The bootstrap

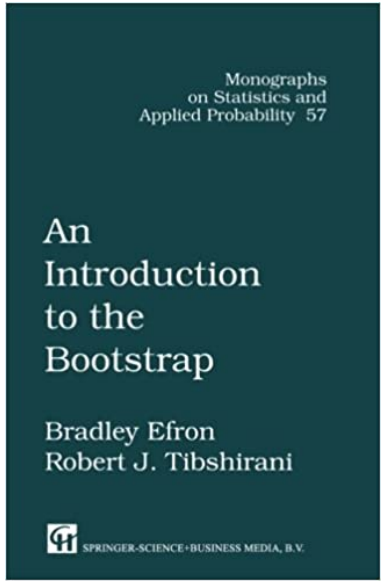
Confidence intervals for QTL location

Karl Broman

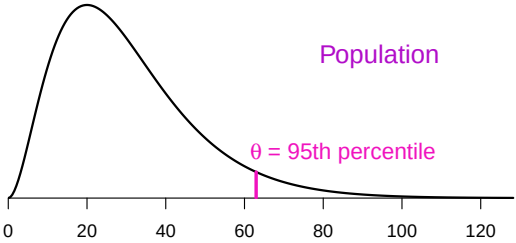
Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman

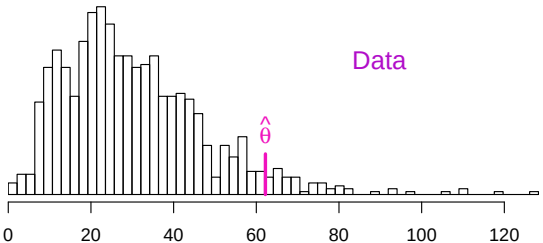
Course web: kbroman.org/AdvData


Monographs
on Statistics and
Applied Probability 57
An
Introduction
to the
Bootstrap
Bradley Efron
Robert J. Tibshirani
SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

