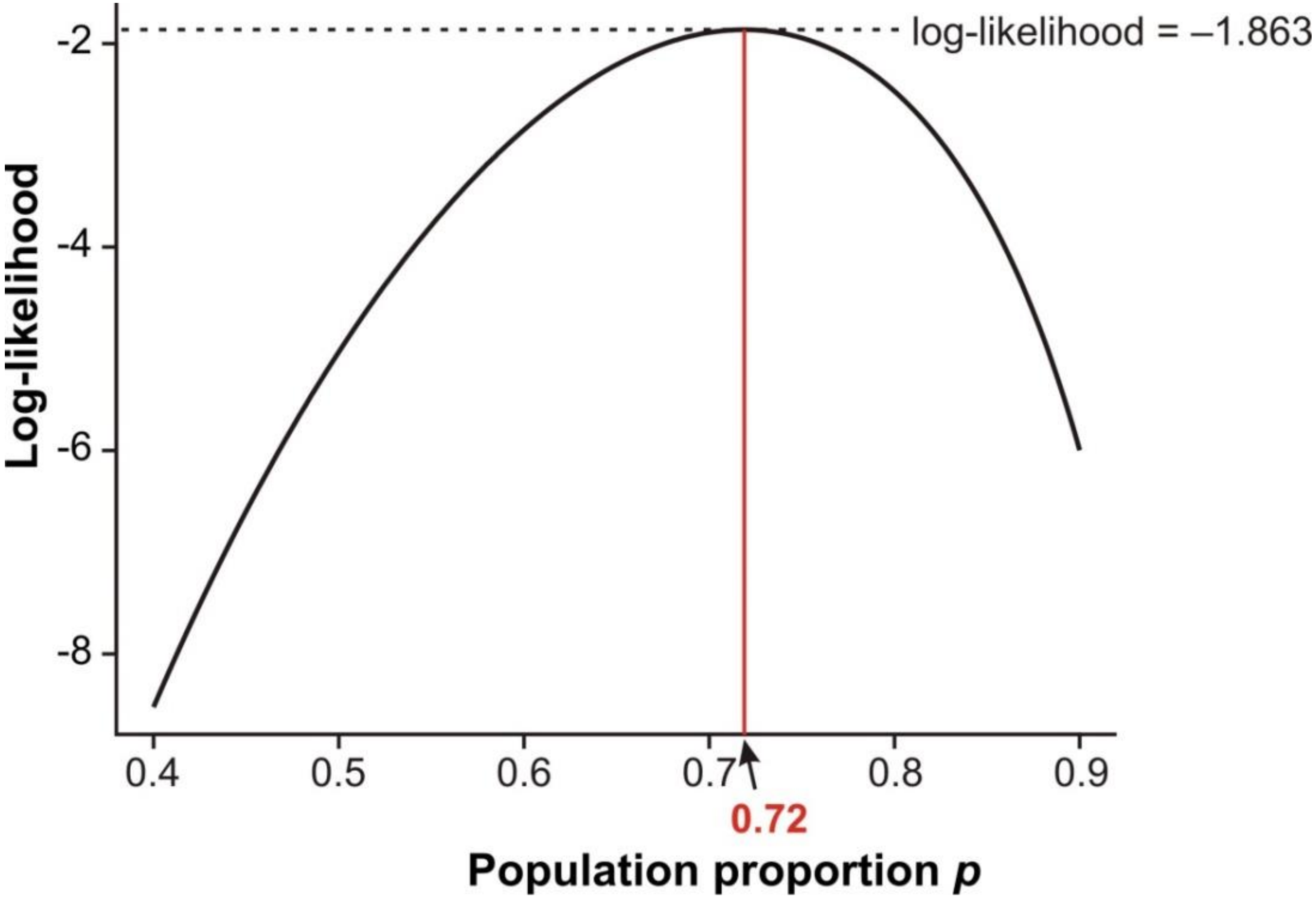
Maximum likelihood estimate

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 ML estimate.

Most formulas you use to estimate parameters yield maximum likelihood estimates.

