Maximum likelihood estimates under k-Allele models with selection can be numerically unstable

by Paul Joyce
University of Idaho

The stationary distributions of allele frequencies under a variety of common population genetics models called Wright-Fisher k-allele models with selection are well studied. However, the statistical properties of maximum likelihood estimates of parameters under these models are not well understood. Under each of these models there is a point in data space which carries the strongest possible signal for selection, yet, at this point, the maximum likelihood estimate for selection intensity does not exist. This result remains valid even if all other parameters in the model are assumed to be known. We will show that this singularity in data space can cause the parametric bootstrap to produces inaccurate and unreliable error estimates of the selection intensity . We describe two alternative methods to build interval estimates for the selection intensity.

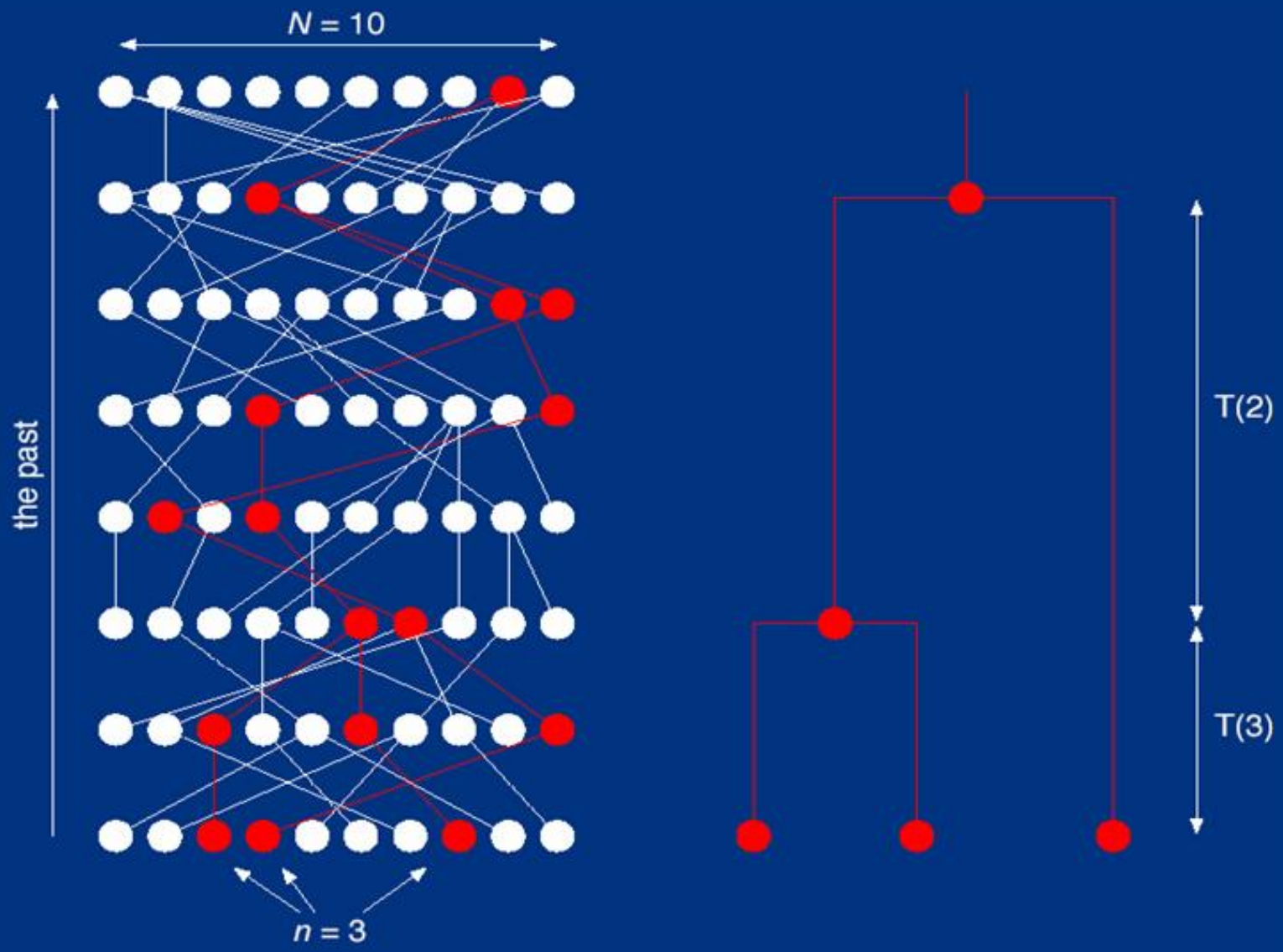
Modern World of Stochastic Models and Statistical Analysis in Genetics

1. Mathematical Models Sophisticated mathematical descriptions are proposed. While they are simplifications of the true biological process, they are often robust descriptions.

2. Statistical Analysis A small part of the process is observed and statisticians are faced with overcoming the missing data problem, or the likelihood is known up to an integration constant which is difficult to calculate.

3. Evaluating the statistical methods Data are simulated under a set of parameters, the statistical procedure is applied and the relationship between true parameters and estimated parameters is investigated.

Wright-Fisher Model



Heterozygote Advantage

Population size—N

Fitness $w(A_i, A_j) = \begin{cases} 1-s & i=j \\ 1 & i \neq j \end{cases}$

Allele Frequencies $x_1, x_2, \cdots, x_k$ where $\sum_j x_j = 1$

Mean Fitness

$$\bar{w} = \sum_{i,j} w(A_i, A_j) x_i x_j = 1 - s \sum_i x_i^2$$

Mutation Rate u

Scaling $\sigma = 2Ns$ and $\theta = 4Nu$.

Frequency Dependent Selection

An allele with frequency x_i has fitness $1 - sx_i$ then the mean fitness for the population will be

$$\sum_{i=1}^{k}(1 - sx_i)x_i = 1 - s\sum_{i=1}^{k} x_i^2$$

Stationary Distribution

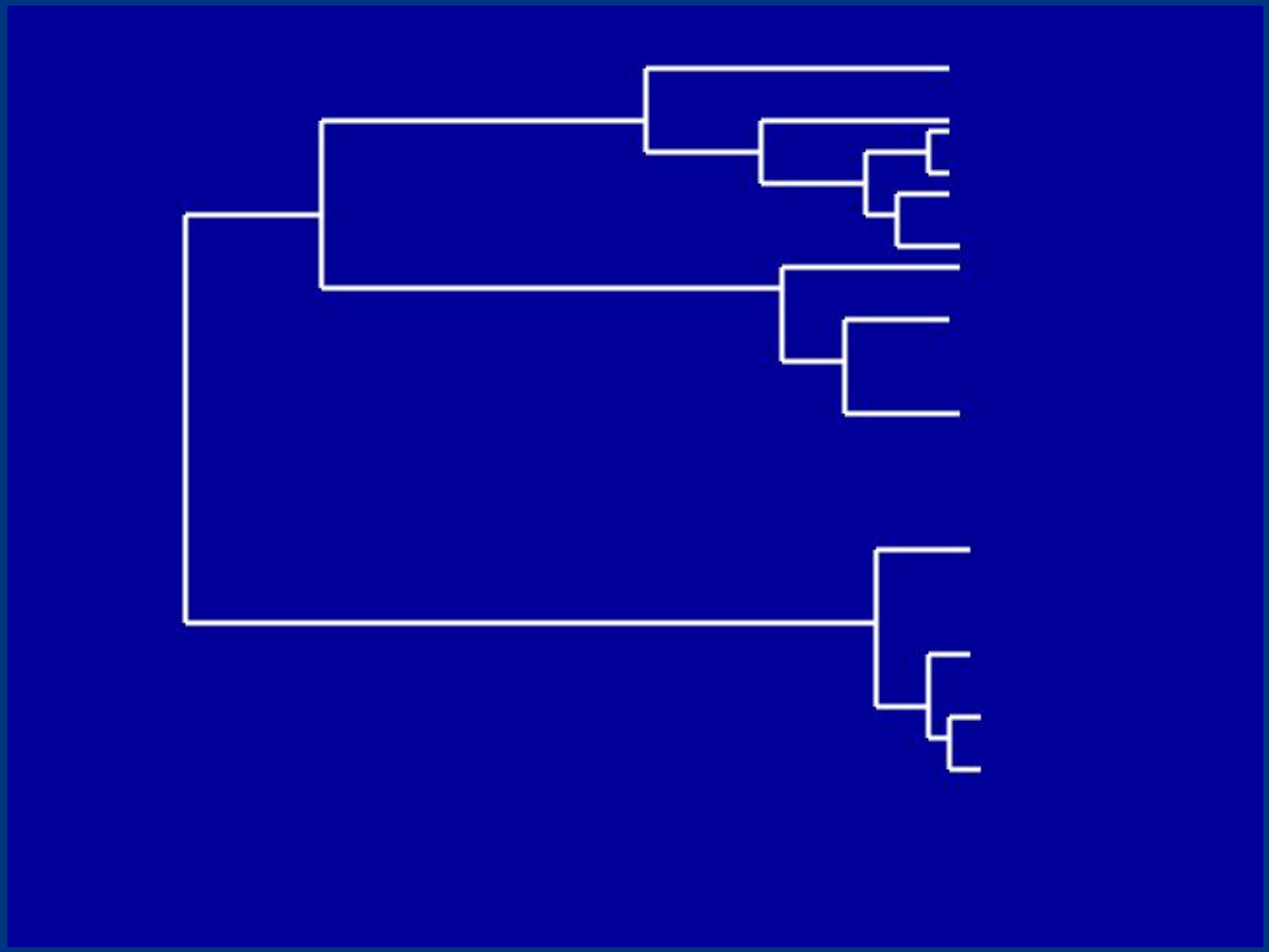
Define $\mathbf{x} = (x_1, x_2, \cdots, x_k)$ to be the allele frequencies, where $\sum x_i = 1$