Example

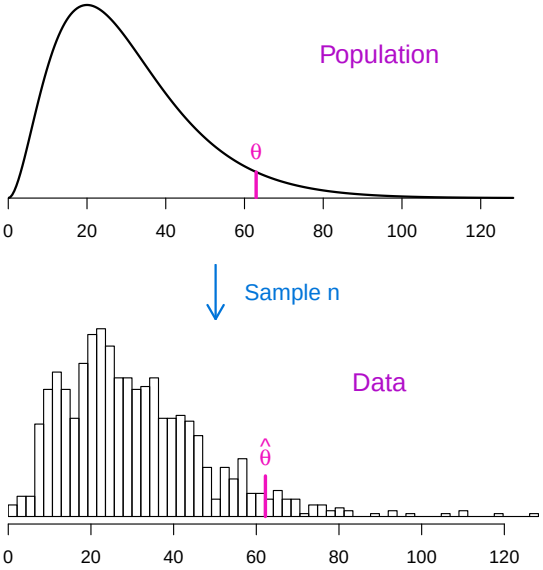


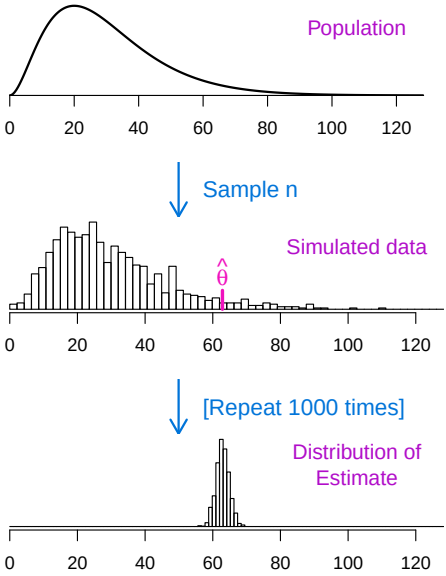
Sample n



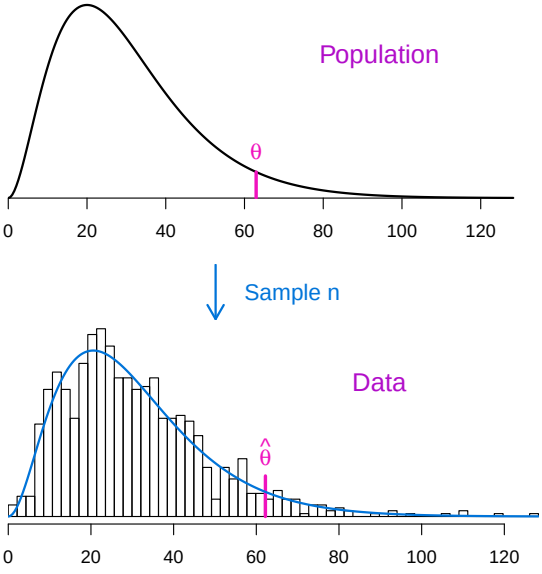
What is the standard error of $\hat{\theta}$?

Example

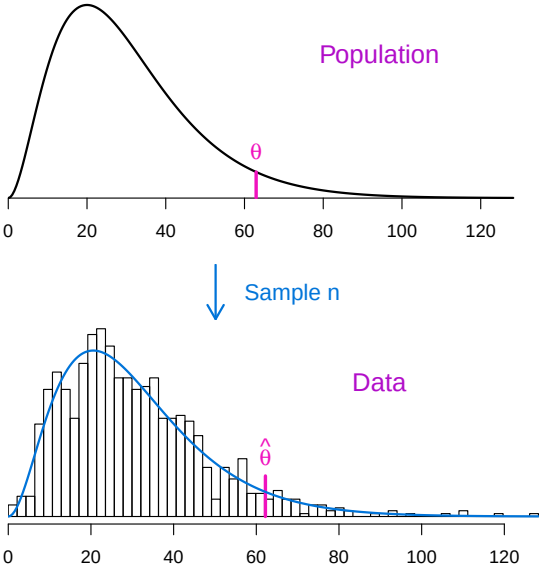

Population
θ
0
20
40
60
80
100
120
Sample n
Data
θ̂
0
20
40
60
80
100
120


Population
0
20
40
60
80
100
120
Sample n
Simulated data
θ̂
0
20
40
60
80
100
120
[Repeat 1000 times]
Distribution of
Estimate
0
20
40
60
80
100
120

