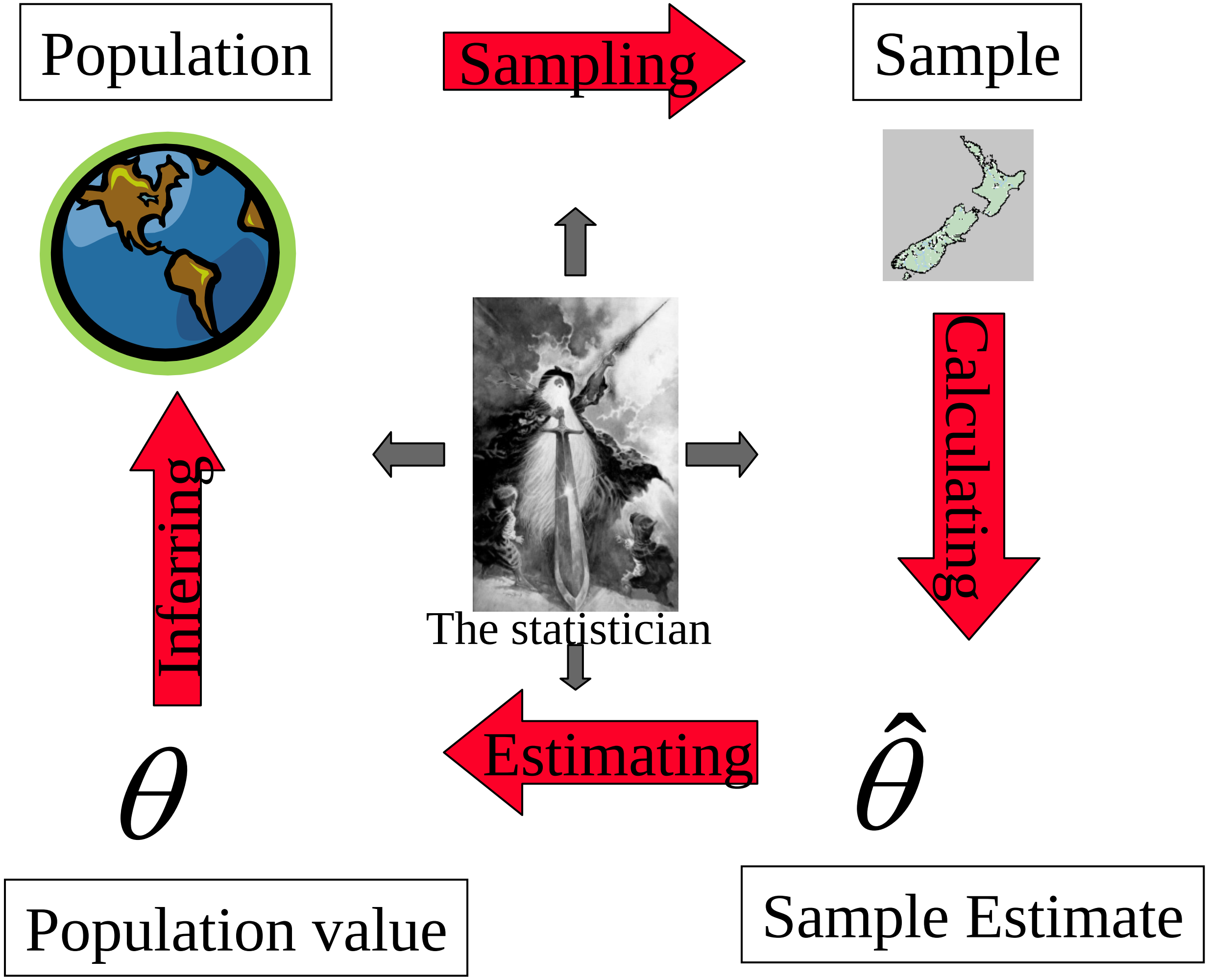
Sampling uncertainty

DNA Statistics Workshop

ISFG

2007

The statistical process



Population

Sampling

Sample

Calculating

Inferring

The statistician

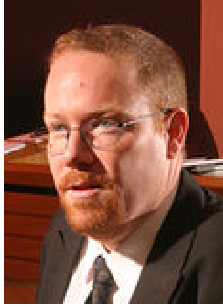
Estimating

θ

$\hat{\theta}$

Population value

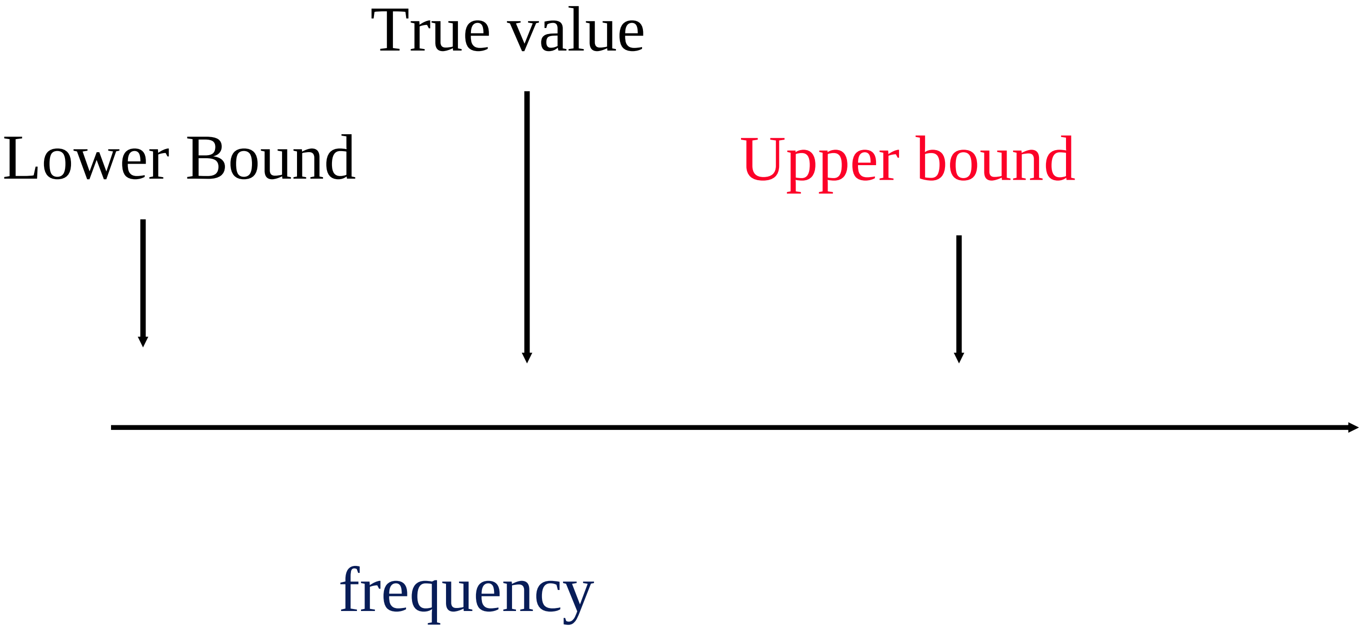
Sample Estimate



Dr James Curran

Modern sampling error corrections

- One sided vs. Two sided
- 99% 95% ?
- How?



True value

Lower Bound

Upper bound

frequency

The product of confidence intervals

is not

the confidence interval of a product

Options

- Factor of 10 (NRC II)
- NRC II (Chakraborty)
- Bootstrap (Efron)
- Size bias correction (Balding)
- Bayesian posterior (Weir et al)

Factor of 10 rule

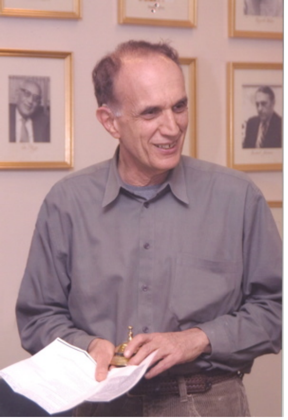
- Very easy to apply
- Is not a general rule
- Cannot handle zeroes without an additional (arbitrary) rule.
- Has no statistical credibility

NRC II (Chakraborty)

Chakraborty, Srinvasan and Daiger "Evaluation of standard error and confidence interval of estimated multilocus genotype probabilities and their implications in DNA Forensics" Am J Hum Genet 1993;52:60-70

NRC II (Chakraborty)