As $N \to \infty$

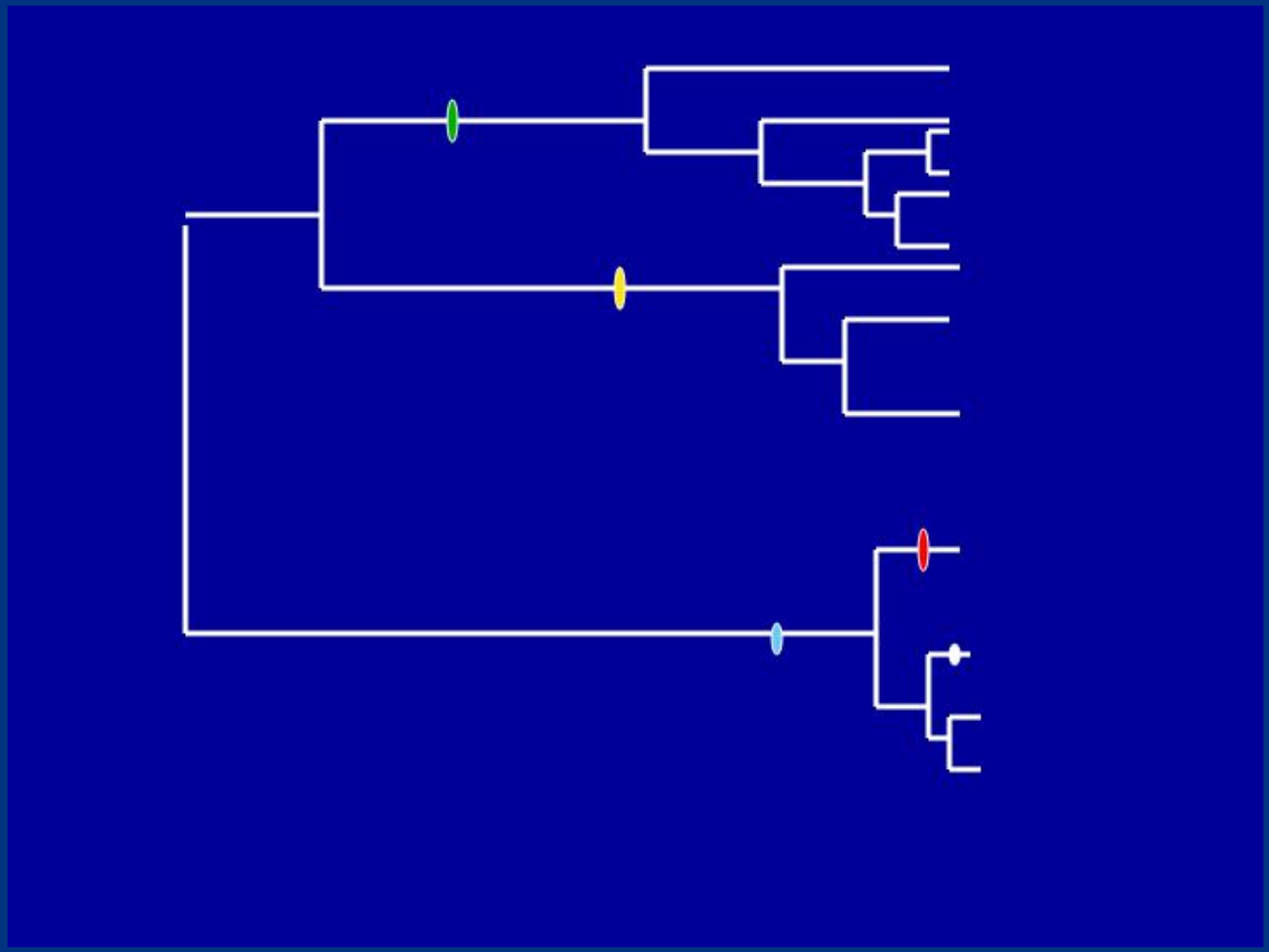
$$f_{\mathrm{Sel}}(\mathbf{x}|\theta, \sigma) = \frac{e^{-\sigma \sum_{i=1}^{k} x_i^2}}{c(\theta, \sigma)} (x_1 x_2 \cdots x_k)^{\theta/k-1}. \qquad (1)$$

When $\sigma = 0$ the population is neutral, no selective advantage.

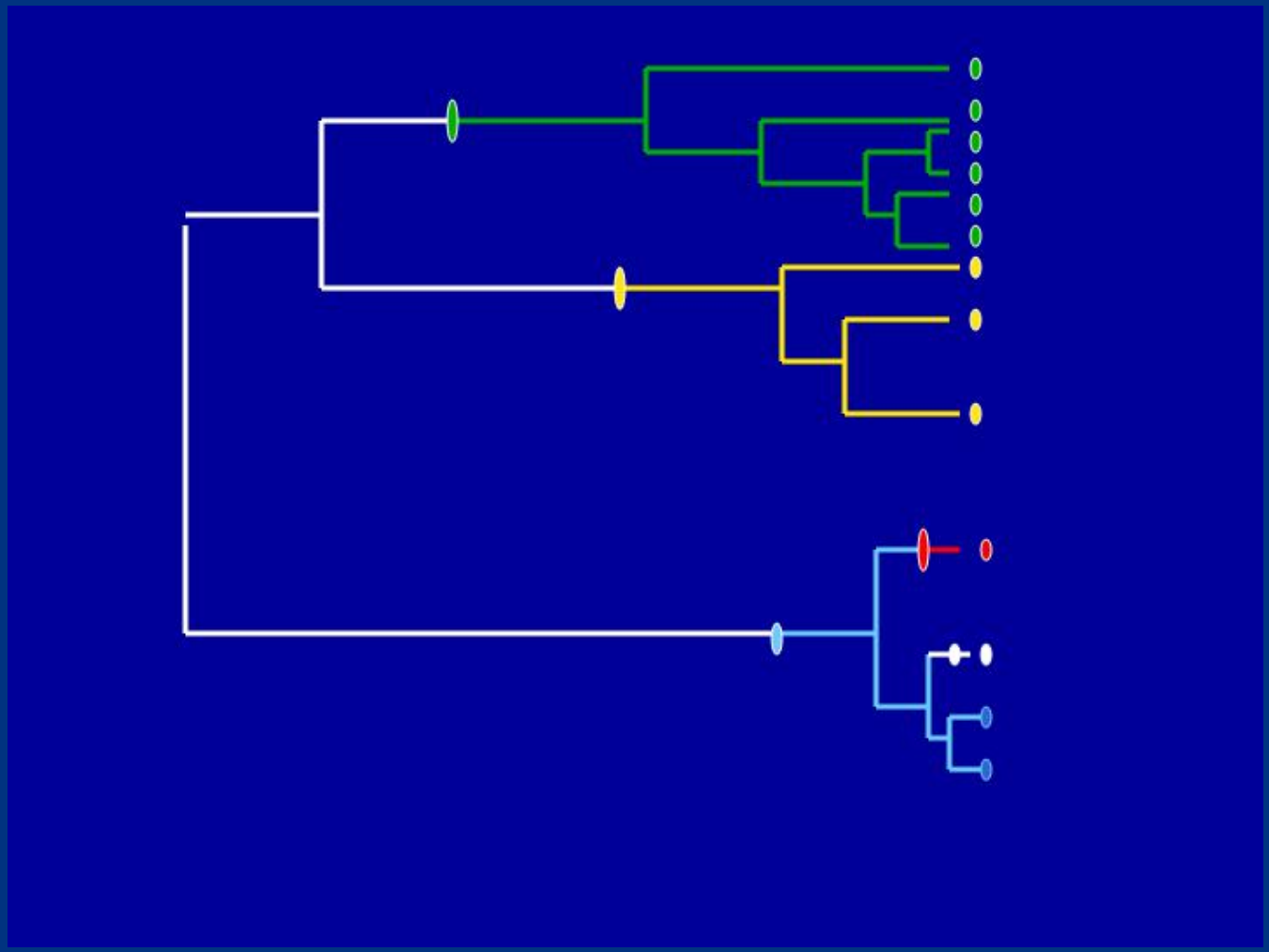
Ancestry of a Alleles



Ancestry of a Alleles



Ancestry of a Alleles



What does a neutral population look like?

type	relative frequencies
1	0.7800
2	0.1730
3	0.0200
4	0.0133
5	0.0067
6	0.0033
7	0.0033
8	0.0003

Older types tend to have larger frequency than younger types. Above is a simulated data set with $k = 8$ and $\theta = 0.3$.

The Effects of Selection

The probability that two individuals chosen at random are the same type is

$$H = \sum_{i=1}^{k} X_i^2.$$

The Selective overdominance model penalizes homozygote, thus decreasing H.

Recall from Calculus

The minimum value of

$$H = \sum_{i=1}^{k} X_i^2$$

subject to the constraint that $\sum_{i=1}^{k} X_i = 1$ occurs when $X_i = \frac{1}{k}$. Selection tends to make the allele frequencies 'more evenly distributed'. It is sometimes referred to as balancing selection.

How would the allele frequencies differ under heterozygote advantage selection?

Simulated Sample under Selection versus Neutrality with (relatively low) Mutation Rate

$\theta = .3$	$\sigma = 200$ relative frequency	$\sigma = 0$ relative frequency
1	0.17	0.7800
2	0.13	0.1730
3	0.13	0.0200
4	0.12	0.0133
5	0.12	0.0068
6	0.12	0.0033
7	0.12	0.0033
8	0.1	0.0003

History of the problem

- Donnelly, Nordborg, Joyce (2001) developed a likelihood framework for analyzing data under various k allele models with selection. They use rejection method for simulating data and importance sampling for calculating the likelihood
- Joyce Genz (2003, 2005) use a numerical method for calculating the constant of integration, and develop methods for sampling from the distribution directly. Thus, simulating the sampling distributions for parameter estimates is now possible.
- Buzbas, Joyce (2009) show that the mle is numerically unstable, do to a the fact that the likelihood as a singularity.

