Compare models with known likelihood function: a workflow

Research question

- Does it matter to use model A or B in parameter estimation from aligned molecular data?
- Prerequisites:
  - The likelihood function of both models must be known (and be put in BEAST2)
  - The creation of null phylogenies of at least one model must be known
  - The parameters of both models can be estimated from a phylogeny

Models

Model	Birth death	Coalescent Constant Population	Coalescent Exponential Population
Parameter #1	Speciation rate	Number of species[1]	Number of species[1]
Parameter #2	Extinction rate	-	Growth rate[2]

[1] Called 'Population size' in BEAST2

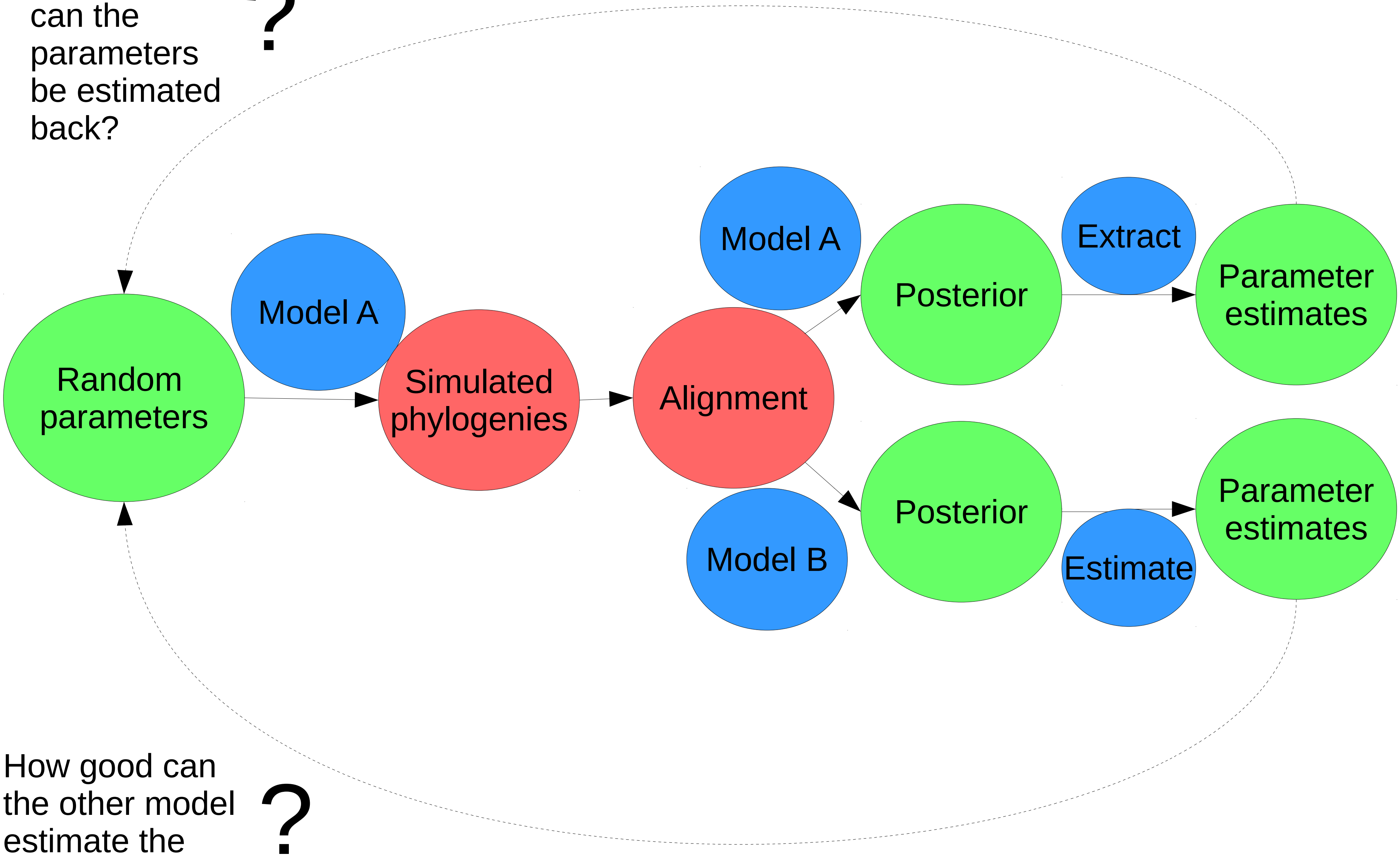
[2] Growth rate = speciation rate – extinction rate

Workflow

1. Using model A:
   - 1.1 create random parameter values
   - 1.2 with those values, create simulated phylogeny
   - 1.3 from that phylogeny, create simulated aligned molecular data
2. Using model A and B:
   - 2.1 create posterior[1] from that data using BEAST2
   - 2.2 check posterior using Tracer for convergence
   - 2.3 for model A: obtain parameter distribution from posterior
   - 2.4 for model B: estimate parameters used by model A
   - 2.5 Does model A give a higher likelihood?
   - ?OR: USE BAYES FACTOR!
3. Do the same vice versa
4. Statistical analysis

[1] posterior = phylogenies + parameter estimates

Does model matter?



How good can the parameters be estimated back? ?

How good can the other model estimate the parameters? ?