- Easily done with a calculator or spreadsheet, but time consuming
- Assumes normality of the log of the frequency
- Cannot handle zeroes without an additional (arbitrary) rule. (no issue if $\theta > 0$)
- Not yet extended to paternity, mixtures or missing persons and may have to be done on a case by case basis (BSW is working on it)
- Strong support in the statistical community
- EXCEL sheet available



The Bootstrap (Efron)

- Requires a purpose written program (one is available)
- Program must be run for each case
- Cannot handle zeroes without an additional (arbitrary) rule. (no issue if $\theta > 0$)
- Easily extended to paternity, mixtures or missing persons and may have to be done on a case by case basis
- Few, if any, modeling assumptions.
- Strong support in the statistical community

The Size Bias Correction

“Estimating products in forensic identification”
Balding J Am Stat Assoc 1995;90(431):839-844

Use x+2/N+4 for hets frequency

Use x+4/N+4 for homs frequency

Try frequency of AB type if there are 98 A alleles and 98 B alleles out of 396 alleles sampled?

The Size Bias Correction

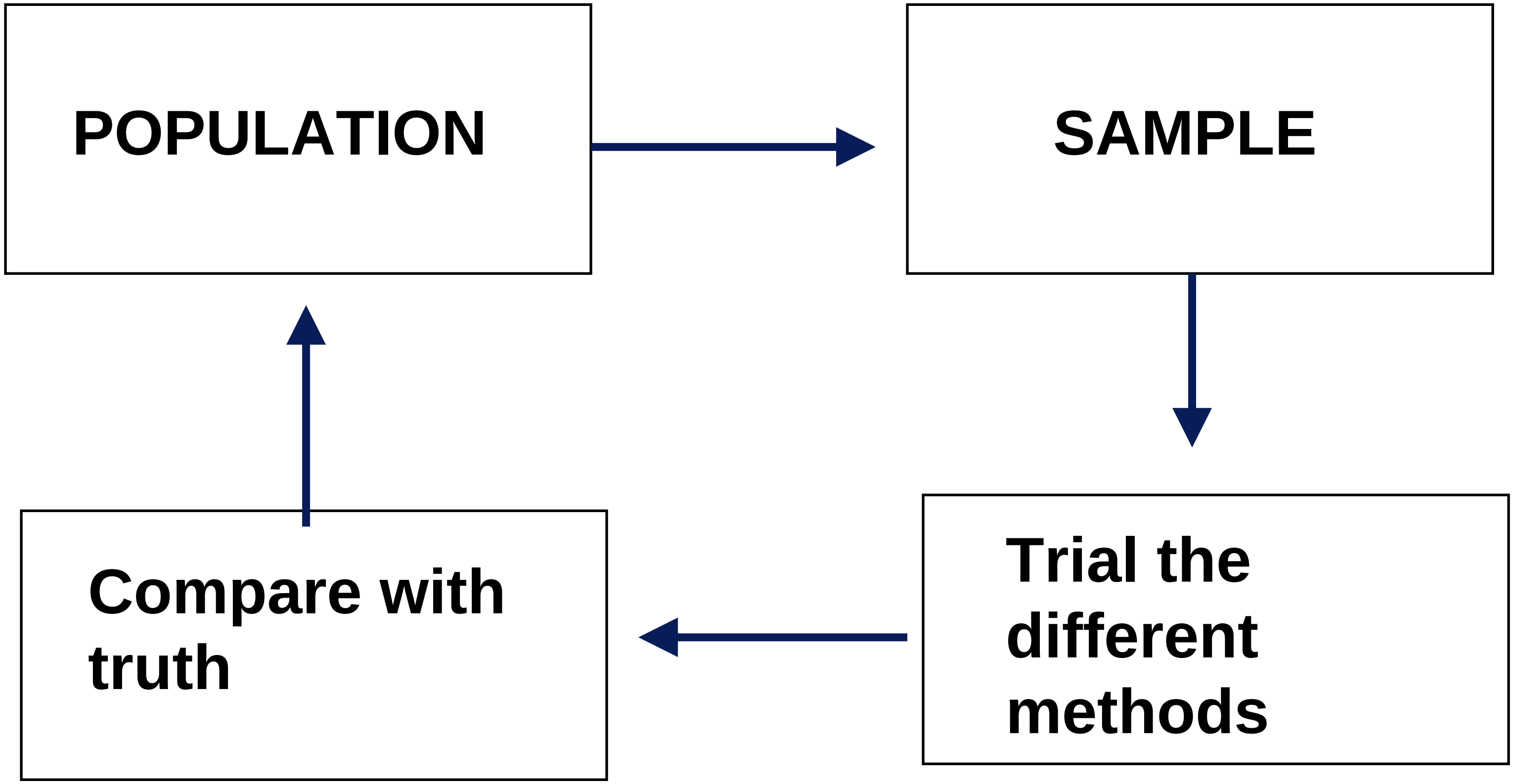
- Can handle zeroes
- Can be done with a calculator or spreadsheet
- Not easily extended to paternity, mixtures or missing persons and may have to be done on a case by case basis
- Derivation appears flawed
- A partial correction appears on pg 138 and 139 of Evett and Weir
- A more substantial correction in Curran et al.
- No support in the statistical community