The Sampling Distribution of $\hat{\sigma}$

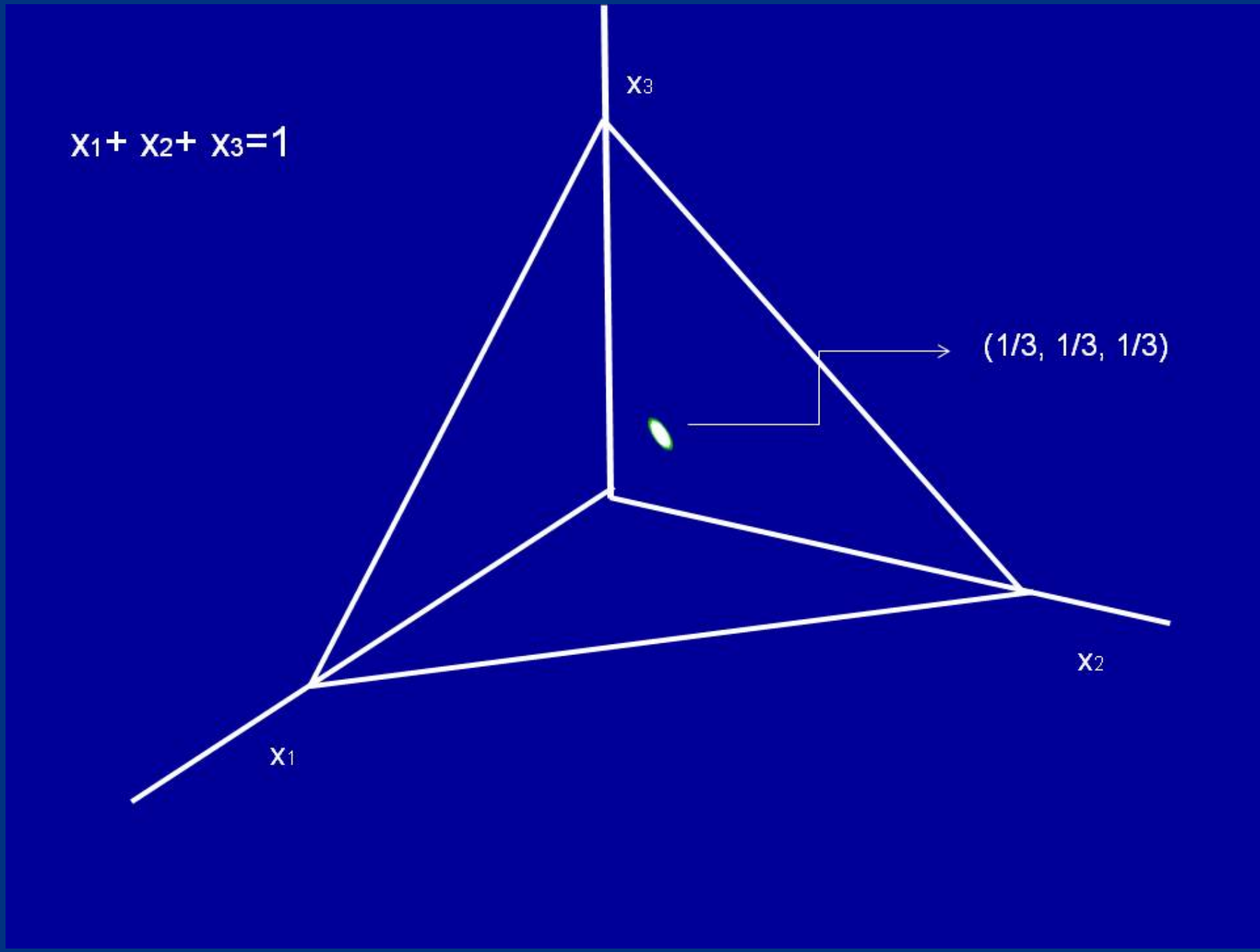
1. For a given value of σ simulate a data set $x_1, x_2 \cdots, x_k$
2. Calculate the maximum likelihood estimate $\hat{\sigma}$
3. Repeat steps (1) and (2) many times and plot the distribution of $\hat{\sigma}$.

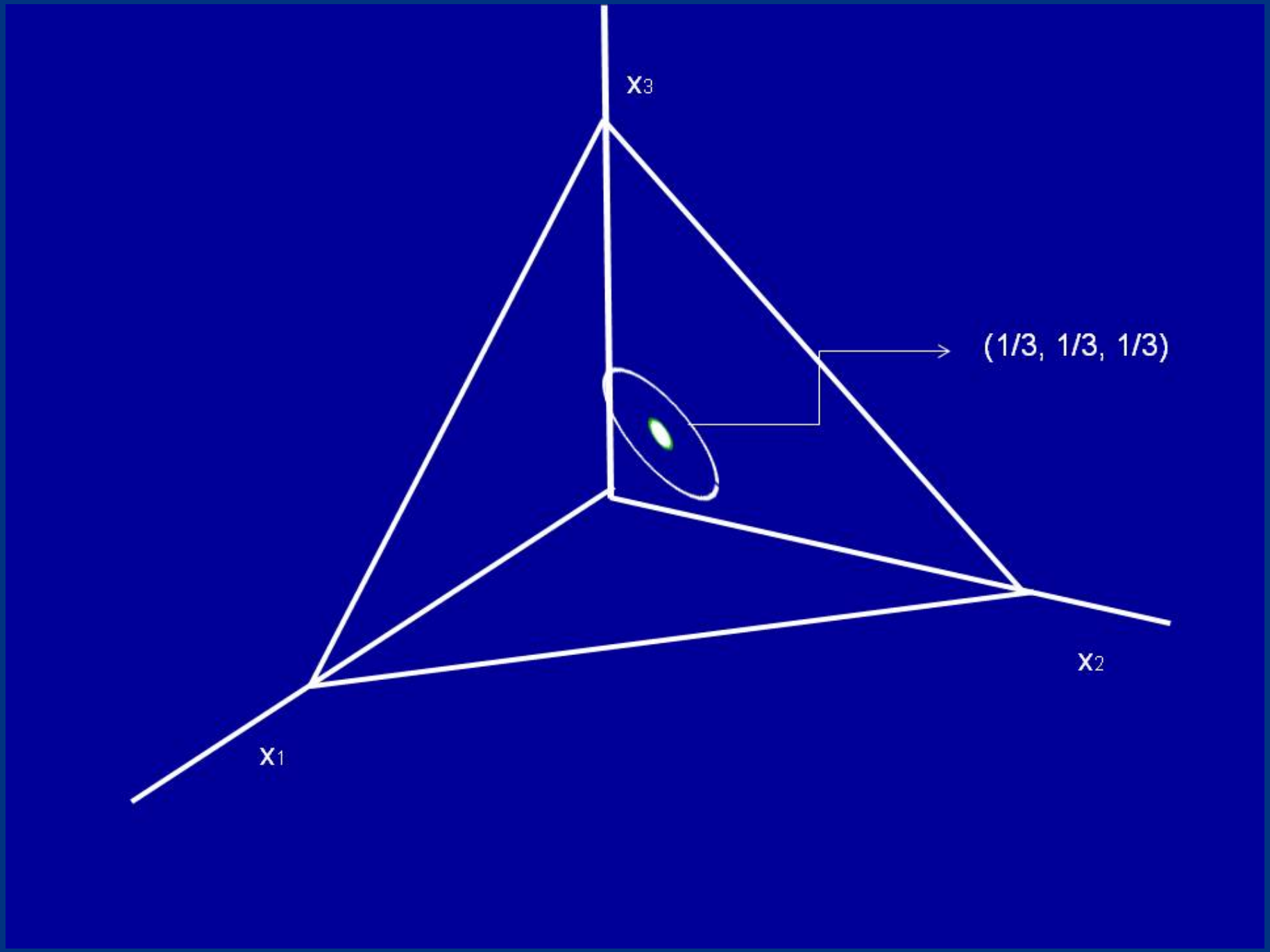
Singularity in Data Space

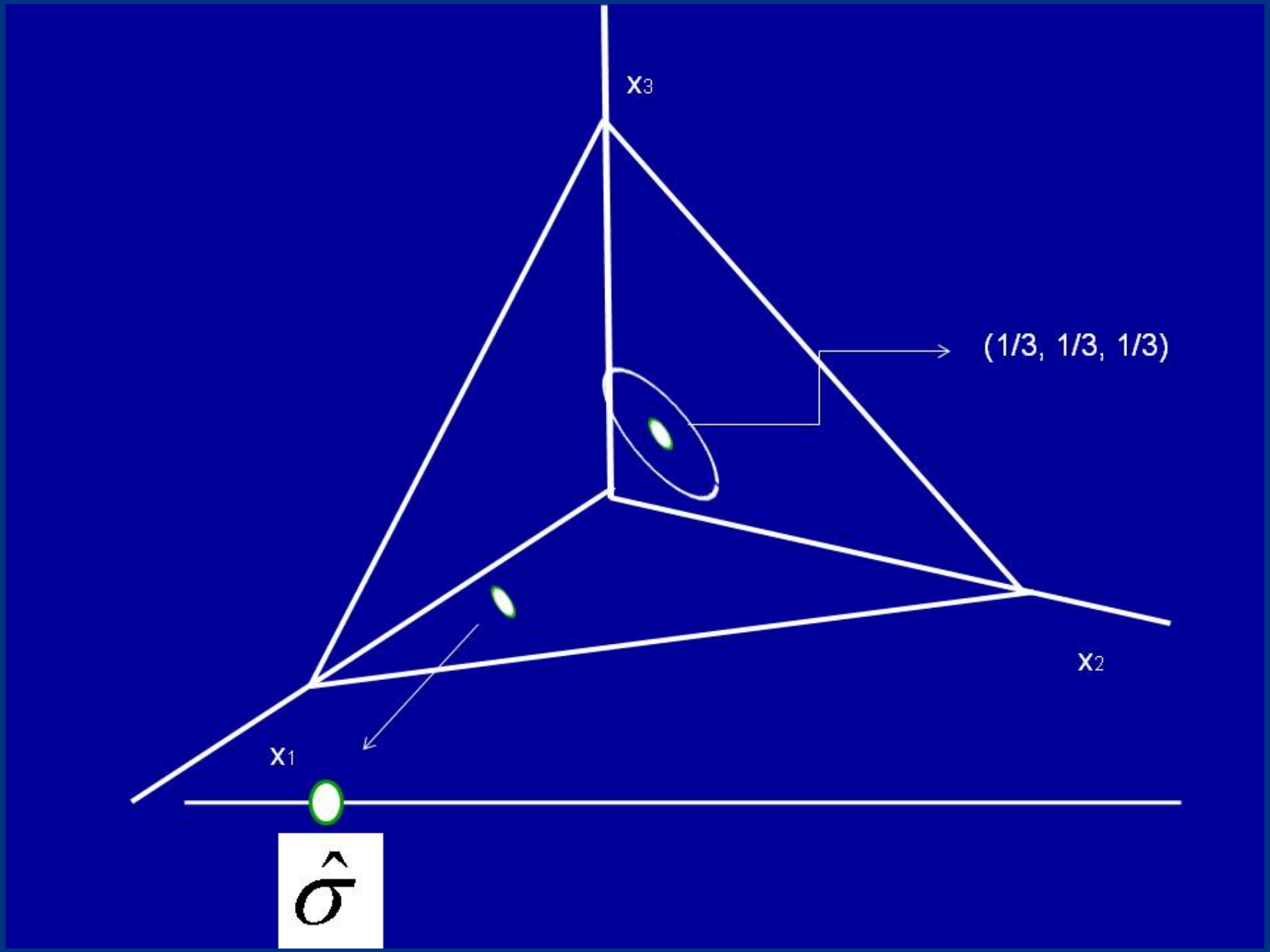
Theorem 1 *Consider the probability density function* $f_{\mathrm{Sel}}(\mathbf{x}|\theta,\sigma)$*, defined by equation (1) that describes the distribution of allele frequencies at stationarity for the Wright-Fisher symmetric selective overdominance model with parent independent mutation. Let* $\mathbf{x}^* = (1/k, \cdots, 1/k)$.