1.1. Create random parameter values

- Different models have a different number of parameters

Model	Birth death	Coalescent Constant Population	Coalescent Exponential Population
Parameter #1	Speciation rate	Number of species[1]	Number of species[1]
Parameter #1	Extinction rate		Growth rate[2]

1: Called 'Population size' in BEAST2

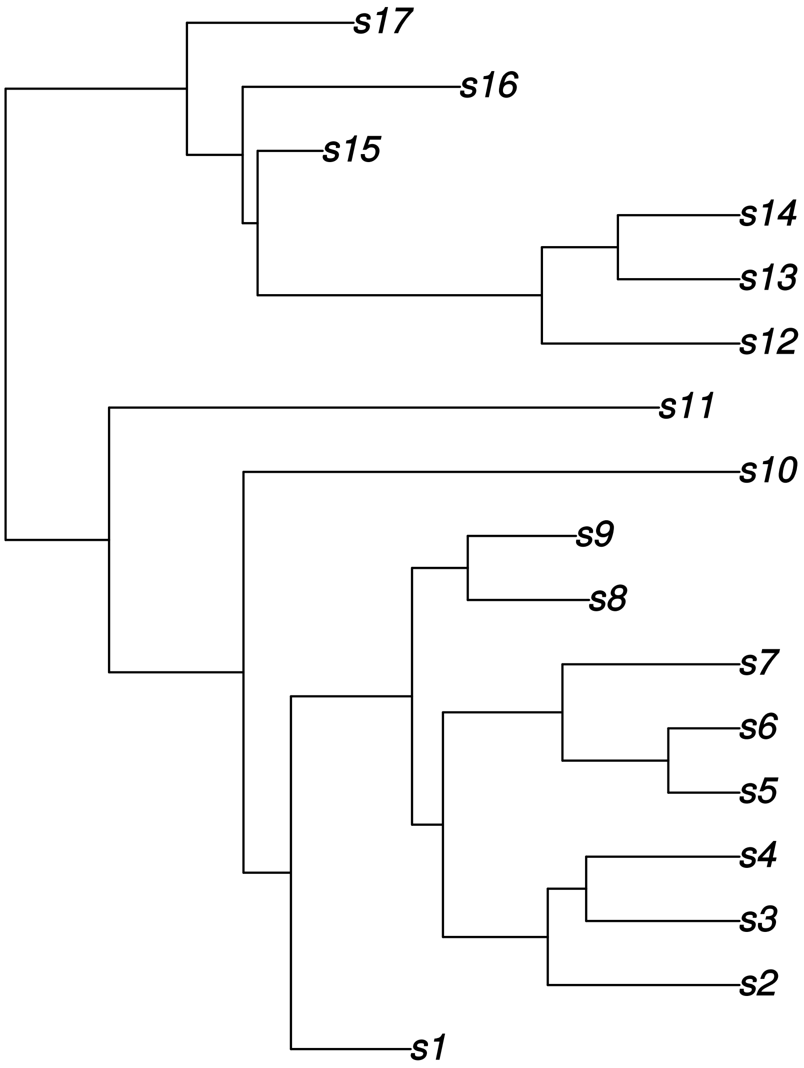
2: Growth rate = speciation rate – extinction rate

1.2. Generating random phylogenies

- Can be done in R:

```
library(geiger);
p = sim.bdtree(
  birth_rate,
  death_rate
  n_taxa)
plot(p)
```

- Random birth-death tree



1.3. Create simulated aligned molecular data

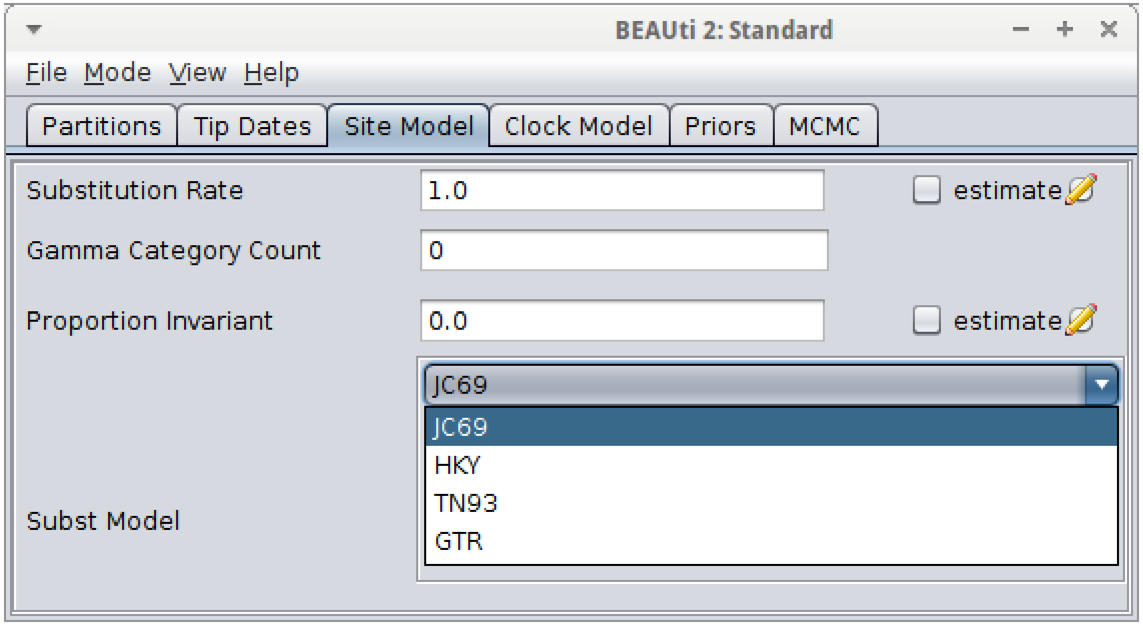
- Can be done in R:

```
library(phangorn);
sequence_length <- 10
data <- simSeq(phylogeny,
  l=sequence_length)
write.phyDat(data,
  file="t.nexus",
  format="nexus")
```

Or other formats

2.1. Create posterior from that alignments for different models using BEAST2

- BEAST2: platform for Bayesian evolutionary analysis
- FOSS (LGPL 2.1)