Bayesian posterior

- Can handle zeroes
- Requires a purpose written program, one is available
- Easily extended to paternity, mixtures or missing persons and may have to be done on a case by case basis
- Excellent sampling properties

Bayesian Posterior vs. CI

- BP: It is 99% probable that the true value lies between x and y
- CI: With 99% confidence the interval x to y contains the true value
- The difference is a bit subtle but the top one seems a bit easier

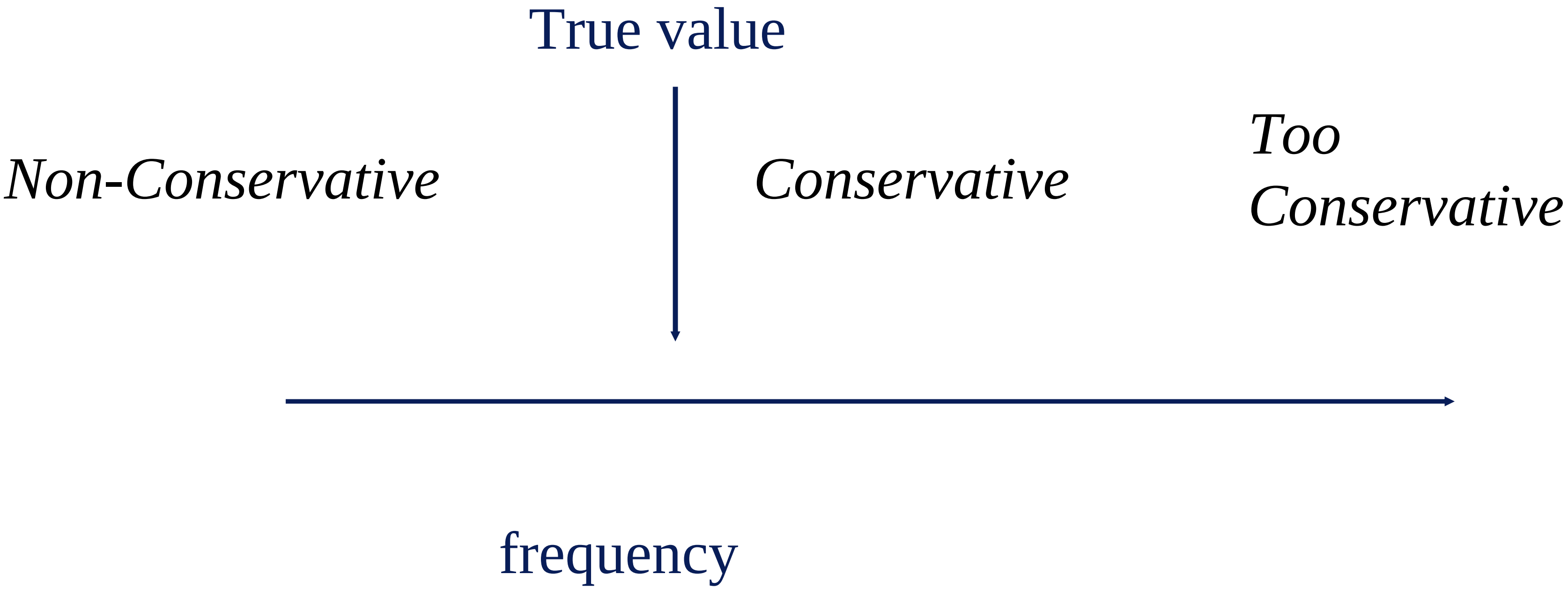


POPULATION

SAMPLE

Trial the different methods

Compare with truth



True value

Non-Conservative

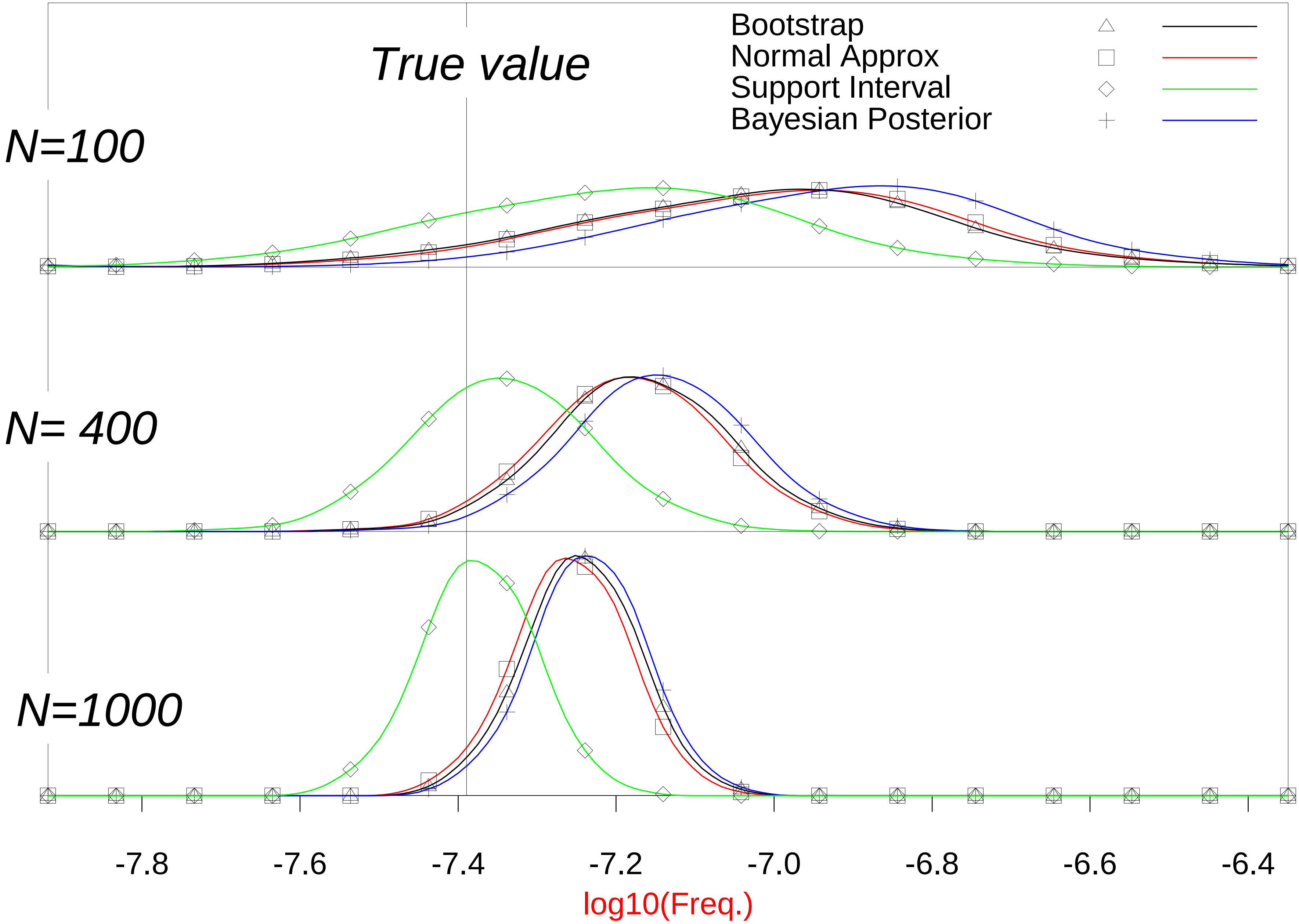