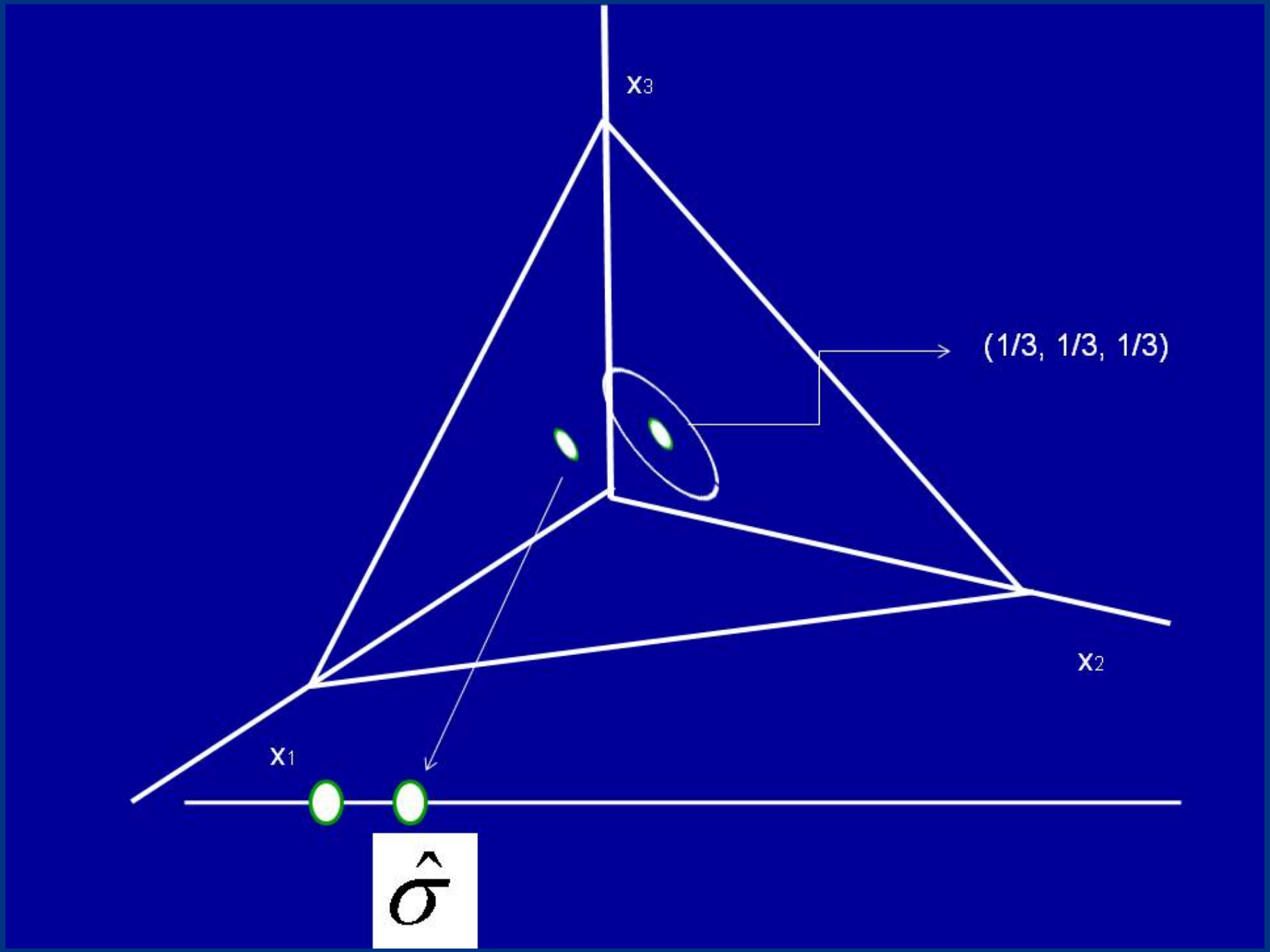
- *If* θ *is assumed to be known, then, for all allele frequencies* $\mathbf{x} \neq \mathbf{x}^*$*, the maximum likelihood estimate for* σ *is finite. Denote the MLE as a function of the homozygosity* $h = \sum_{i=1}^{k} x_i^2$ *by* $\hat{\sigma}(h)$*. Then,*

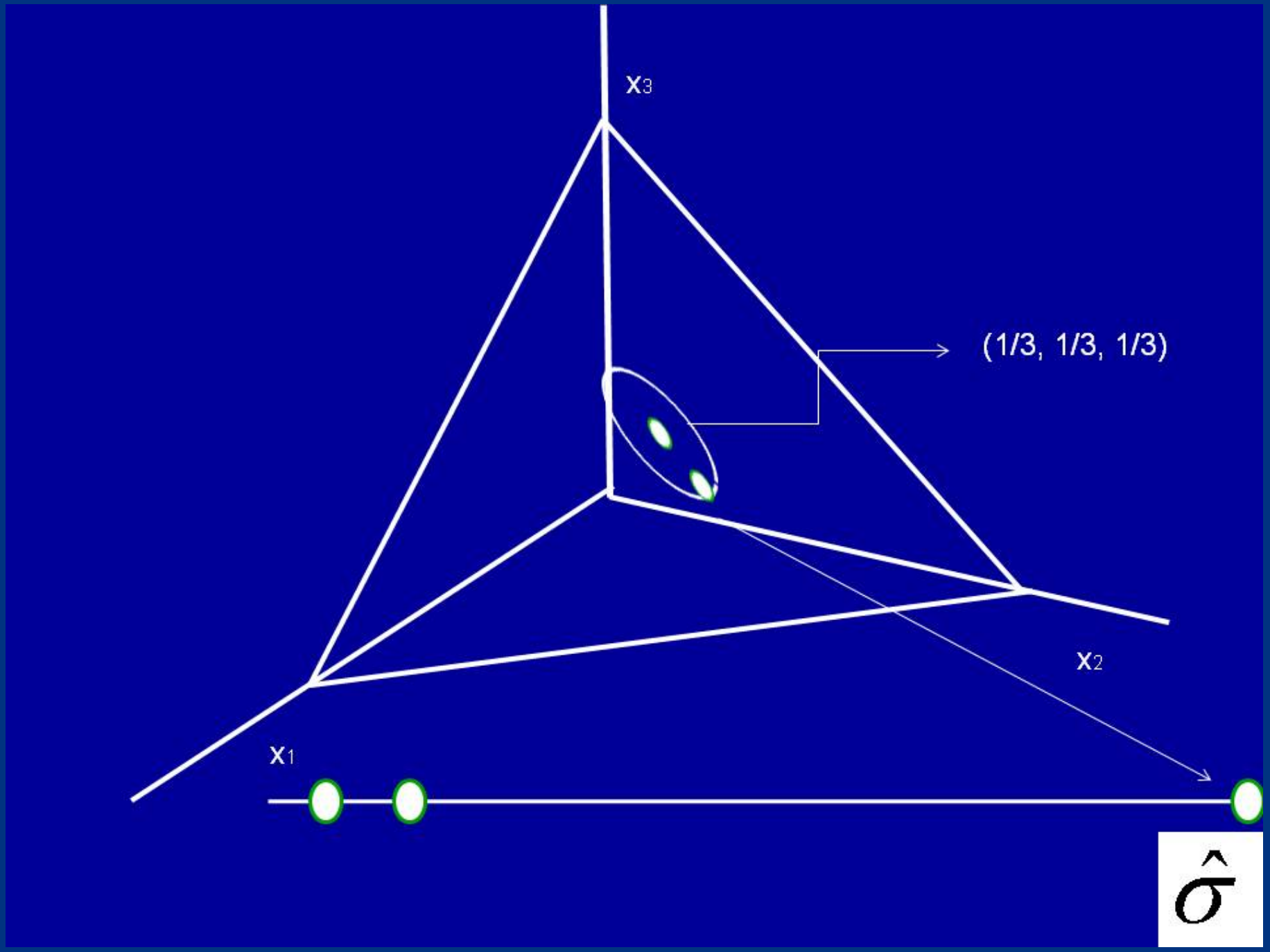
$$\lim_{h \to (1/k)^+} \hat{\sigma}(h) = \infty \tag{2}$$


$x_1 + x_2 + x_3 = 1$
x_3
(1/3, 1/3, 1/3)
x_2
x_1


x3
(1/3, 1/3, 1/3)
x2
x1


x3
(1/3, 1/3, 1/3)
x2
x1
$\hat{\sigma}$


x3
(1/3, 1/3, 1/3)
x2
x1
$\hat{\sigma}$


x3
(1/3, 1/3, 1/3)
x2
x1
σ̂

Consider for example a highly polymorphic locus with $k = 20$

$1/k = 0.05$

A 38% decrease in homozygosity ($h = 0.13$ to $h = 0.08$) corresponds to an approximate 300% increase in $\hat{\sigma}(h)$