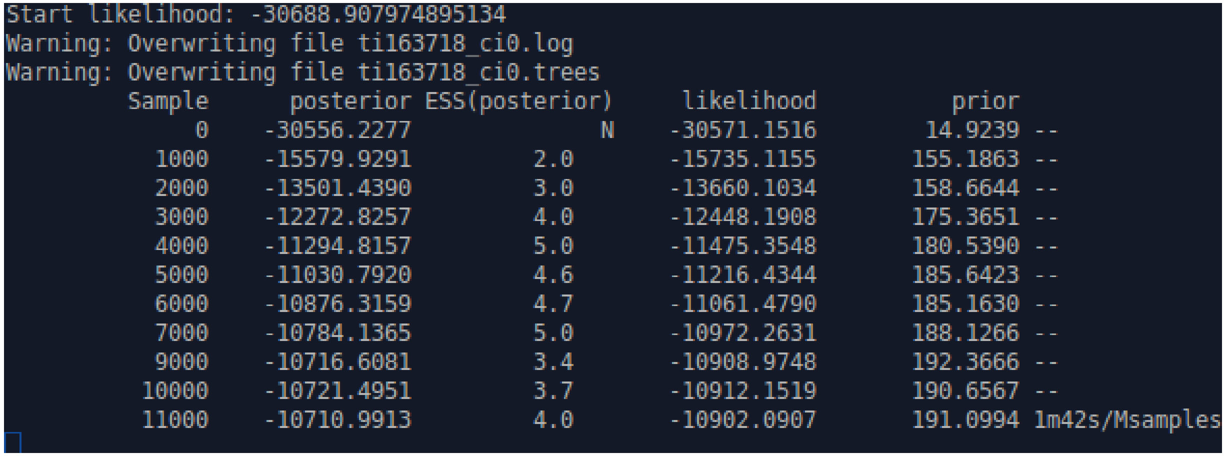


Substitution Rate 1.0 □ estimate

Gamma Category Count 0

Proportion Invariant 0.0 □ estimate

Subst Model JC69

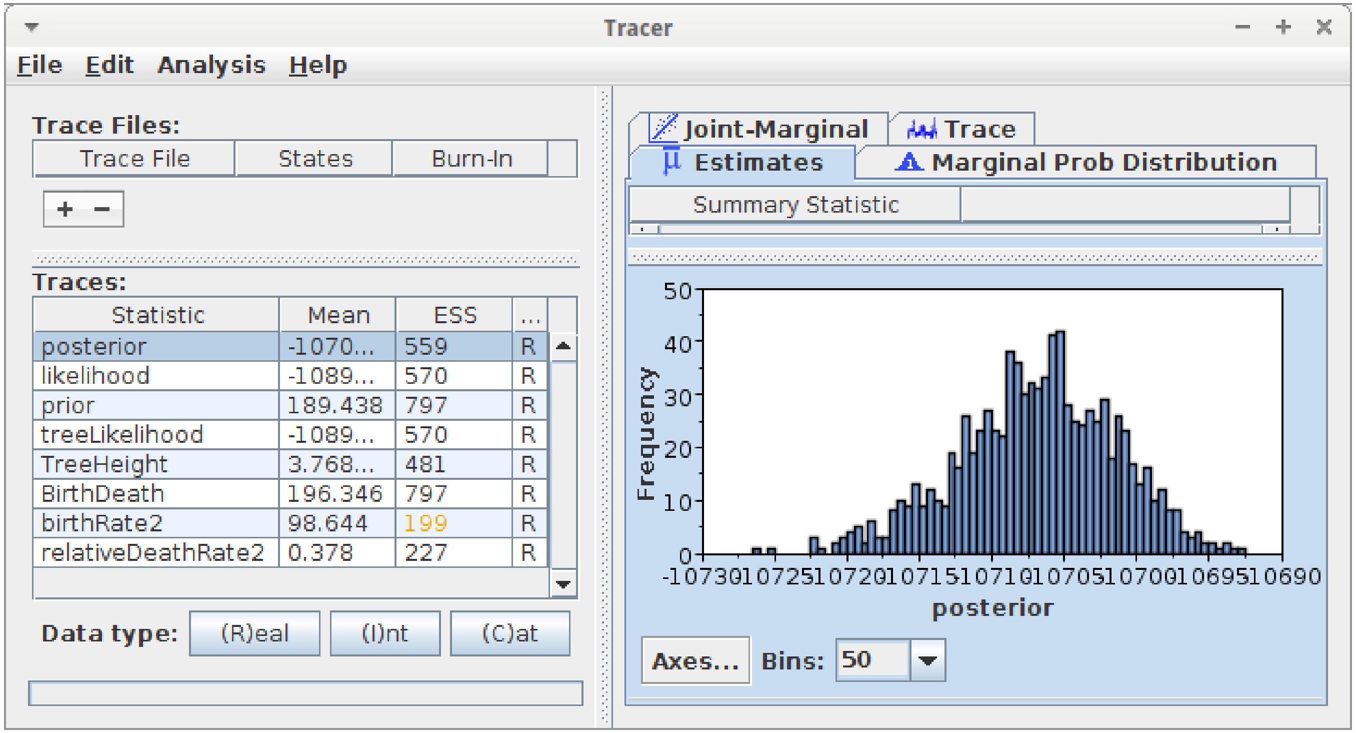


```
Start likelihood: -30688.907974895134
Warning: Overwriting file ti163718_ci0.log
Warning: Overwriting file ti163718_ci0.trees
         Sample      posterior ESS(posterior)     likelihood          prior
              0    -30556.2277              N    -30571.1516        14.9239 --
           1000    -15579.9291            2.0    -15735.1155       155.1863 --
           2000    -13501.4390            3.0    -13660.1034       158.6644 --
           3000    -12272.8257            4.0    -12448.1908       175.3651 --
           4000    -11294.8157            5.0    -11475.3548       180.5390 --
           5000    -11030.7920            4.6    -11216.4344       185.6423 --
           6000    -10876.3159            4.7    -11061.4790       185.1630 --
           7000    -10784.1365            5.0    -10972.2631       188.1266 --
           9000    -10716.6081            3.4    -10908.9748       192.3666 --
          10000    -10721.4951            3.7    -10912.1519       190.6567 --
          11000    -10710.9913            4.0    -10902.0907       191.0994 1m42s/Msamples
```