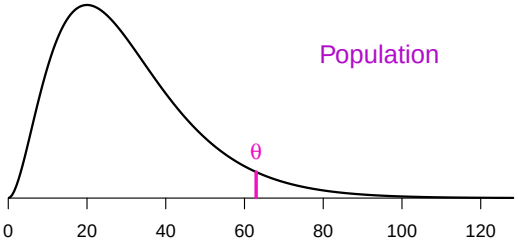
Parametric bootstrap

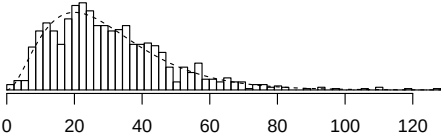


Parametric bootstrap

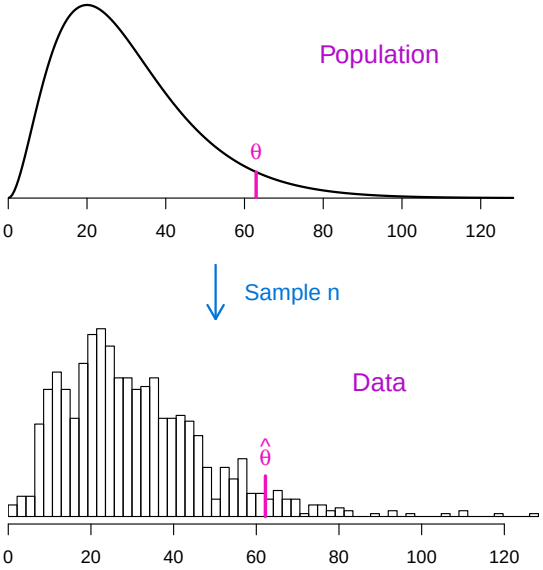


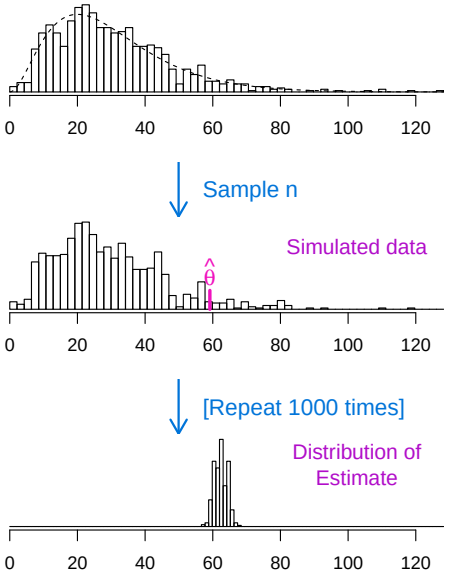
Non-parametric bootstrap





Non-parametric bootstrap





Don’t think “re-sampling”

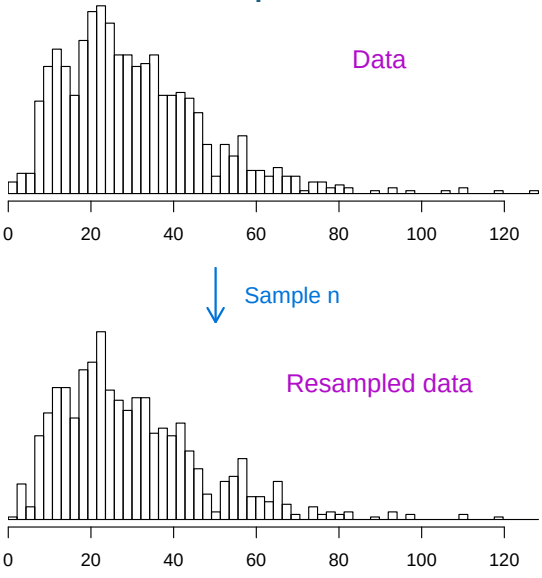
Think “simulate from an estimate of the population”

How can we tell if the bootstrap works?

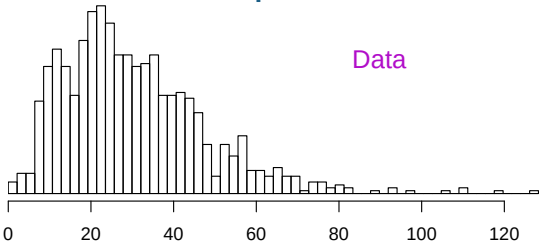
How can we tell if the bootstrap works?

Simulate!

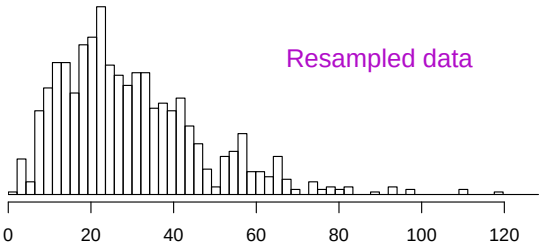
Nested bootstrap

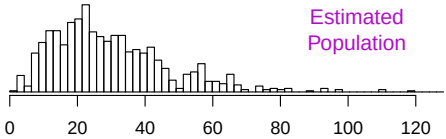


Nested bootstrap

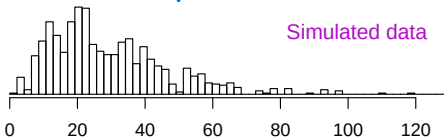


Sample n

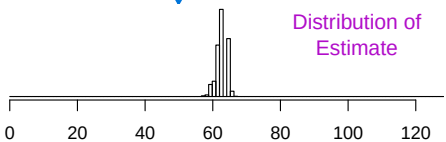




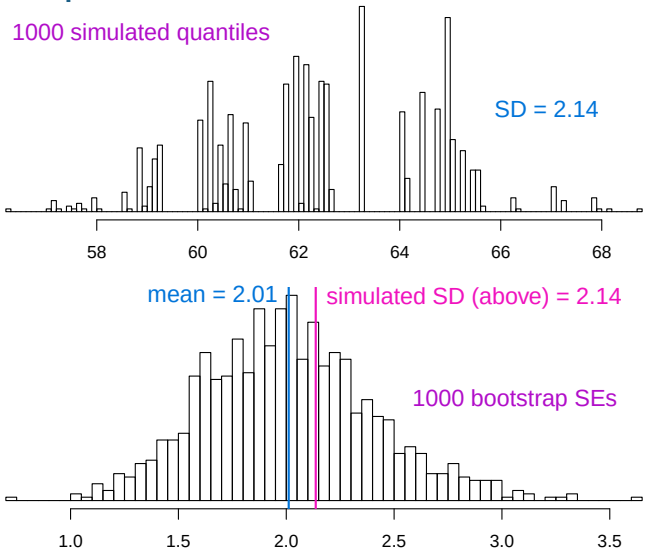
Sample n



[Repeat 1000 times]