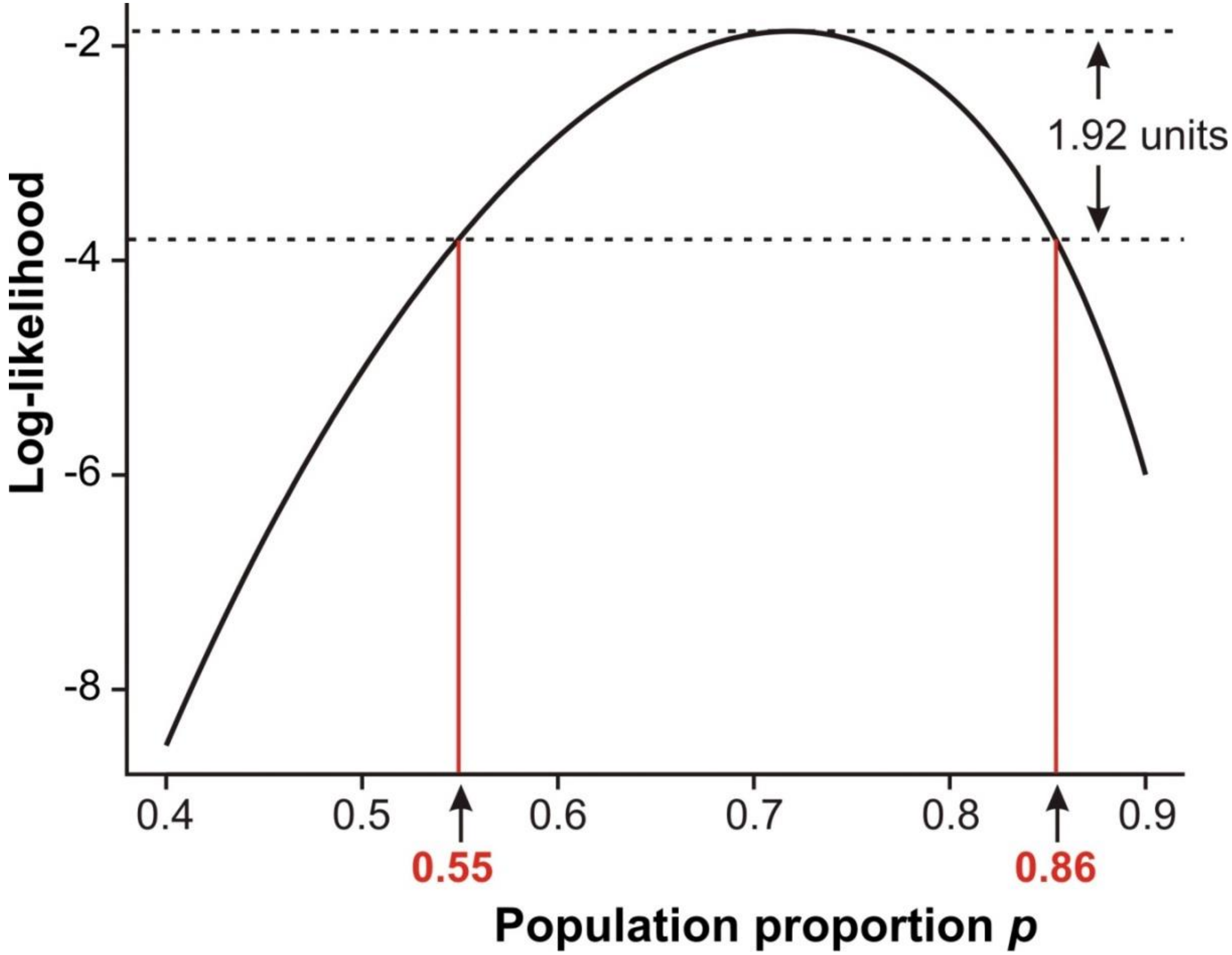


Likelihood-based confidence intervals

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.

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

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



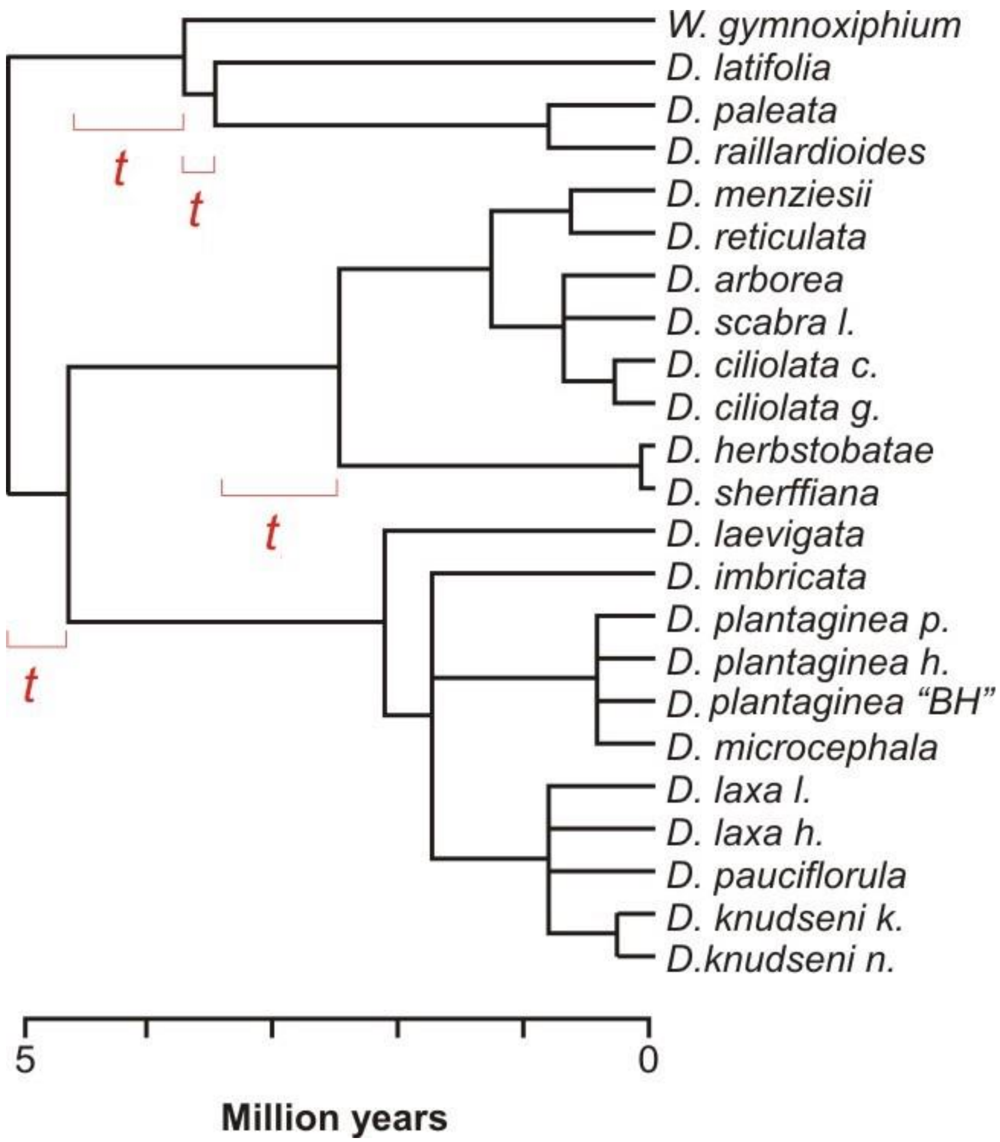
Example 2: Speciation and extinction rates in the Hawaiian silverswords.

You don’t need to be a mathematician to use likelihood in your data analysis. You just need to find a formula for the probability distribution of outcomes for your particular situation. I tried this myself in the following example.

My interest was in using a dated phylogeny to estimate rates of speciation and extinction in the past.

Dubautia scabra

Example 2: Speciation and extinction rates in the Hawaiian silverswords.



λ is the rate at which new species form per lineage per million years.

μ is the rate at which species go extinct per lineage per million years.

Dubautia scabra

Example 2: Speciation and extinction rates in the Hawaiian silverswords.

No simple calculation will give the ML estimates of λ and μ. What to do?