R Bouckaert R et al., 2014

2.2. Check posterior using Tracer

- No burn-in visible
- ESS must be in range 700-800
- All parameters follow a smooth distribution



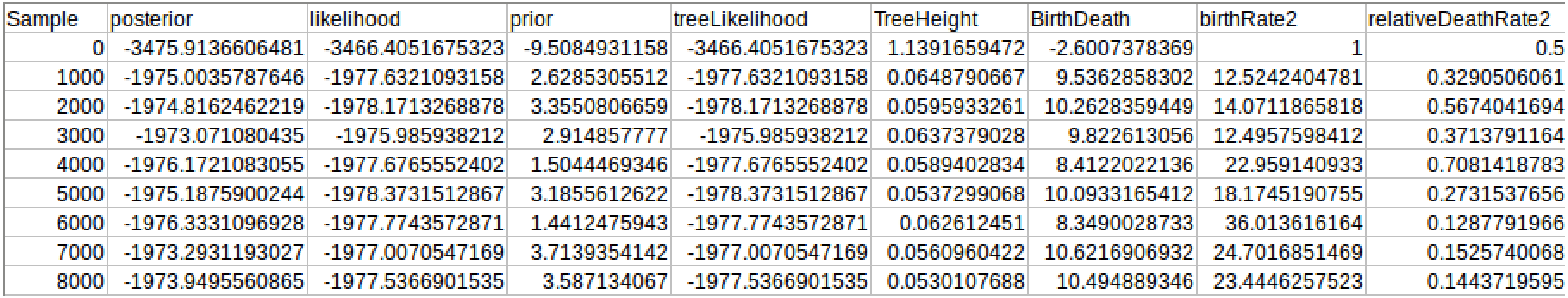
Traces:

Statistic	Mean	ESS	...
posterior	-1070...	559	R
likelihood	-1089...	570	R
prior	189.438	797	R
treeLikelihood	-1089...	570	R
TreeHeight	3.768...	481	R
BirthDeath	196.346	797	R
birthRate2	98.644	199	R
relativeDeathRate2	0.378	227	R

Drummond & Bouckaert, 2015

2.3. For model A: obtain parameter distribution from posterior

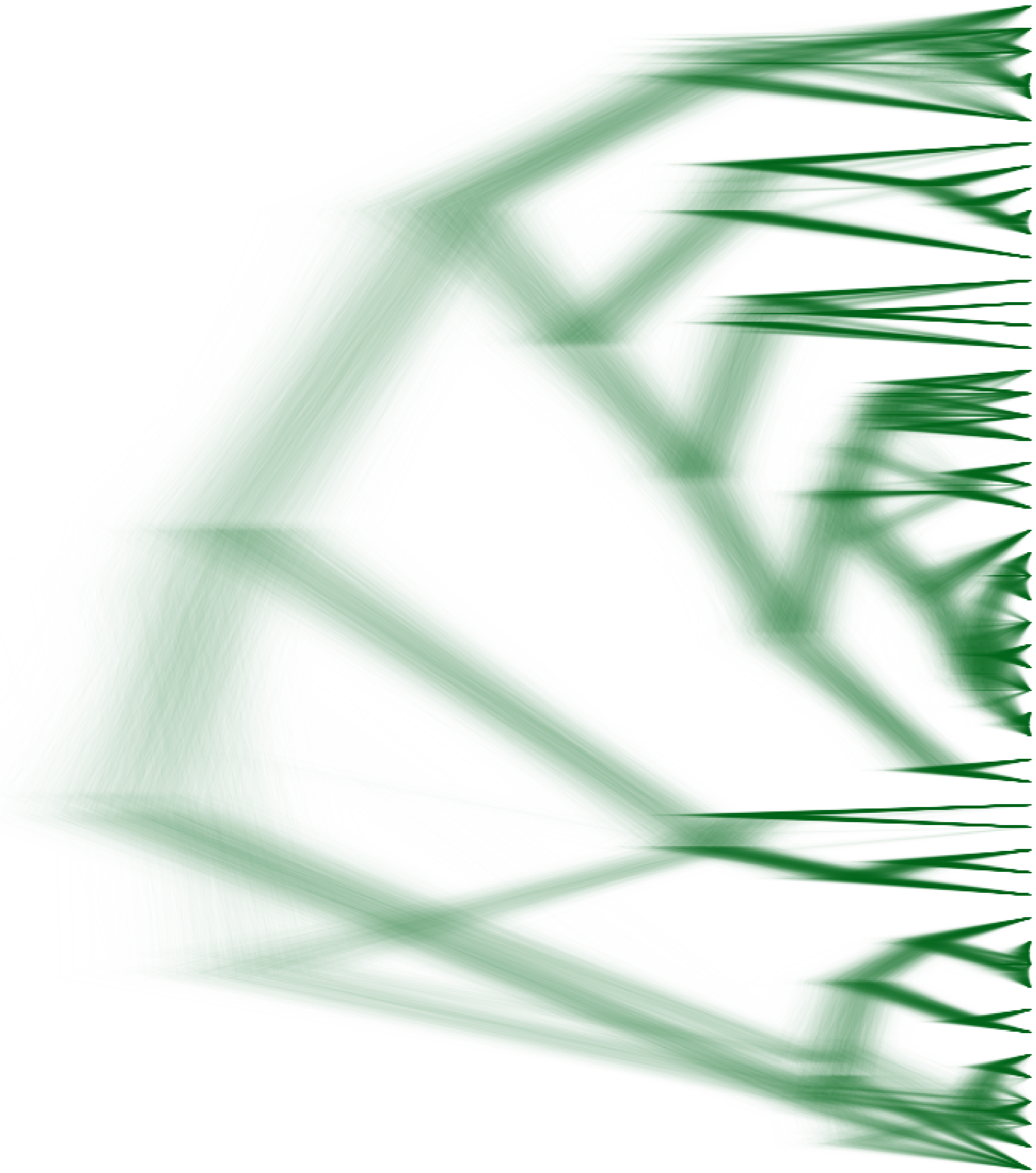
- Read the parameter from a text file
- Easy!