$\hat{\sigma}(0.13) \approx 350$

$\hat{\sigma}(0.08) > 900$

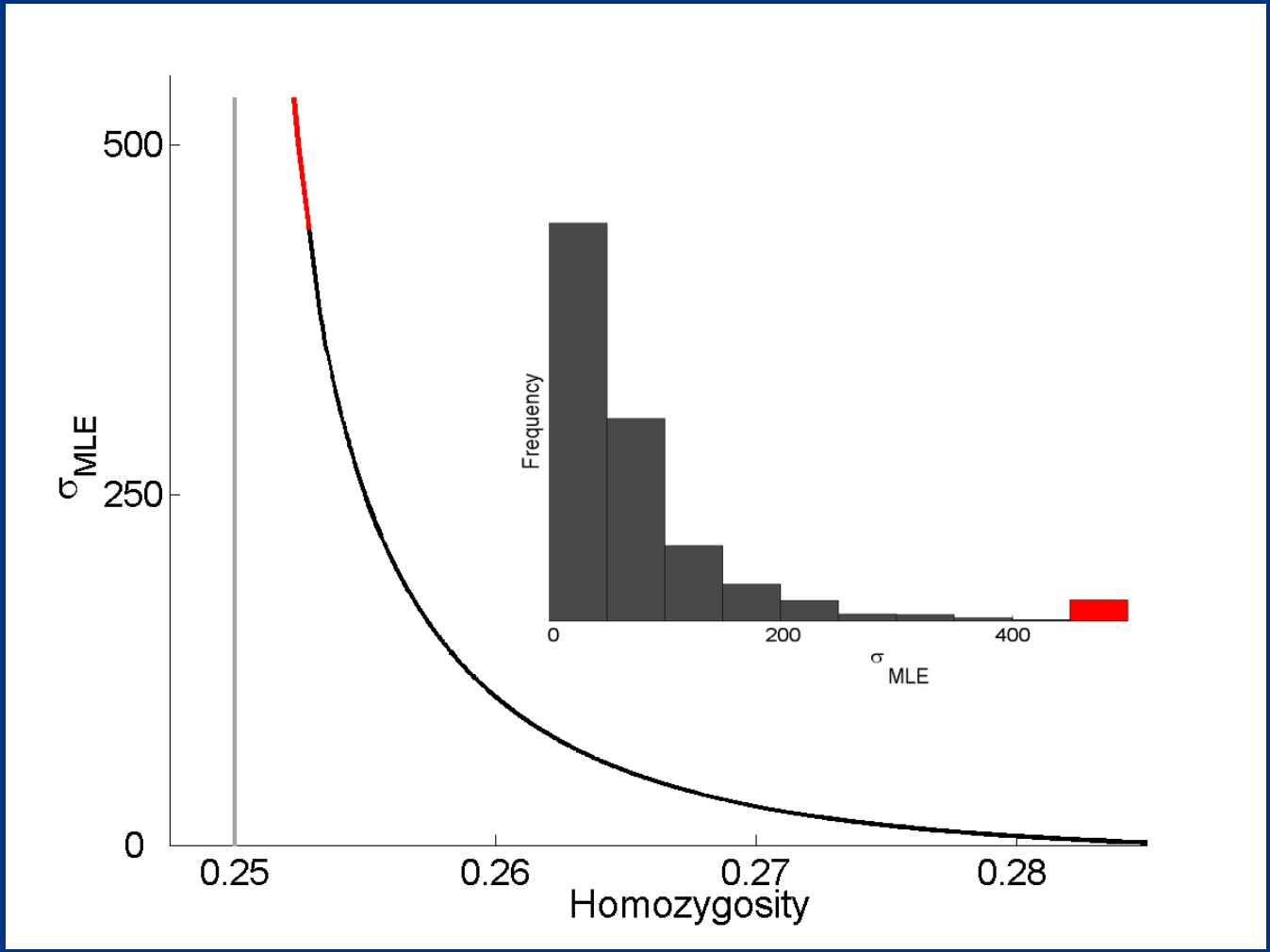
Lyme disease sample

The following data was collected by Qui *et al.* (1997) Hereditas **127**: 203-216 on *B. burgdorferi* (the cause of Lyme disease) from eastern Long Island, New York.

	relative frequency
1	0.10
2	0.37
3	0.26
4	0.27

The observed homozygosity is $h = 0.288$ relatively close to the minimum 0.25 under $k = 4$.

Lyme disease sample



$K = 4$

Lyme disease sample

The maximum likelihood estimate is

$\hat{\sigma} = 35.1$

Based on the simulated sampling distribution for $\hat{\sigma}$ we get an estimated standard error of 176.4.

The 2.5 percentile of the simulated sampling distribution of $\hat{\sigma}$ corresponds to 17.2 and the 97.5 percentile is 681.3. Therefore, an approximate 95% interval estimate based on the parametric bootstrap associated with $\hat{\sigma}$ is $(17.2, 681.3)$.

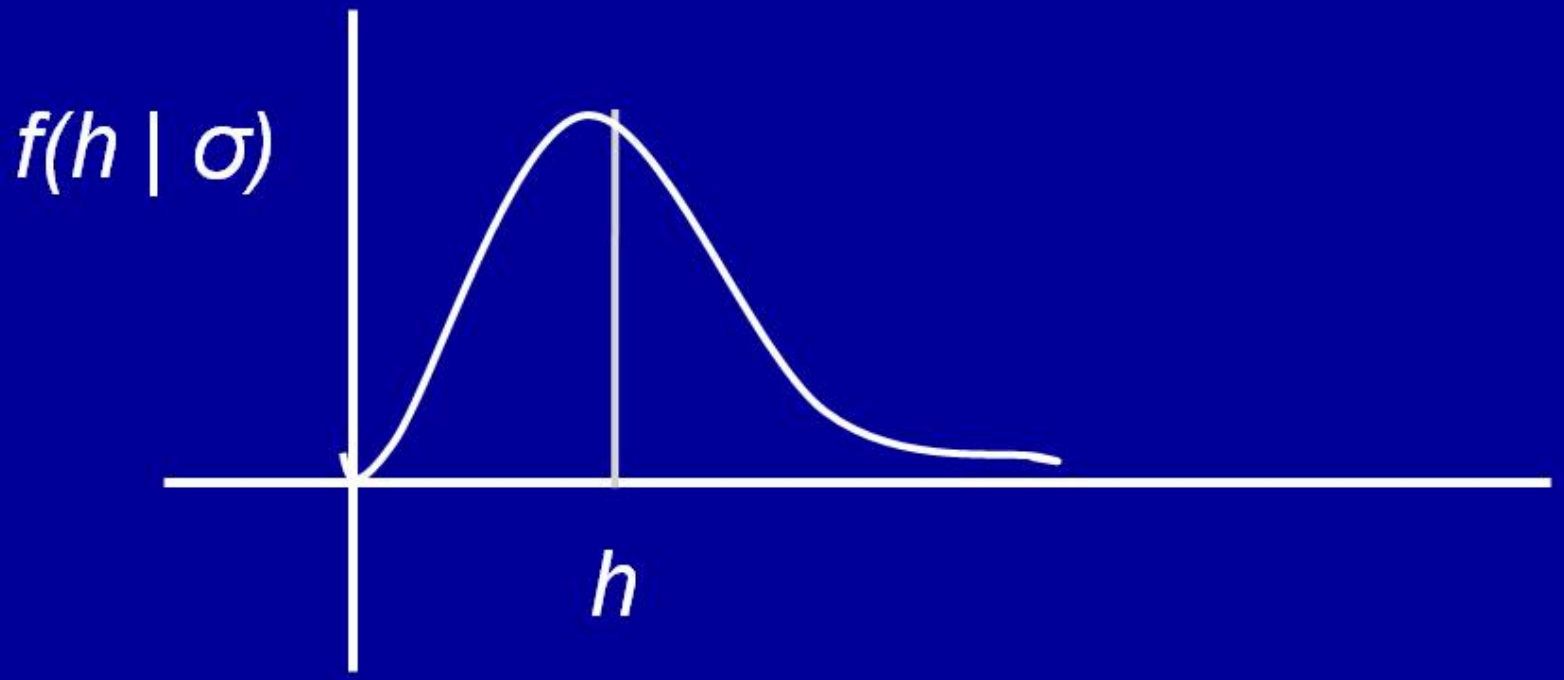
Exact Confidence Interval

For a given confidence level $(1-\alpha)$ and an observed homozygosity $H = h$, we choose $\hat{\sigma}_L$ and $\hat{\sigma}_U$ so that

$$F_{\mathrm{H}}(h|\hat{\sigma}_L) = \alpha_1, \quad F_{\mathrm{H}}(h|\hat{\sigma}_U) = 1 - \alpha_2, \tag{3}$$

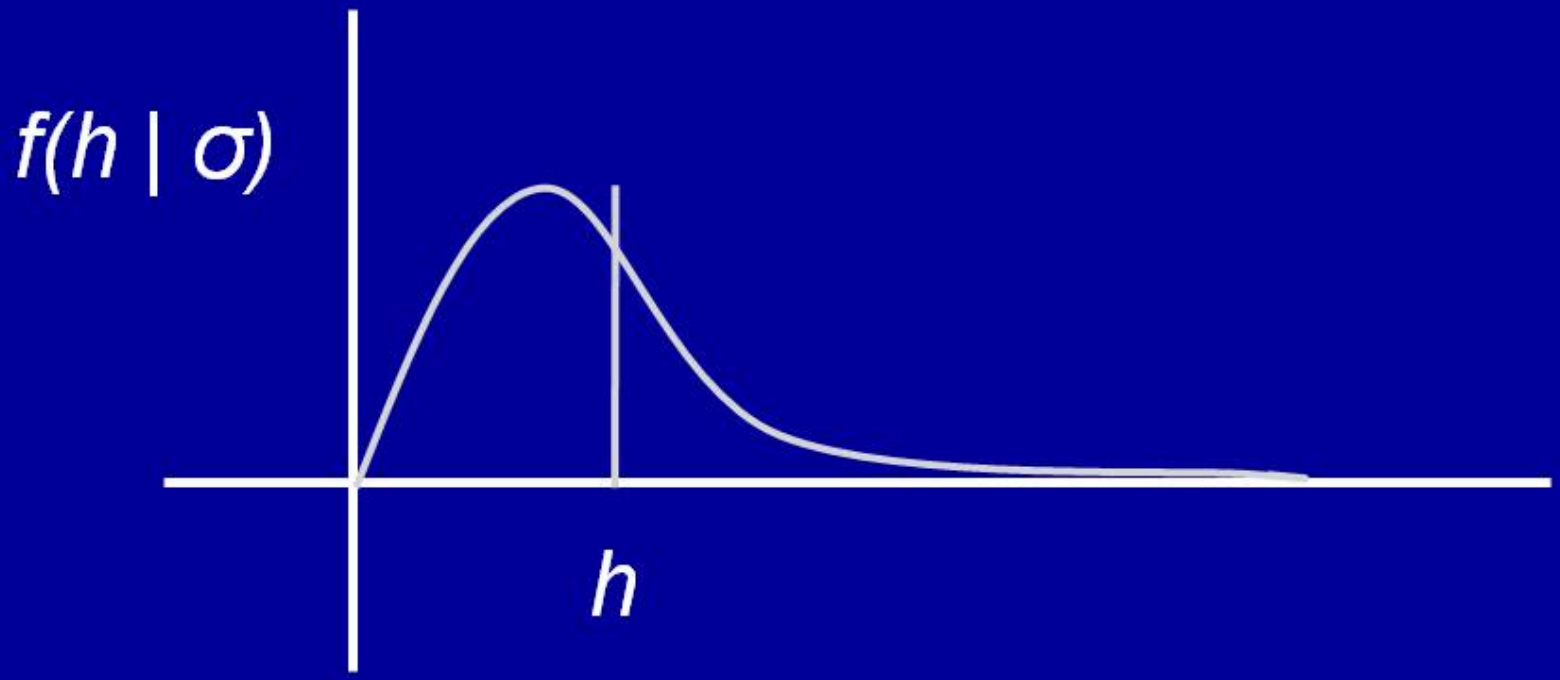
where $\alpha = \alpha_1 + \alpha_2$.