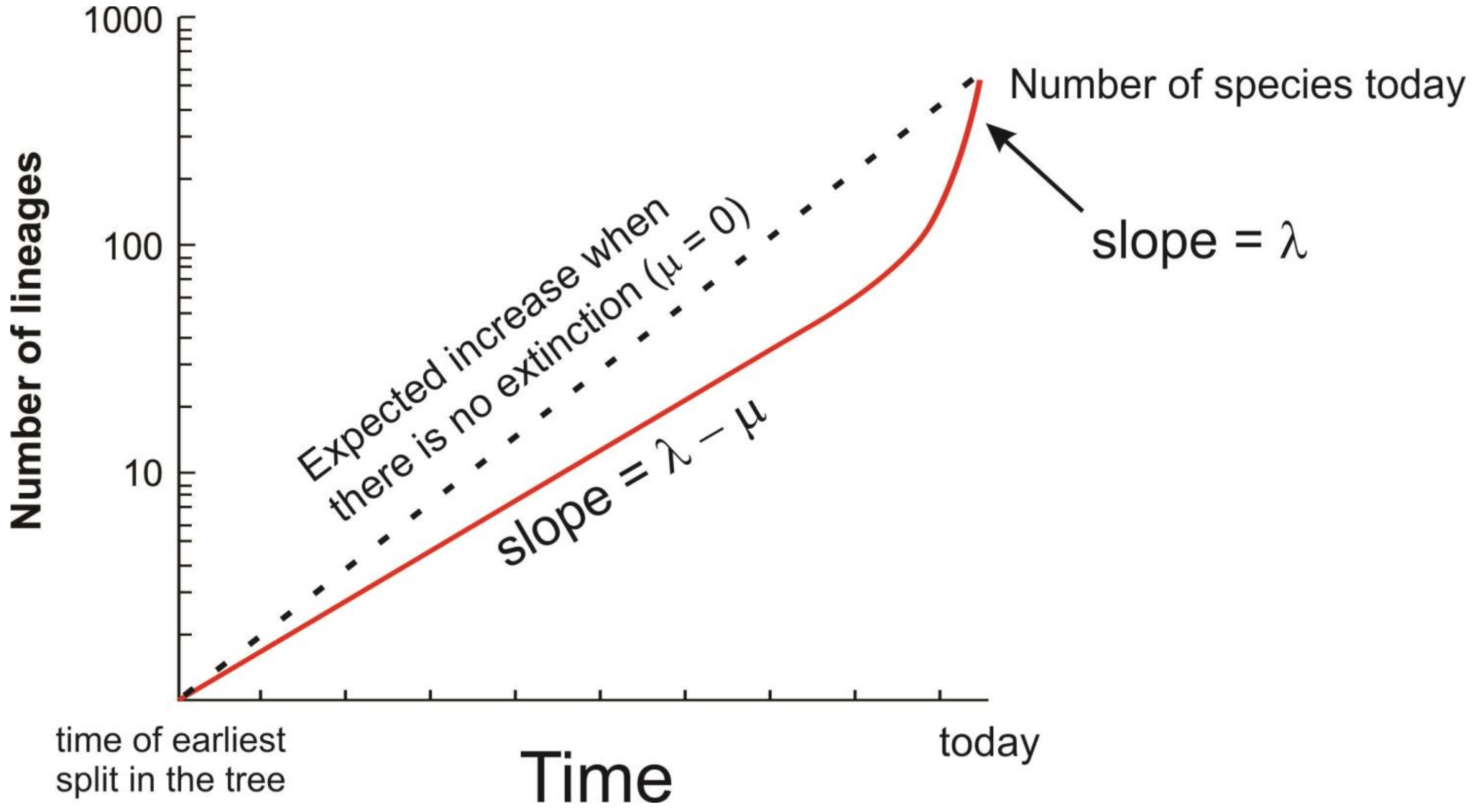
I found a formula for the probability distribution of times between speciation events on reconstructed trees (Nee et al 1994).

I used R to calculate log-likelihoods for a range of possible parameter values (grid search).

Then I made a contour plot to get the log-likelihood surface, the maximum likelihood estimate, and an approximate 95% confidence region.

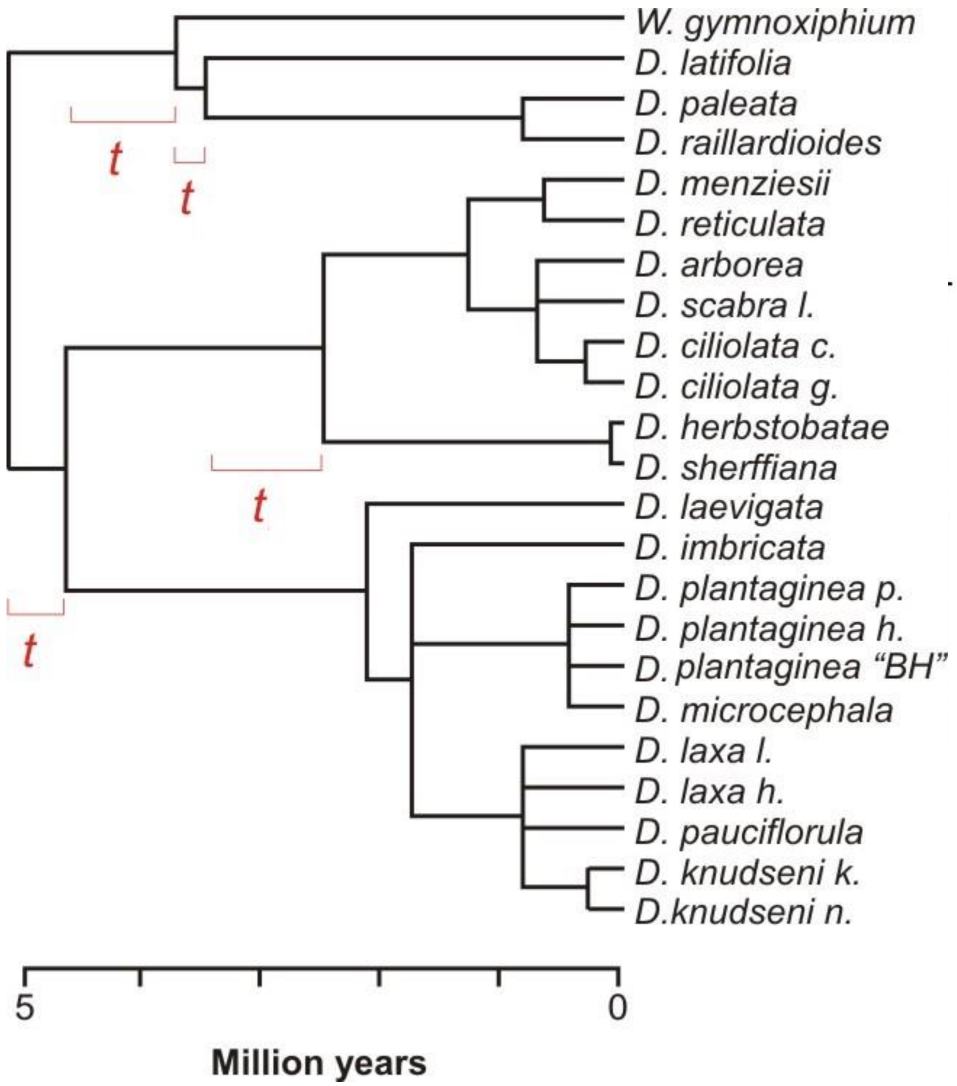
Example 2: Speciation and extinction rates in the Hawaiian silverswords.