Sample	posterior	likelihood	prior	treeLikelihood	TreeHeight	BirthDeath	birthRate2	relativeDeathRate2
0	-3475.9136606481	-3466.4051675323	-9.5084931158	-3466.4051675323	1.1391659472	-2.6007378369	1	0.5
1000	-1975.0035787646	-1977.6321093158	2.6285305512	-1977.6321093158	0.0648790667	9.5362858302	12.5242404781	0.3290506061
2000	-1974.8162462219	-1978.1713268878	3.3550806659	-1978.1713268878	0.0595933261	10.2628359449	14.0711865818	0.5674041694
3000	-1973.071080435	-1975.985938212	2.914857777	-1975.985938212	0.0637379028	9.822613056	12.4957598412	0.3713791164
4000	-1976.1721083055	-1977.6765552402	1.5044469346	-1977.6765552402	0.0589402834	8.4122022136	22.959140933	0.7081418783
5000	-1975.1875900244	-1978.3731512867	3.1855612622	-1978.3731512867	0.0537299068	10.0933165412	18.1745190755	0.2731537656
6000	-1976.3331096928	-1977.7743572871	1.4412475943	-1977.7743572871	0.062612451	8.3490028733	36.013616164	0.1287791966
7000	-1973.2931193027	-1977.0070547169	3.7139354142	-1977.0070547169	0.0560960422	10.6216906932	24.7016851469	0.1525740068
8000	-1973.9495560865	-1977.5366901535	3.587134067	-1977.5366901535	0.0530107688	10.494889346	23.4446257523	0.1443719595

2.4 for model B: estimate parameters used by model A

- Model A has different parameters than model B
- Parameters for model A can be estimated from the phylogenies in the posterior of model B



2.4. Estimate parameters from phylogenies

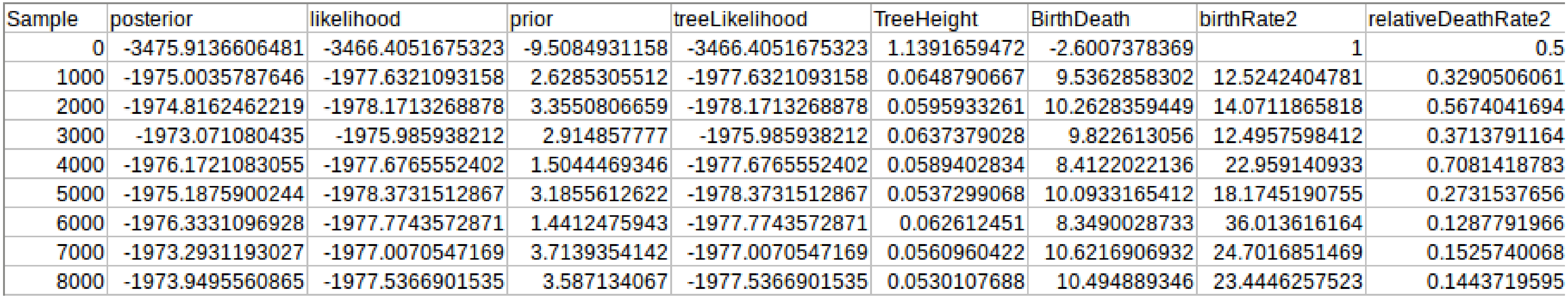
- Effective population size (Kuhner et al. 1995)
- Rate of population growth or decline (Kuhner et al. 1998, Drummond et al. 2002)
- Migration rates and population structure (Beerli & Felsenstein 1999, Beerli & Felsenstein 2001, Ewing et al. 2004, Ewing & Rodrigo 2006)
- Recombination rates and reticulate ancestry (Kuhner et al. 2000, Bloomquist & Suchard 2010)

3. Do the same vice versa

- Easy

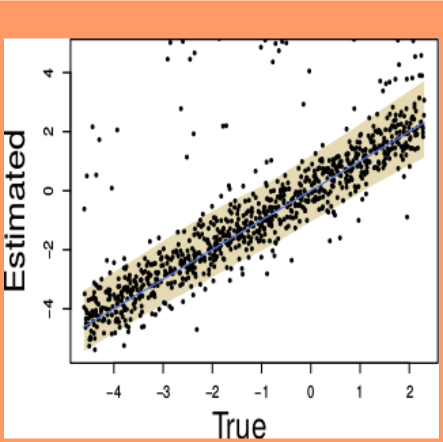
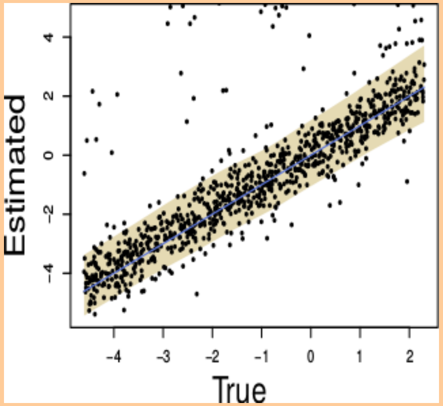
2.3. For model A: obtain parameter distribution from posterior

- Read the parameter from a text file
- Easy!