σ→ increase
decrease ←E(H|σ)



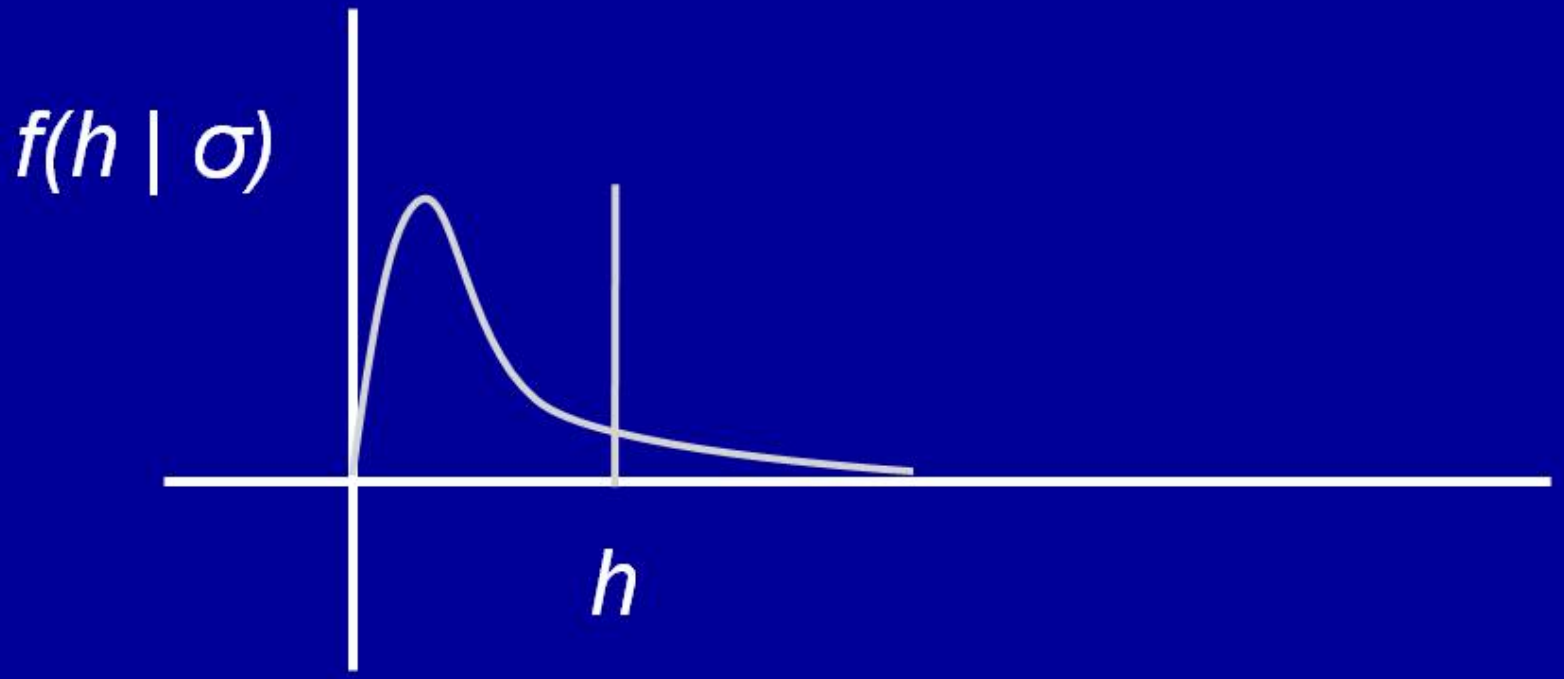
σ→ increase

decrease ←E(H|σ)



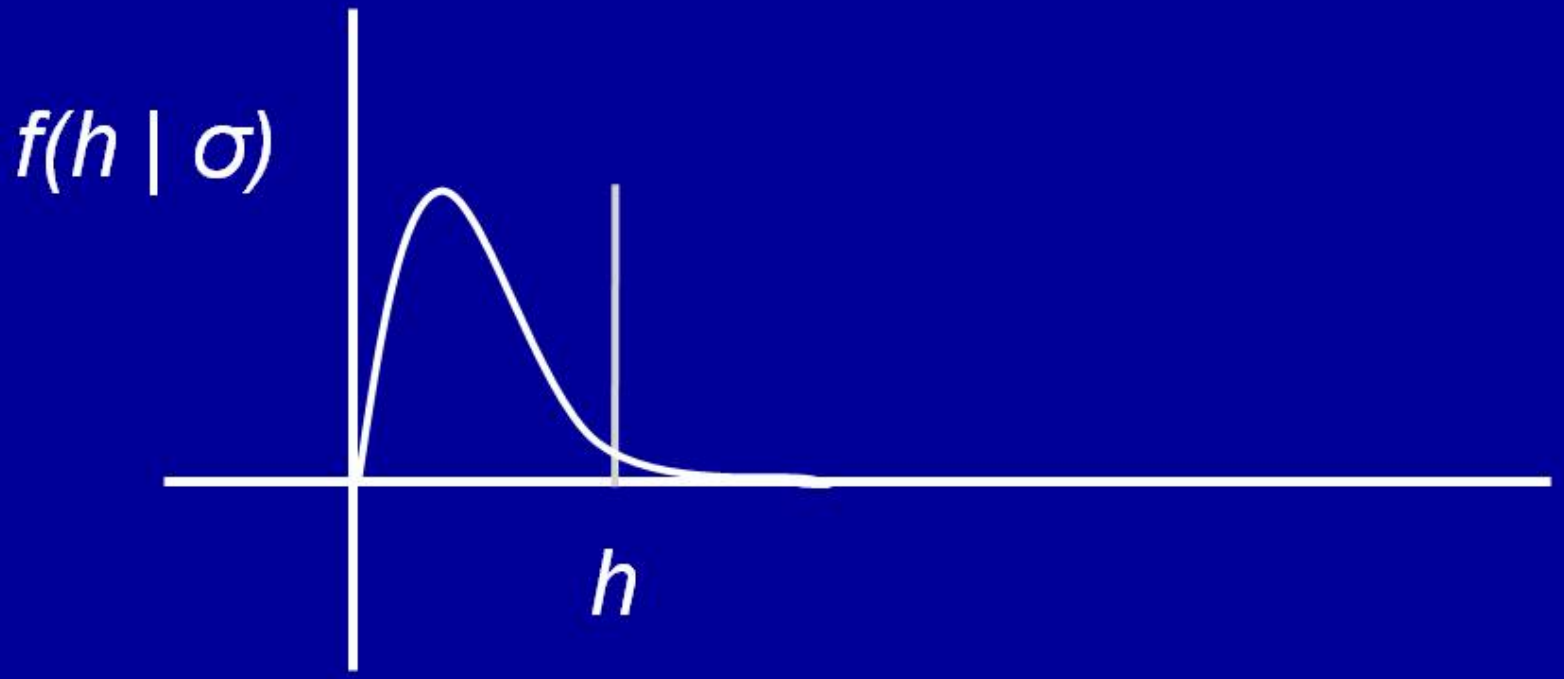
σ→ increase

decrease ←E(H|σ)

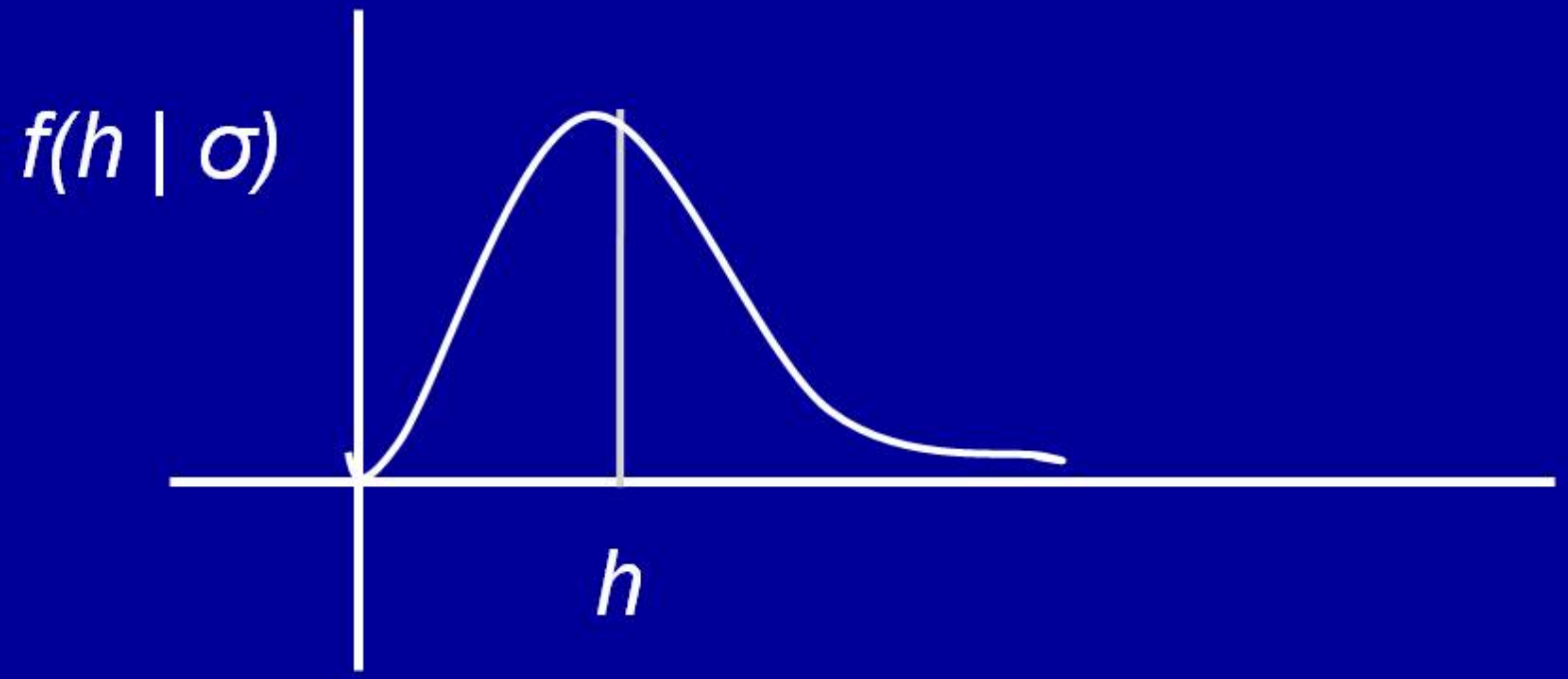


σ→ increase

decrease ←E(H|σ)