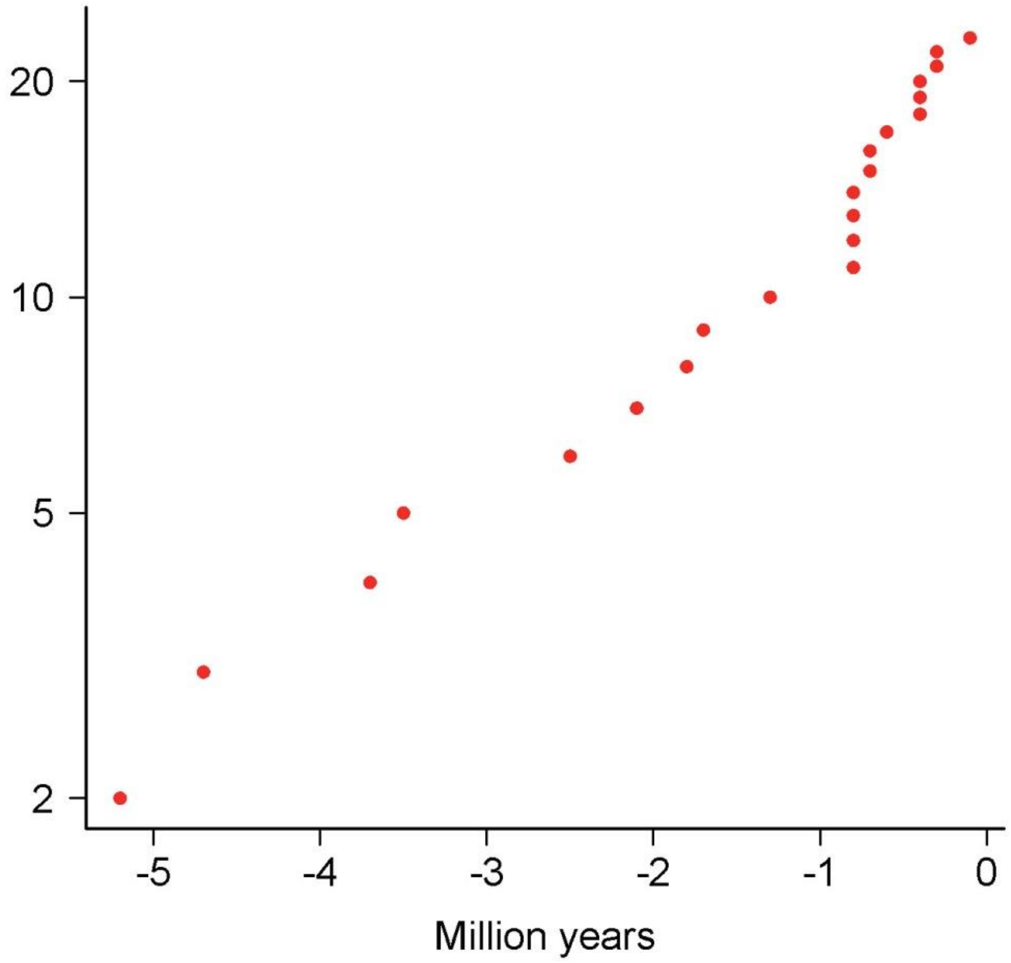
The lineages-through-time curve, which count lineages that survived to the present time, supposedly contains information about speciation and extinction rates, assuming that speciation and extinction can be modeled as a “birth-death process” (a well-understood probability model). Assumes λ and μ are constant through time.



Example 2: Speciation and extinction rates in the Hawaiian silverswords.

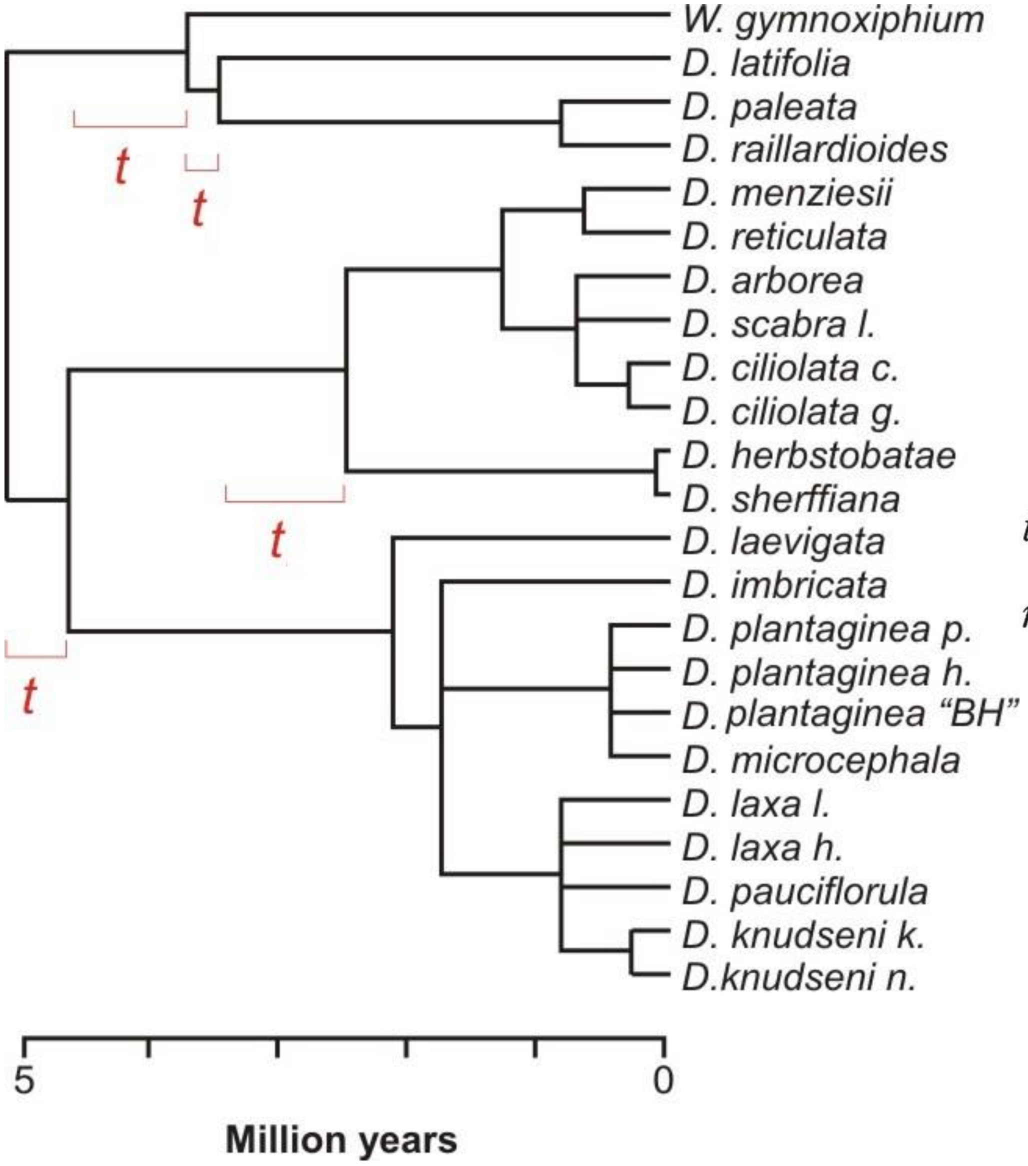
The Hawaiian silversword data:



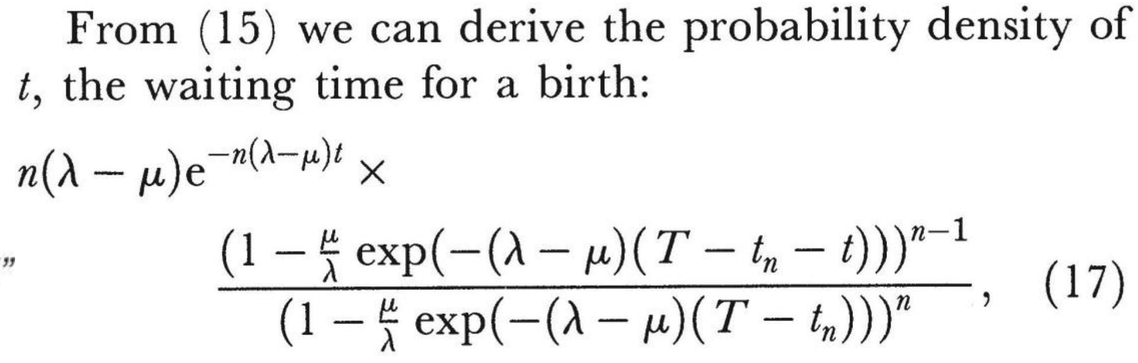


Example 2: Estimating speciation and extinction rates

To estimate, need a formula for the probability distribution of outcomes



Nee et al (1994) found the probability distribution for the waiting time between branching events, P(t | λ, μ), on the "reconstructed" phylogeny of extant species, assuming a constant birth-death process.



From (15) we can derive the probability density of t, the waiting time for a birth:

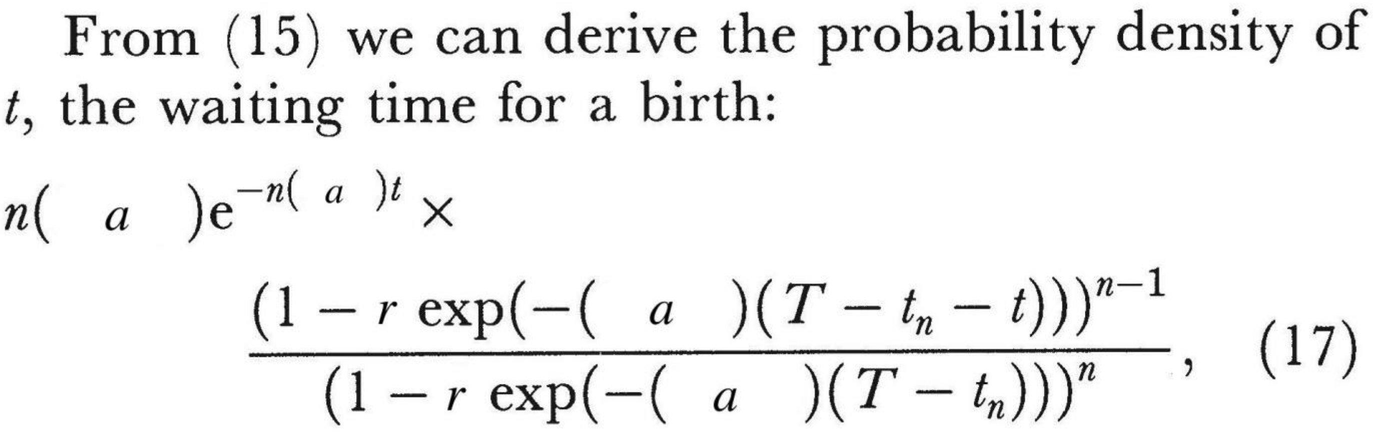
$$n(\lambda-\mu)\mathrm{e}^{-n(\lambda-\mu)t} \times \frac{(1-\frac{\mu}{\lambda}\exp(-(\lambda-\mu)(T-t_n-t)))^{n-1}}{(1-\frac{\mu}{\lambda}\exp(-(\lambda-\mu)(T-t_n)))^{n}}, \quad (17)$$

t is the waiting time to next split in tree (observed data).
λ is the birth (speciation) rate to be estimated.
μ is the death (extinction) rate to be estimated.
n is the number of lineages present at the time.
$T - t_n$ is the time to the present.

Example 2: Estimating speciation and extinction rates

Let *a* = $(\lambda - \mu)$ and *r* $= (\mu/\lambda)$. I carried out a grid search for best *a* and *r*

a	r	loglike
0.1	0.1	-1.937178
0.1	0.2	-0.212507
0.1	0.3	1.662159
0.1	0.4	3.712741
0.1	0.5	5.969948
0.1	0.6	8.46619
0.1	0.7	11.21831
0.1	0.8	14.13799
0.1	0.9	16.39656
0.2	0.1	8.510571
0.2	0.2	9.701548
0.2	0.3	10.94532
0.2	0.4	12.23591
0.2	0.5	13.55520
0.2	0.6	14.85593
0.2	0.7	16.01336
0.2	0.8	16.65578
0.2	0.9	15.30007
0.3	0.1	13.02322
0.3	0.2	13.81377