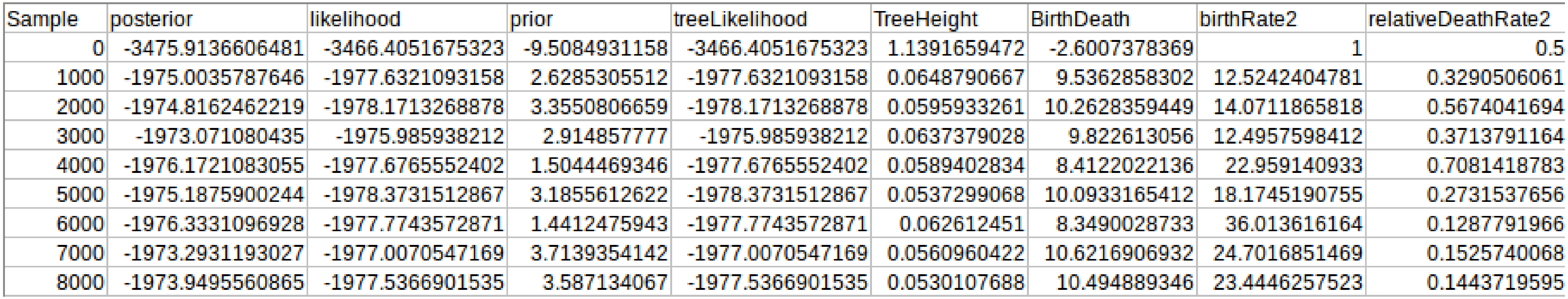
Sample	posterior	likelihood	prior	treeLikelihood	TreeHeight	BirthDeath	birthRate2	relativeDeathRate2
0	-3475.9136606481	-3466.4051675323	-9.5084931158	-3466.4051675323	1.1391659472	-2.6007378369	1	0.5
1000	-1975.0035787646	-1977.6321093158	2.6285305512	-1977.6321093158	0.0648790667	9.5362858302	12.5242404781	0.3290506061
2000	-1974.8162462219	-1978.1713268878	3.3550806659	-1978.1713268878	0.0595933261	10.2628359449	14.0711865818	0.5674041694
3000	-1973.071080435	-1975.985938212	2.914857777	-1975.985938212	0.0637379028	9.822613056	12.4957598412	0.3713791164
4000	-1976.1721083055	-1977.6765552402	1.5044469346	-1977.6765552402	0.0589402834	8.4122022136	22.959140933	0.7081418783
5000	-1975.1875900244	-1978.3731512867	3.1855612622	-1978.3731512867	0.0537299068	10.0933165412	18.1745190755	0.2731537656
6000	-1976.3331096928	-1977.7743572871	1.4412475943	-1977.7743572871	0.062612451	8.3490028733	36.013616164	0.1287791966
7000	-1973.2931193027	-1977.0070547169	3.7139354142	-1977.0070547169	0.0560960422	10.6216906932	24.7016851469	0.1525740068
8000	-1973.9495560865	-1977.5366901535	3.587134067	-1977.5366901535	0.0530107688	10.494889346	23.4446257523	0.1443719595

4. Statistical analysis

	Data created by model A	Data created by model B
Parameters estimated best by model A		Next slide
Parameters estimated best by model B	Next slide	

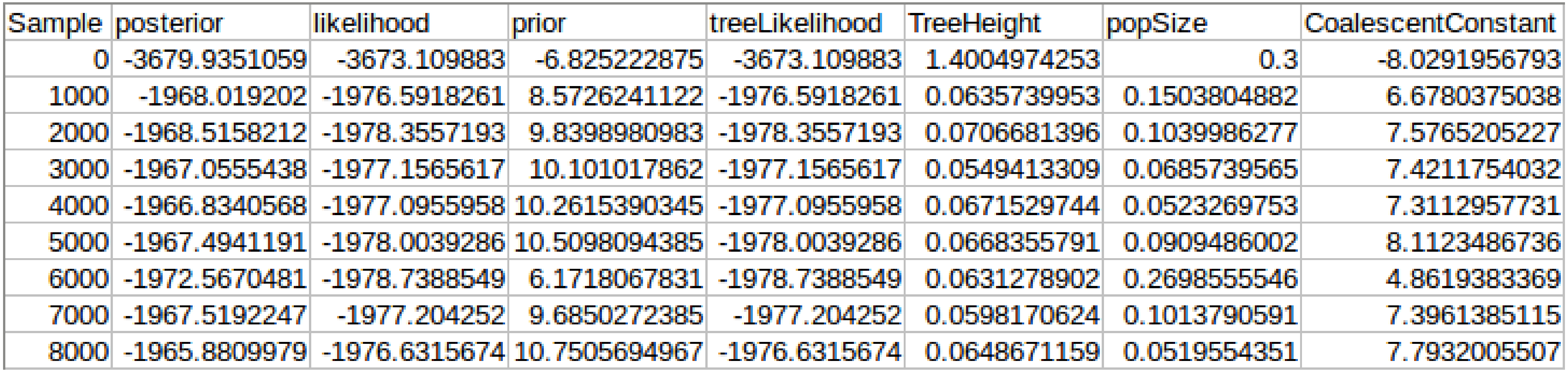
4. Statistical analysis

Model A on data simulated by model A (example values)