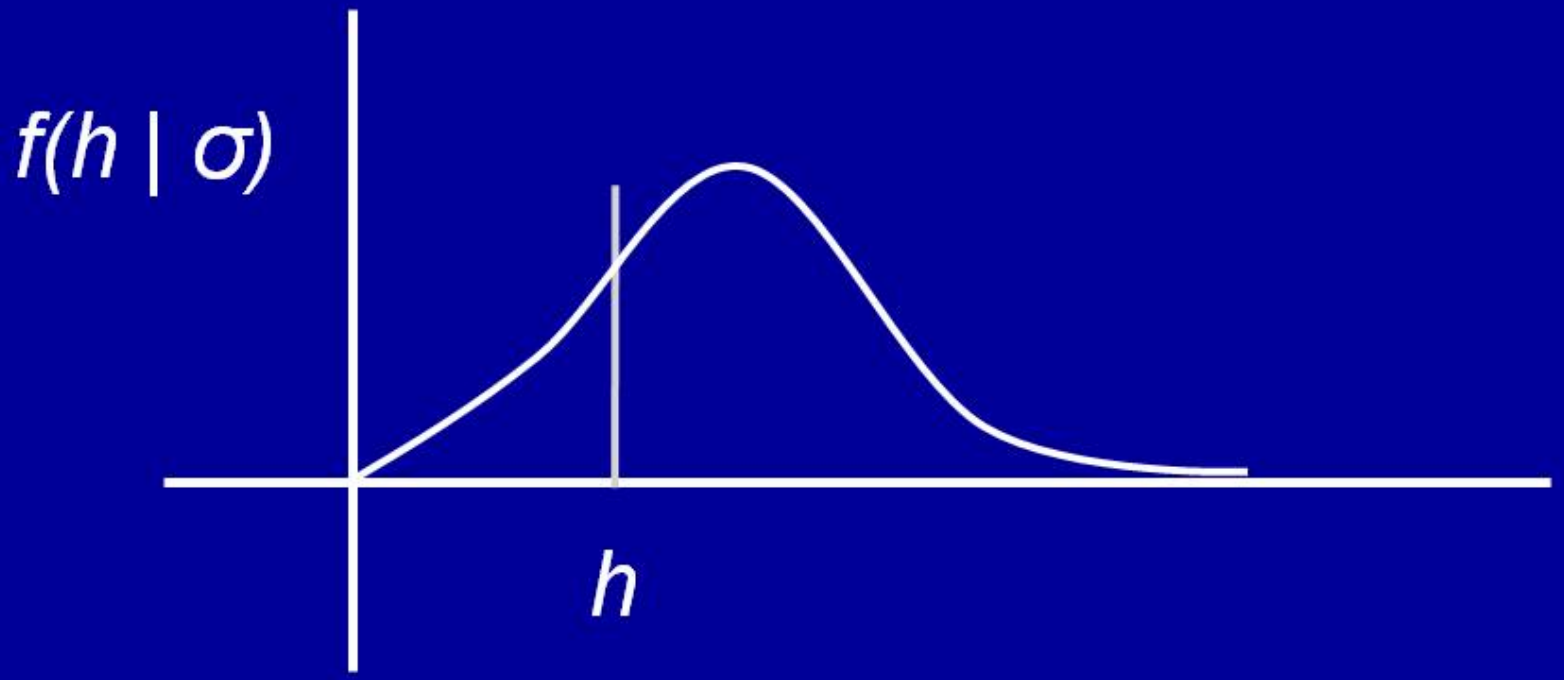


σ→ increase
decrease ←E(H|σ)



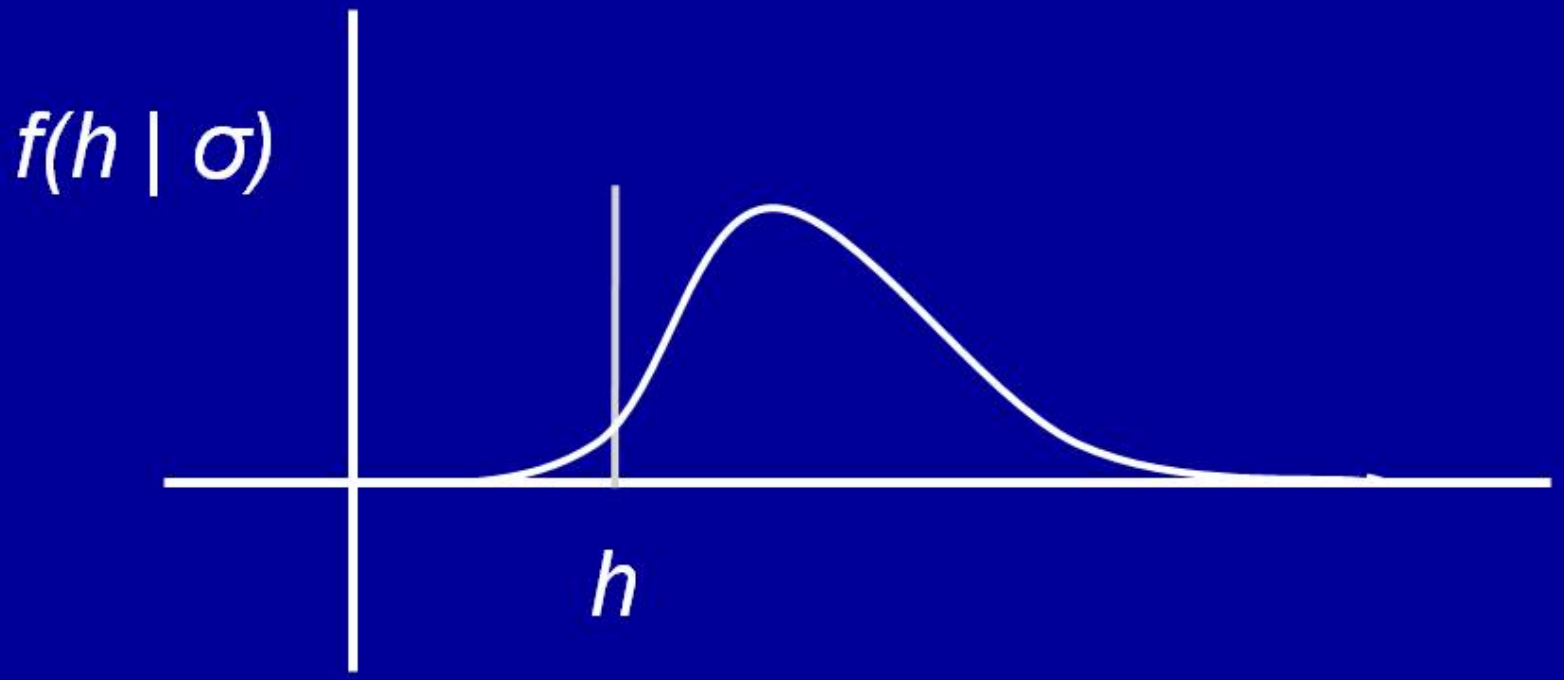
decrease ← σ

E(H|σ) → increase



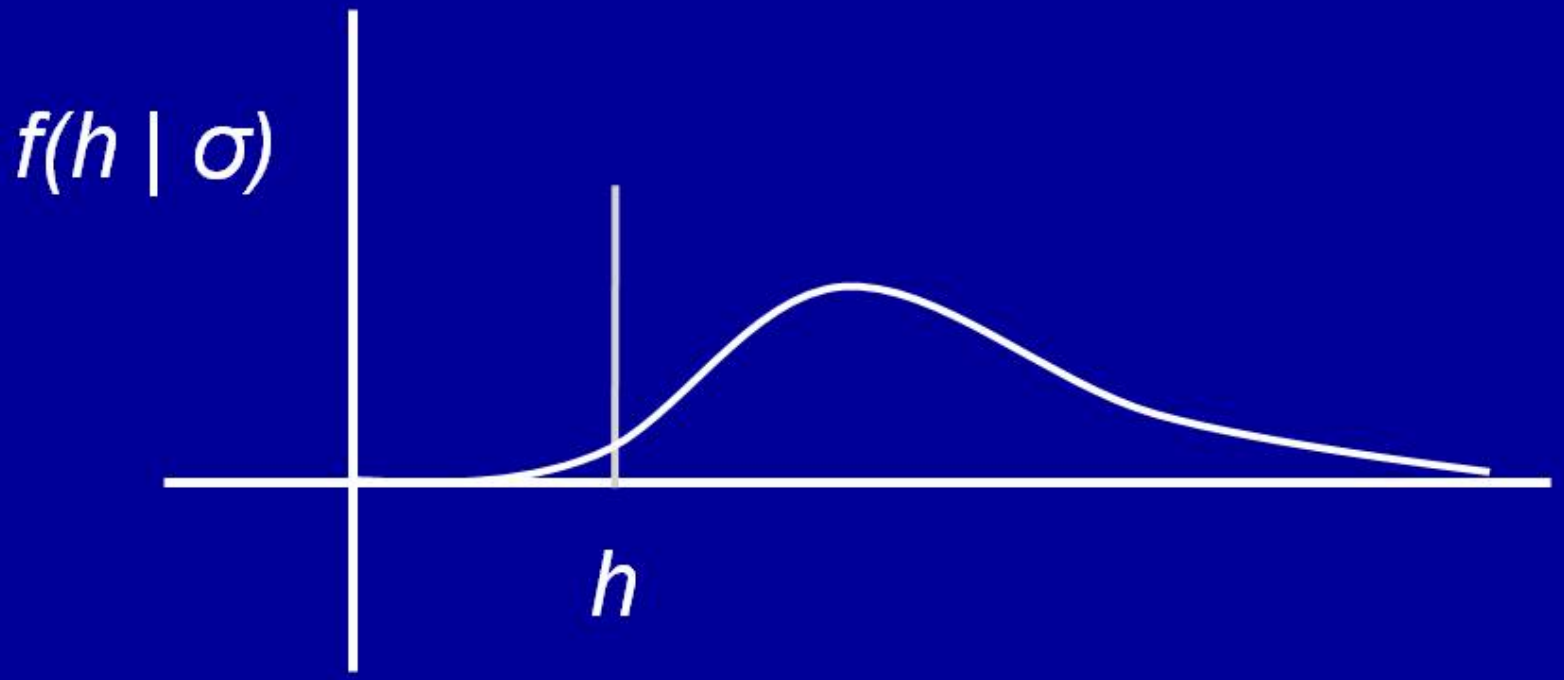
decrease ← σ

E(H|σ) → increase



decrease ← σ

E(H|σ) → increase



Reliability and Precision

True Confidence Level for Parametric Bootstrap for Lyme Disease Data

Recall that $\hat{\sigma}_{LB} = 17.2$ and $\hat{\sigma}_{UB} = 681.3$.

Using the monotonicity of the homozygosity gives $\alpha_{1B} = 0.354$, and $\alpha_{2B} < 0.001$. Thus the *true* confidence level is $1 - \alpha_{1B} - \alpha_{2B} \approx 0.65$.

Exact Confidence Interval for Lyme Disease Data

Using the monotonicity of the homozygosity method with $\alpha_1 = \alpha_2 = 0.025$ produces an exact 95% confidence interval of $(-8, 105)$ for the Lyme disease data.