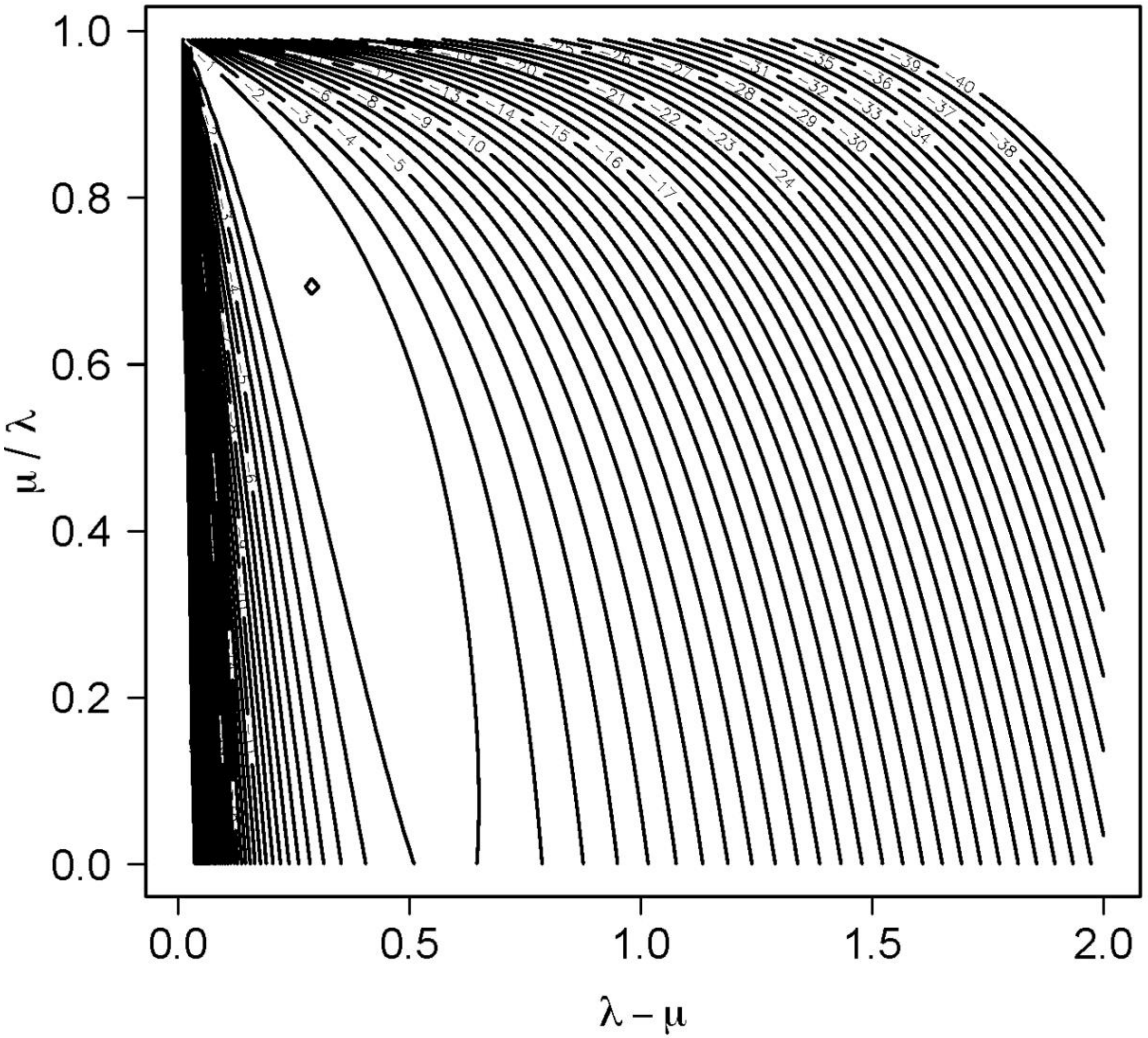


From (15) we can derive the probability density of t, the waiting time for a birth:

$$n(\ a\ )\mathrm{e}^{-n(\ a\ )t} \times \frac{(1 - r\exp(-(\ a\ )(T - t_n - t)))^{n-1}}{(1 - r\exp(-(\ a\ )(T - t_n)))^{n}}, \quad (17)$$

Example 2: Estimating speciation and extinction rates

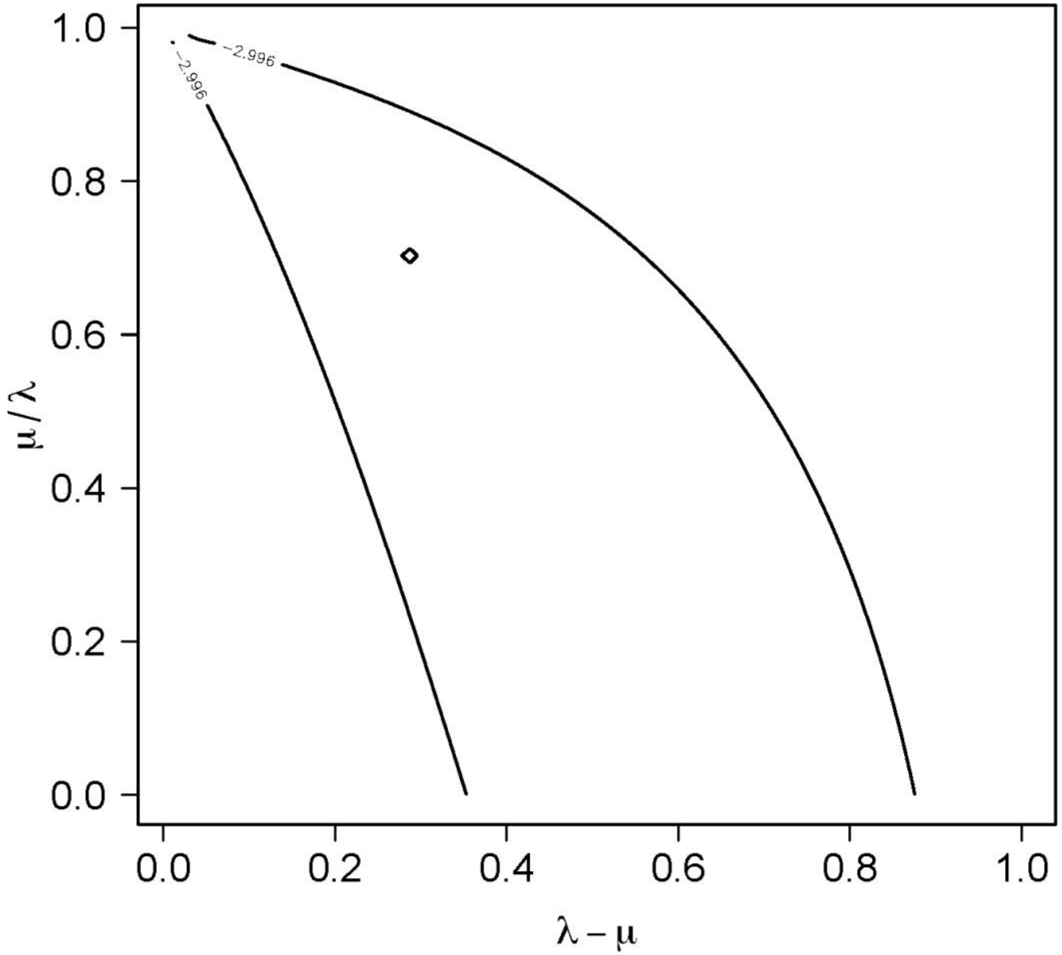
Contour plot of the log-likelihood surface



Example 2: Estimating speciation and extinction rates

Likelihood-based 95% confidence region

When estimating two parameters jointly, an approximate 95% confidence region along any one axis is obtained by the values corresponding to 2.996 log-likelihood units below the maximum



2.996 is $\chi^2_{0.05,2}/2$

Very few silversword species in phylogeny, so the confidence limits are wide for this example.