Sample	posterior	likelihood	prior	treeLikelihood	TreeHeight	BirthDeath	birthRate2	relativeDeathRate2
0	-3475.9136606481	-3466.4051675323	-9.5084931158	-3466.4051675323	1.1391659472	-2.6007378369	1	0.5
1000	-1975.0035787646	-1977.6321093158	2.6285305512	-1977.6321093158	0.0648790667	9.5362858302	12.5242404781	0.3290506061
2000	-1974.8162462219	-1978.1713268878	3.3550806659	-1978.1713268878	0.0595933261	10.2628359449	14.0711865818	0.5674041694
3000	-1973.071080435	-1975.985938212	2.914857777	-1975.985938212	0.0637379028	9.822613056	12.4957598412	0.3713791164
4000	-1976.1721083055	-1977.6765552402	1.5044469346	-1977.6765552402	0.0589402834	8.4122022136	22.959140933	0.7081418783
5000	-1975.1875900244	-1978.3731512867	3.1855612622	-1978.3731512867	0.0537299068	10.0933165412	18.1745190755	0.2731537656
6000	-1976.3331096928	-1977.7743572871	1.4412475943	-1977.7743572871	0.062612451	8.3490028733	36.013616164	0.1287791966
7000	-1973.2931193027	-1977.0070547169	3.7139354142	-1977.0070547169	0.0560960422	10.6216906932	24.7016851469	0.1525740068
8000	-1973.9495560865	-1977.5366901535	3.587134067	-1977.5366901535	0.0530107688	10.494889346	23.4446257523	0.1443719595

Model B on data simulated by model A (example values)



Sample	posterior	likelihood	prior	treeLikelihood	TreeHeight	popSize	CoalescentConstant
0	-3679.9351059	-3673.109883	-6.825222875	-3673.109883	1.4004974253	0.3	-8.0291956793
1000	-1968.019202	-1976.5918261	8.5726241122	-1976.5918261	0.0635739953	0.1503804882	6.6780375038
2000	-1968.5158212	-1978.3557193	9.8398980983	-1978.3557193	0.0706681396	0.1039986277	7.5765205227
3000	-1967.0555438	-1977.1565617	10.101017862	-1977.1565617	0.0549413309	0.0685739565	7.4211754032
4000	-1966.8340568	-1977.0955958	10.2615390345	-1977.0955958	0.0671529744	0.0523269753	7.3112957731
5000	-1967.4941191	-1978.0039286	10.5098094385	-1978.0039286	0.0668355791	0.0909486002	8.1123486736
6000	-1972.5670481	-1978.7388549	6.1718067831	-1978.7388549	0.0631278902	0.2698555546	4.8619383369
7000	-1967.5192247	-1977.204252	9.6850272385	-1977.204252	0.0598170624	0.1013790591	7.3961385115
8000	-1965.8809979	-1976.6315674	10.7505694967	-1976.6315674	0.0648671159	0.0519554351	7.7932005507

So, which model is best?

4. Statistical analysis

- Null expectation

	Data created by model A	Data created by model B
Parameters estimated best by model A	100	0
Parameters estimated best by model B	0	100

4. Statistical analysis

- B is superior

	Data created by model A	Data created by model B
Parameters estimated best by model A	0	0
Parameters estimated best by model B	100	100

4. Statistical analysis

- But else?

	Data created by model A	Data created by model B
Parameters estimated best by model A	60	30
Parameters estimated best by model B	40	70