Bayesian Approach

Assuming independent uniform priors on (θ, σ), the joint posterior distribution of (θ, σ) is proportional to the likelihood,

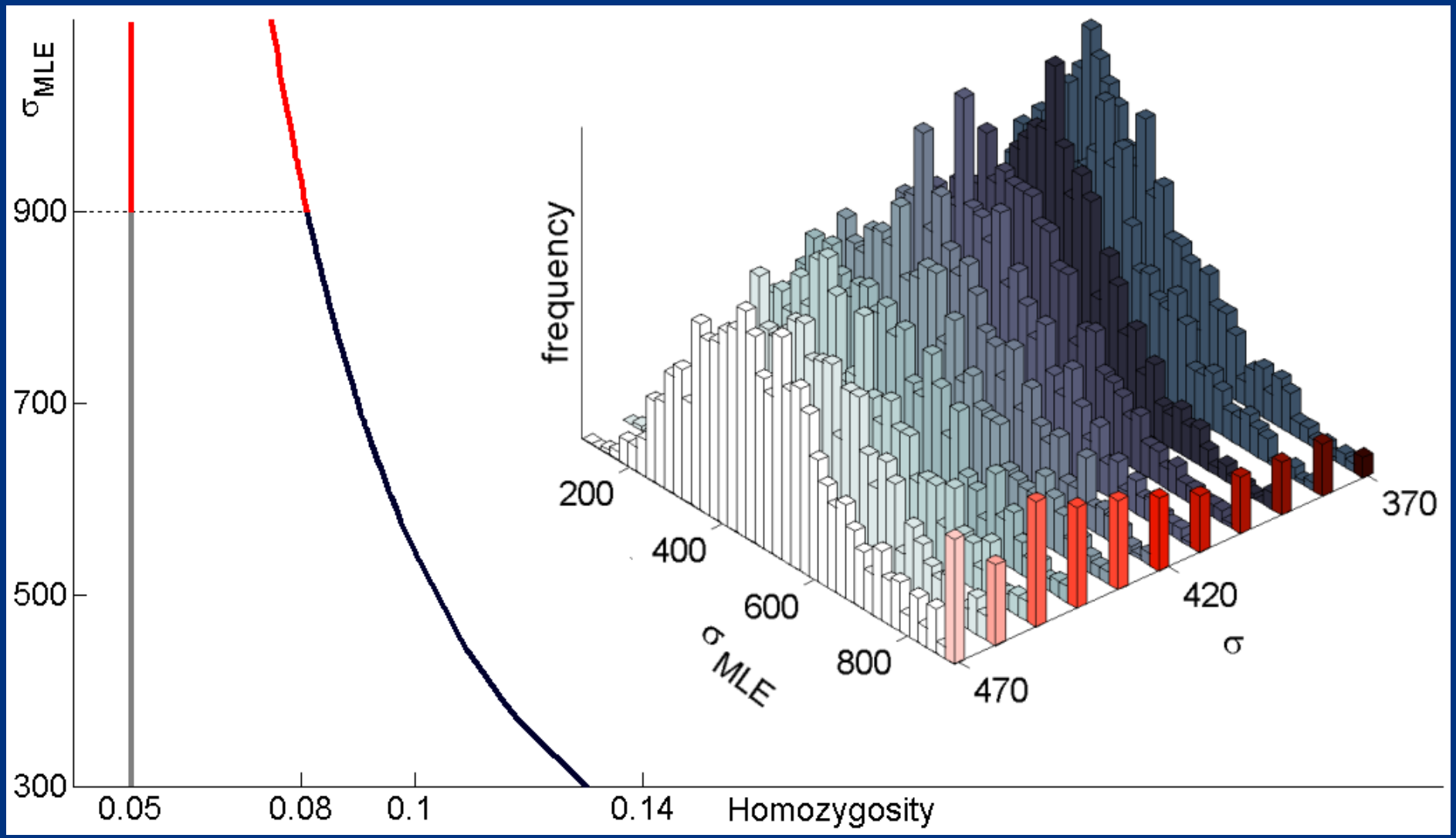
$$P_{\mathrm{Sel}}(\theta, \sigma|\mathbf{x}) \propto \frac{e^{-\sigma \sum_{i=1}^{k} x_i^2}}{E_{\mathrm{Neut}}\left(e^{-\sigma \sum_{i=1}^{k} X_i^2}\right)} (x_1 x_2 \cdots x_k)^{\theta/k-1}, \tag{4}$$

which can be sampled using a standard Markov Chain Monte Carlo approach.

Summary of Results from Lyme Disease Data

	interval estimate for σ	confidence/credibility
P – boot	$(17, 681)$	65%
Exact c.i.	$(-8, 105)$	95%
Bayesian	$(11, 125)$	95%

P-boot for $k = 20$



HLA Data

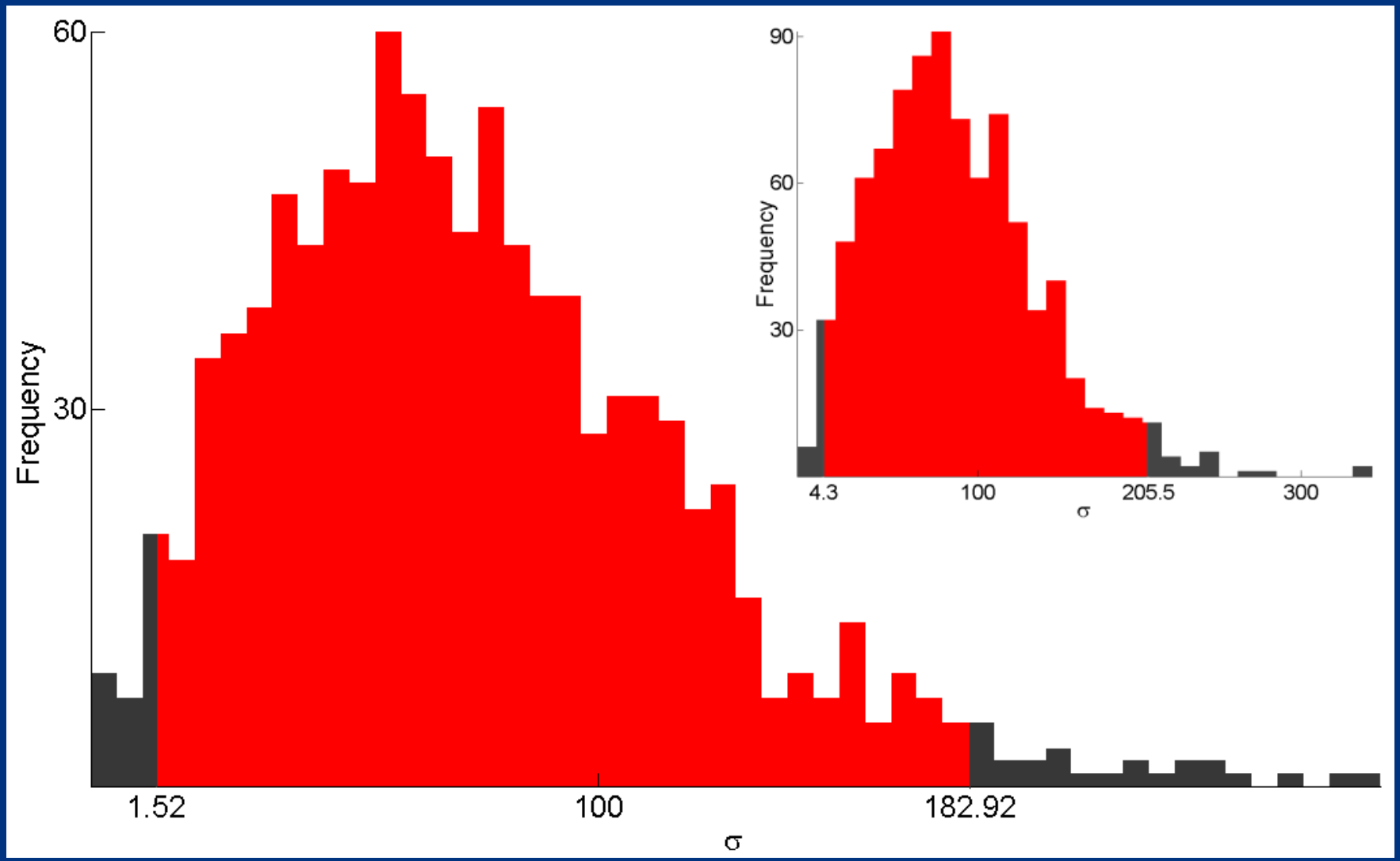
The population frequencies are given by $\mathbf{x}' = (0.22, 0.21, 0.17, 0.16, 0.15, 0.04, 0.03, 0.02)$.

The homozygosity statistic is $h = 0.172$, again close to the minimum $h_{\min} = 0.125$ for $k = 8$.

Summary of Results of HLA Data

	interval estimate for σ	confidence/credibility
P – boot	$((21, 396)$	70%
Exact c.i.	$(-10, 159)$	95%
Bayesian	$(6, 183)$	95%

Posterior Distributions



Full Model

$$f_{\mathrm{Sel}}(\mathbf{x}|\boldsymbol{\theta},\boldsymbol{\Sigma}) = \frac{e^{-\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}}}{E_{\mathrm{Neut}}(e^{-\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}})} f_{\mathrm{Neut}}(\mathbf{x}|\boldsymbol{\theta}) \tag{5}$$

$$f_{\mathrm{Neut}}(\mathbf{x}|\boldsymbol{\theta}) = \frac{\Gamma(\theta_1+\theta_2+\cdots+\theta_k)}{\Gamma(\theta_1)\Gamma(\theta_2)\cdots\Gamma(\theta_k)} x_1^{\theta_1-1} x_2^{\theta_2-1} \cdots x_k^{\theta_k-1}. \tag{6}$$

General Theorem

Theorem 2 *Consider the probability density function* $f_{\mathrm{Sel}}(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\Sigma})$ *defined by equation (5) that describes the distribution of allele frequencies at stationarity under the Wright-Fisher model with selection and parent independent mutation. There exists a vector of allele frequencies* $\mathbf{x}^* = (x_1^*, \cdots, x_k^*)'$ *where* $f_{\mathrm{Sel}}(\mathbf{x}^*|\boldsymbol{\theta}, \boldsymbol{\Sigma})$ *is unbounded as a function of* $\boldsymbol{\Sigma}$ *regardless of* θ

Conclusion

- Assuming a k allele model with heterozygote advantage, the maximum likelihood estimates coupled with the parametric bootstrap approach gives unreliable and imprecise interval estimates of the selection intensity. Even if the mutation parameter is assumed known.
- The problem is caused by a singularity in data space. Since the parametric bootstrap approach samples data space repeatedly, there is good chance of sampling near the singularity.
- Methods that vary the parameters and fix the data produce better estimates. The monotonicity method and the Bayesian method both have this property.