Example 2: Estimating speciation and extinction rates

Likelihood-based 95% confidence region

Final note:

The 95% confidence interval is an approximation based on χ^2. It assumes that sample size is large (not true in the silverswords, because there aren’t very many species).

I could use simulation in R to improve accuracy of confidence region, but did not do this for the present example.

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

$$G = 2\ln\frac{L[\text{full model} \mid \text{data}]}{L[\text{reduced model} \mid \text{data}]}$$

G is the log-likelihood ratio test statistic.

Under H_0, *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.

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

Log-likelihood ratio test

Example 1: Fatouros et al. (2005) 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. Results: 23 successes out of n = 32 trials.

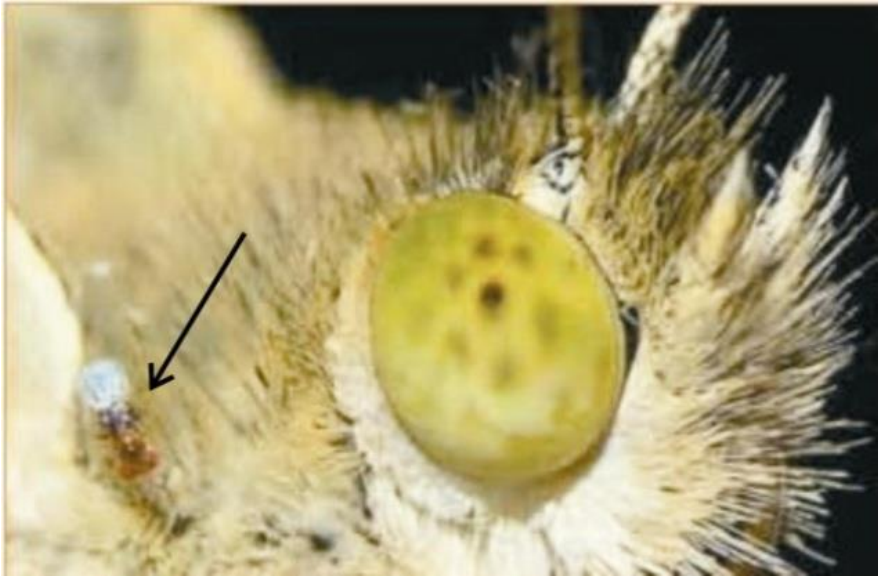
“Reduced” model:

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.



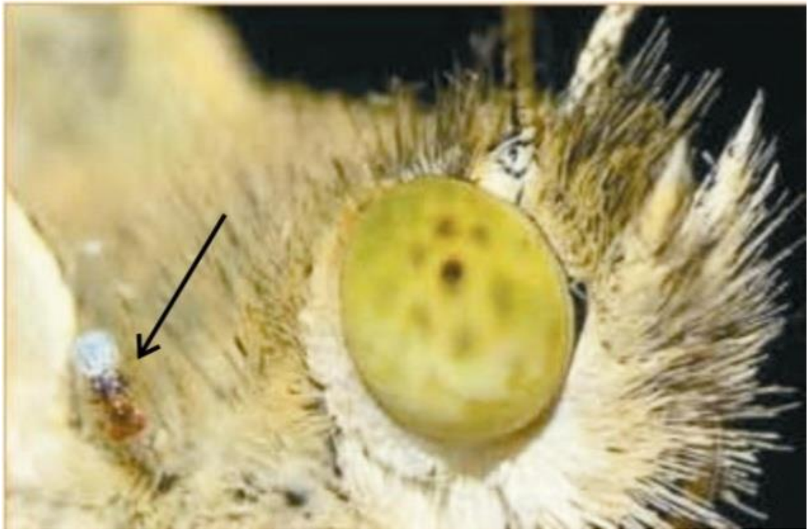
Log-likelihood ratio test

$$G = 2\ln\frac{L[\text{full model} \mid \text{data}]}{L[\text{reduced model} \mid \text{data}]}$$

Applied to the wasp example:

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

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,

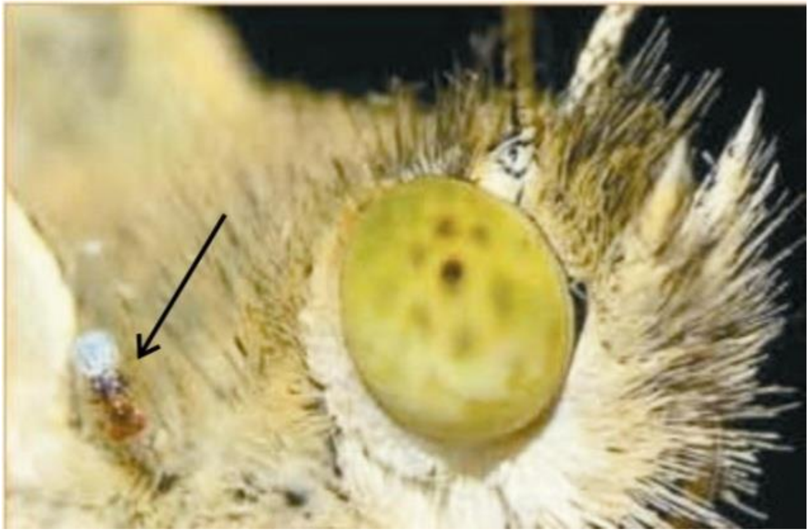
$$L[0.72 \mid 23 \text{ of } 32 \text{ chose mated female}] = 0.1553$$