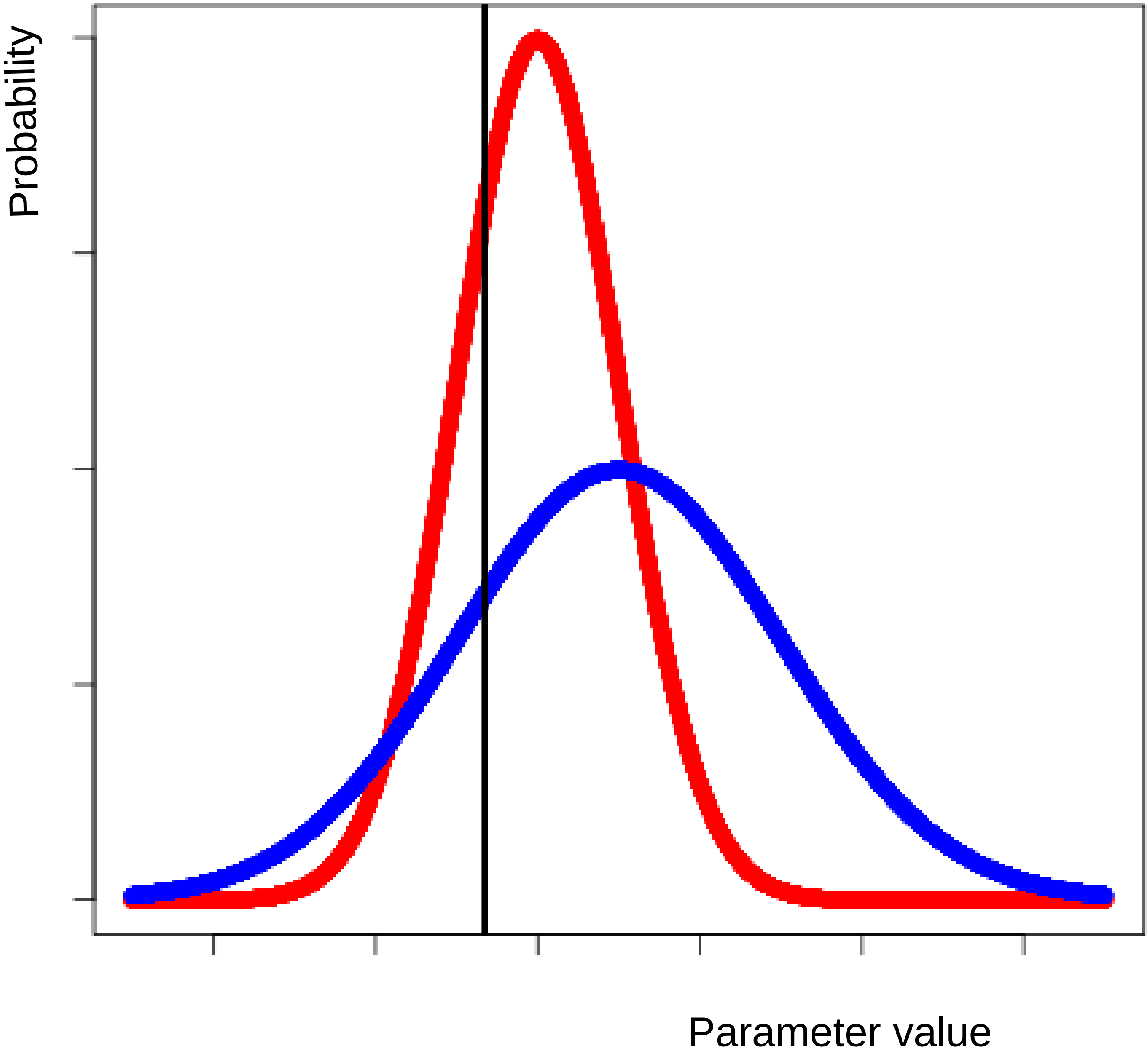
4. Statistical analysis

- ?

	H_0: A = true	H_0: A = false
Accept A	OK	Type I error
Reject A	Type II error	OK

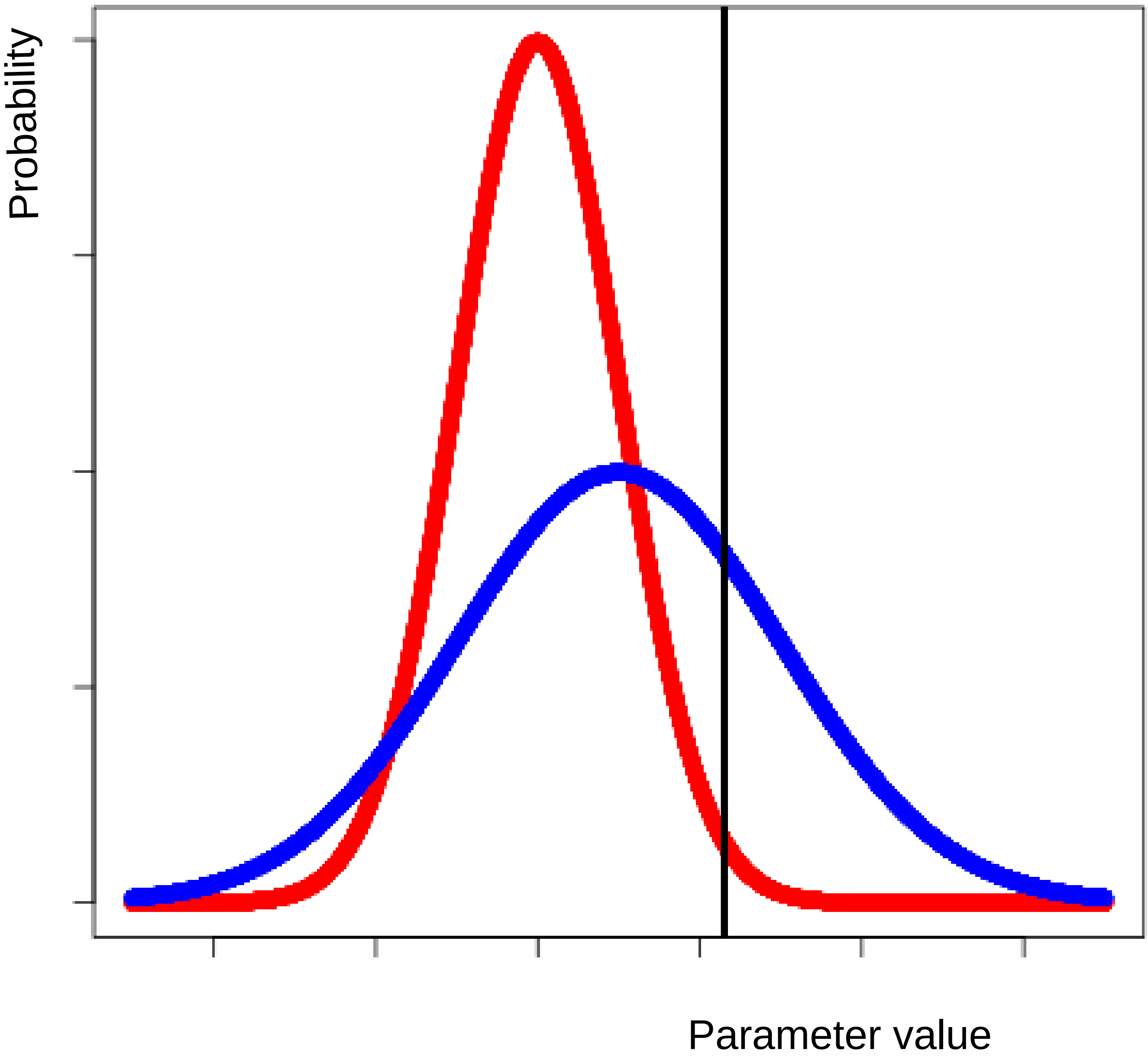
4. Statistical analysis

- If the known parameter value (black bar) is estimated best by its own model (red): OK



4. Statistical analysis

- If the known parameter value (black bar) is estimated best by other model (blue): ?


Probability
Parameter value