Conservative

Too Conservative

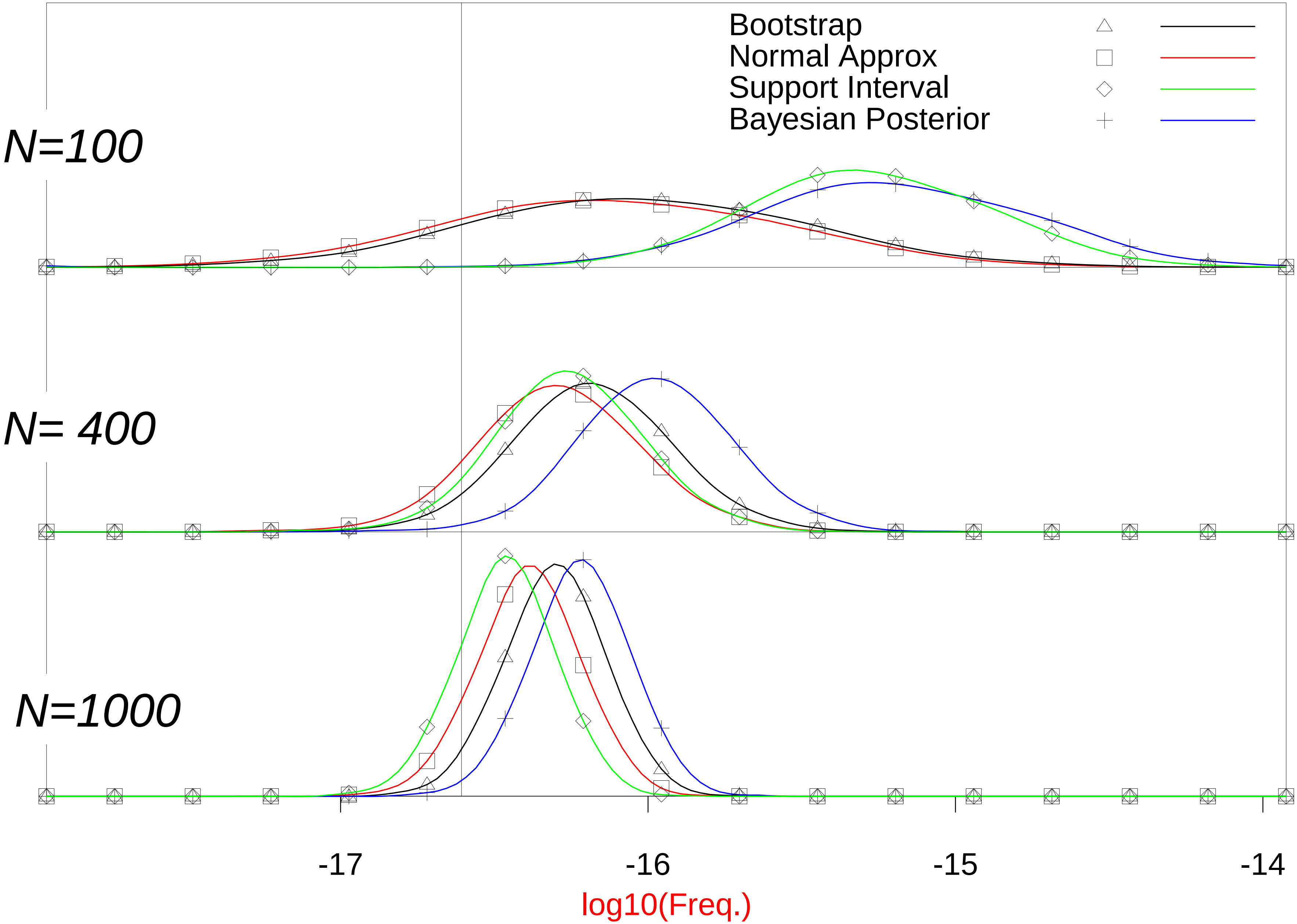
frequency

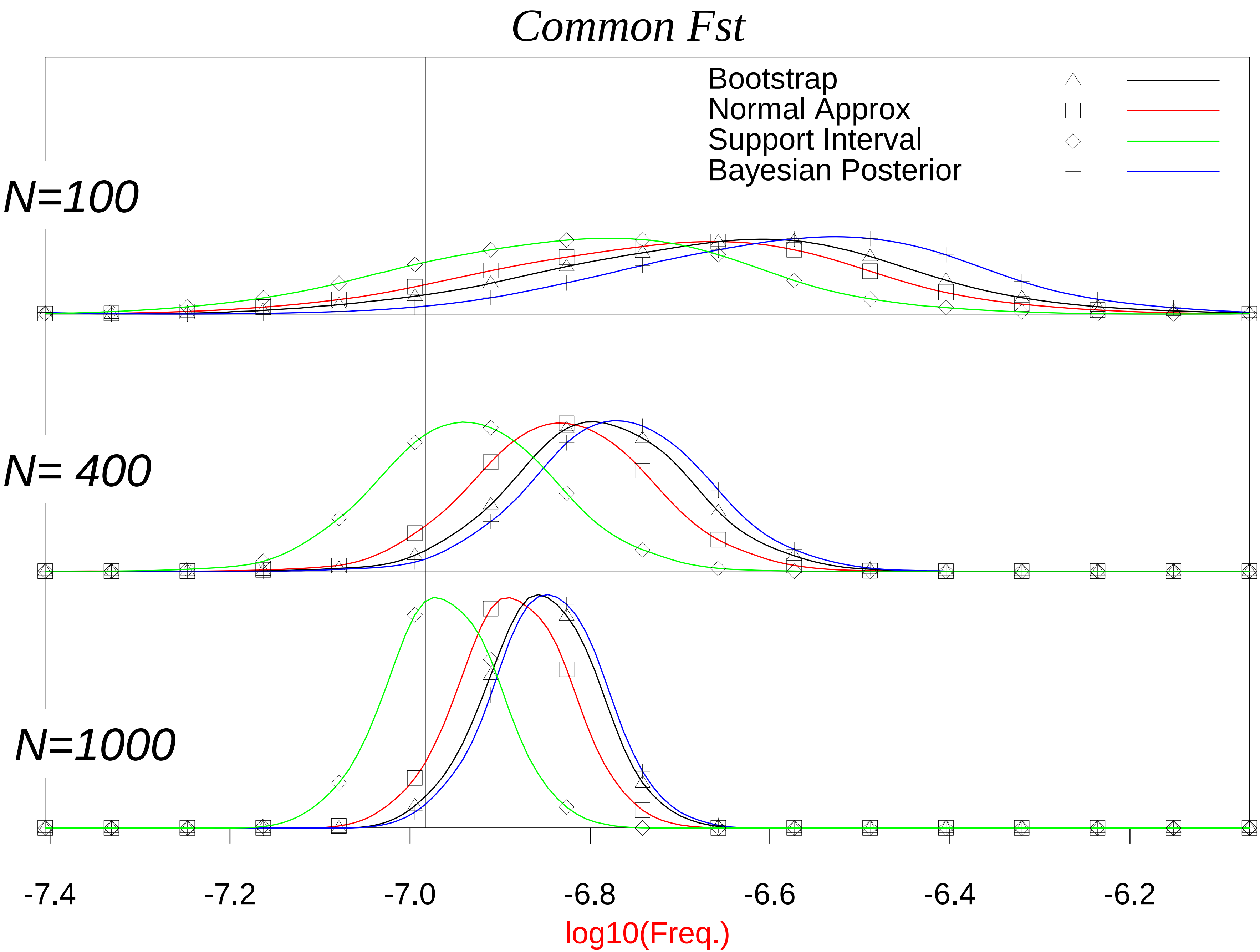
Non-Conservative

Common Product Rule

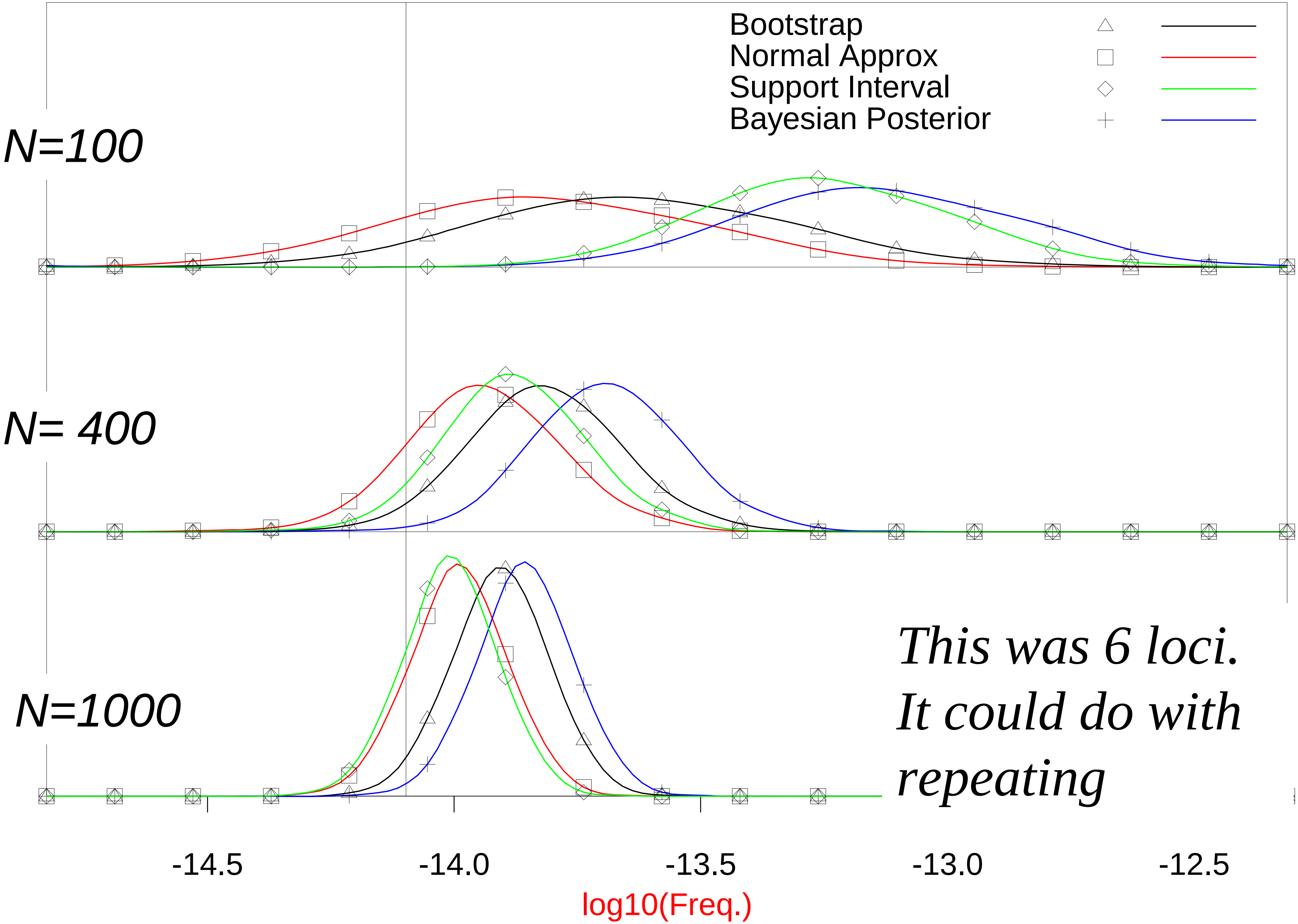


Rare Product Rule




Common Fst
Bootstrap △
Normal Approx □
Support Interval ◇
Bayesian Posterior +
N=100
N= 400
N=1000
-7.4
-7.2
-7.0
-6.8
-6.6
-6.4
-6.2
log10(Freq.)

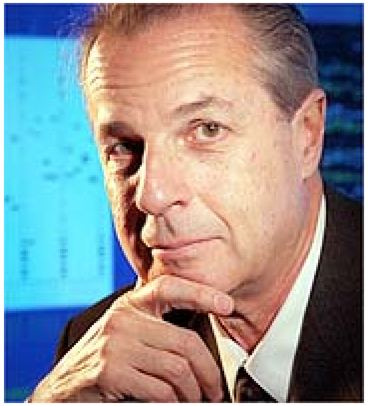
Rare Fst



*This was 6 loci.
It could do with
repeating*

This was mainly about “how” and “how effective”

Do we need to consider whether it is needed at all?



DNA Frequency Uncertainty – Why Bother?

A challenge

Will someone tell me, please, what rational difference it ever can make to know the confidence limits in addition to knowing the best point estimate? Specifically, can you give premises under which, for a fixed point estimate, the decision to convict or not to convict would depend on the size of the confidence interval?

http://dna-view.com/noconfid.htm

END