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

$$G = 2 \ln \frac{0.1553}{0.00653} = 6.336$$

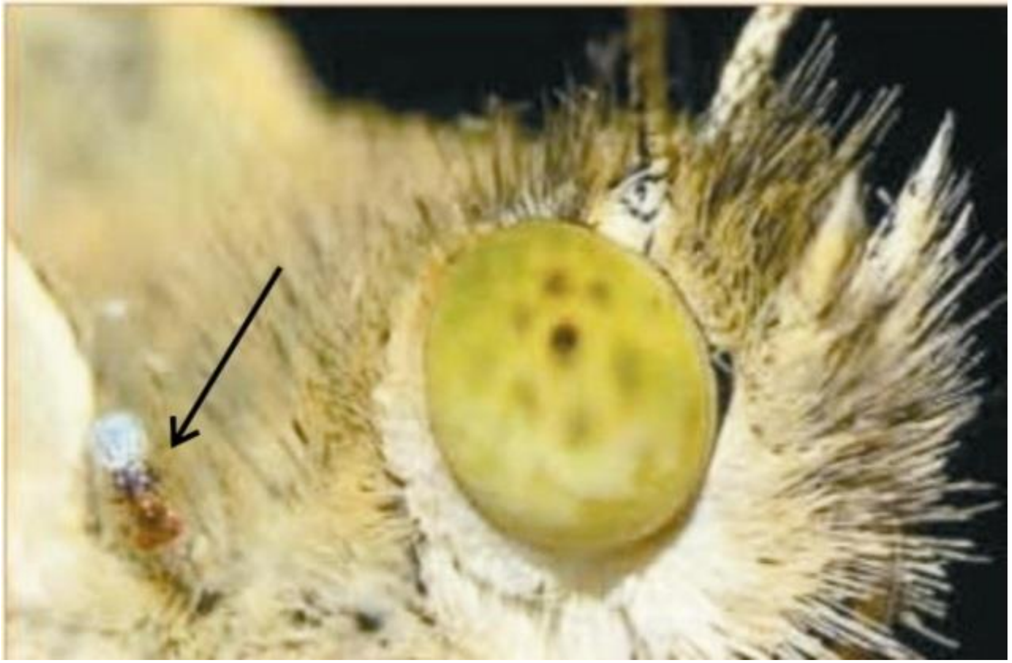
df = 1, so the critical value $\chi^2 = 3.841$

Since 6.336 > 3.841, we reject H_0.



Wasp example: finale

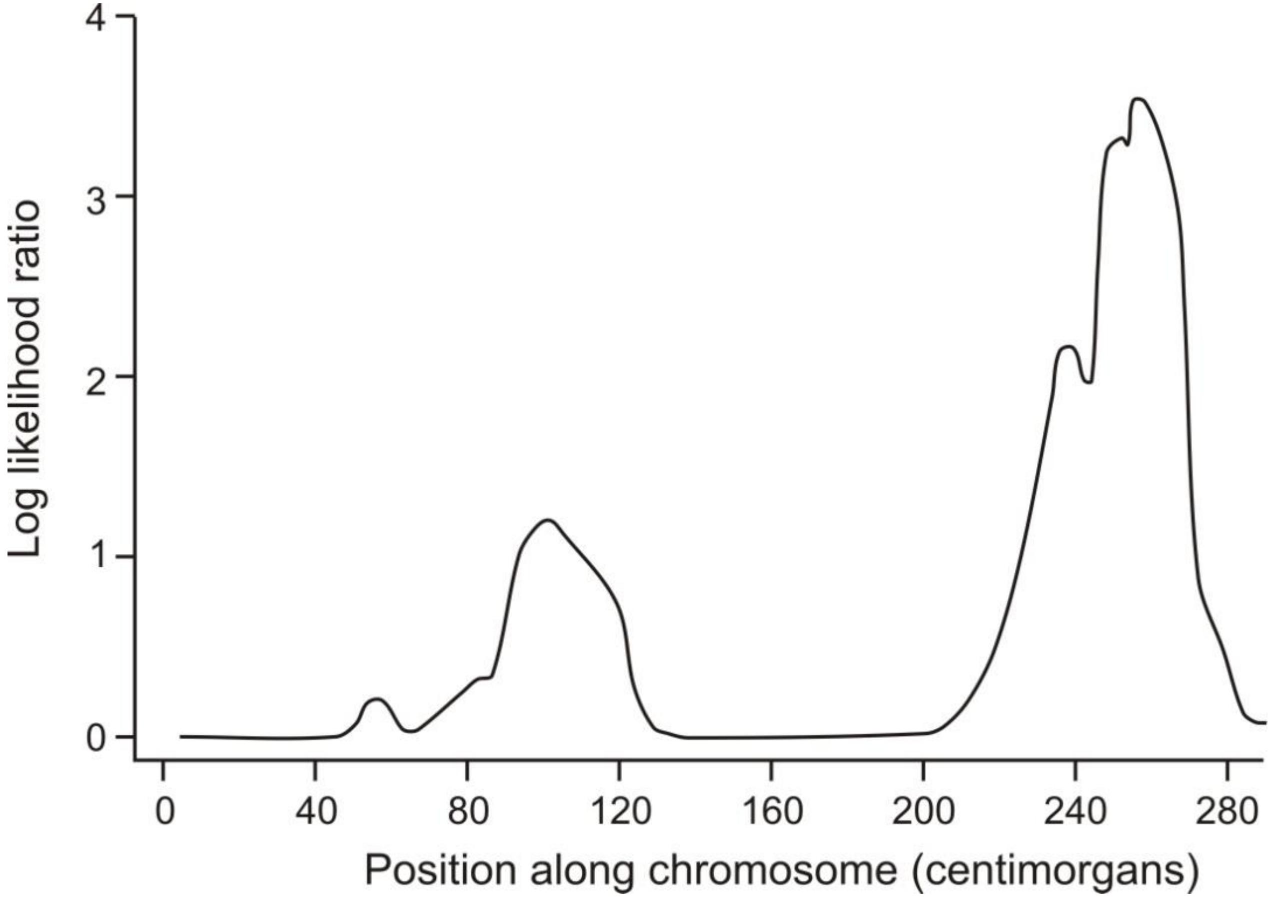
The chemical that the wasps use to distinguish mated from unmated females is benzyl cyanide, which the male butterfly passes to the female during mating. The compound is an “anti-aphrodisiac”, rendering the mated female less attractive to other male butterflies (Fatouros *et al*. 2005).



Log-likelihood ratios are used in gene mapping

At each marker along the chromosome two models are fitted to data on healthy and diseased individuals. The “reduced model” assumes that the frequency of healthy and diseased individuals is the same for every genotype. The “full model” assumes that some genotypes are associated with a higher frequency of diseased individuals than other genotypes. The log of the ratio of the likelihoods of the two models (“full” divided by “reduced”) is called the LOD score, and is a measure of the strength of evidence for a causative mutation near the marker.

Evidence for a gene affecting schizo-affective disorder on human chromosome 1.



Next discussion paper:

Verhoeven et al (2005) Controlling false discovery rate when multiple testing. Oikos.

Download from “**Handouts**” tab on course web site.

Presenters: ____________ and ______________

Moderators: ____________ and ______________