Nested bootstrap results



Confidence Intervals in QTL Mapping by Bootstrapping

Peter M. Visscher, Robin Thompson and Chris S. Haley

Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, Scotland

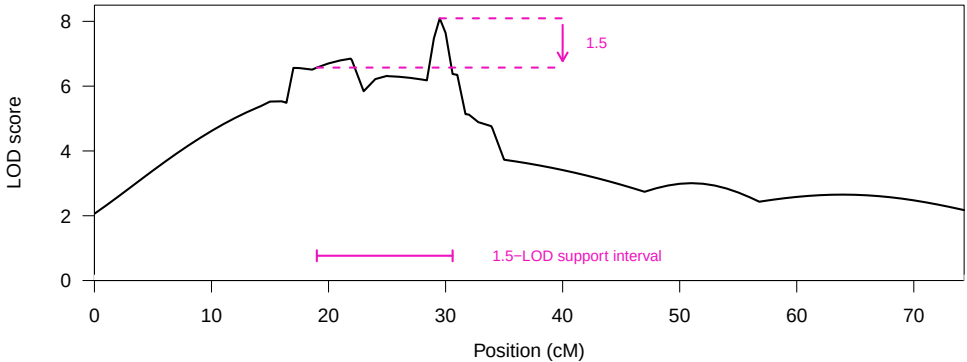
Manuscript received July 24, 1995
Accepted for publication February 24, 1996

ABSTRACT

The determination of empirical confidence intervals for the location of quantitative trait loci (QTLs) was investigated using simulation. Empirical confidence intervals were calculated using a bootstrap resampling method for a backcross population derived from inbred lines. Sample sizes were either 200 or 500 individuals, and the QTL explained 1, 5, or 10% of the phenotypic variance. The method worked well in that the proportion of empirical confidence intervals that contained the simulated QTL was close to expectation. In general, the confidence intervals were slightly conservatively biased. Correlations between the test statistic and the width of the confidence interval were strongly negative, so that the stronger the evidence for a QTL segregating, the smaller the empirical confidence interval for its location. The size of the average confidence interval depended heavily on the population size and the effect of the QTL. Marker spacing had only a small effect on the average empirical confidence interval. The LOD drop-off method to calculate empirical support intervals gave confidence intervals that generally were too small, in particular if confidence intervals were calculated only for samples above a certain significance threshold. The bootstrap method is easy to implement and is useful in the analysis of experimental data.

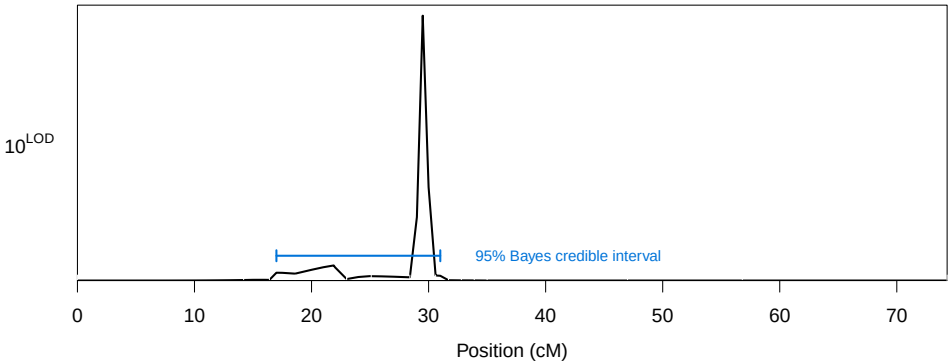
FOR many plant and animal species, genetic maps are available with a large number of highly polymorphic markers. These maps are useful in breeding programs. For example, when using markers to introgress a QTL allele from a donor population

LOD support interval



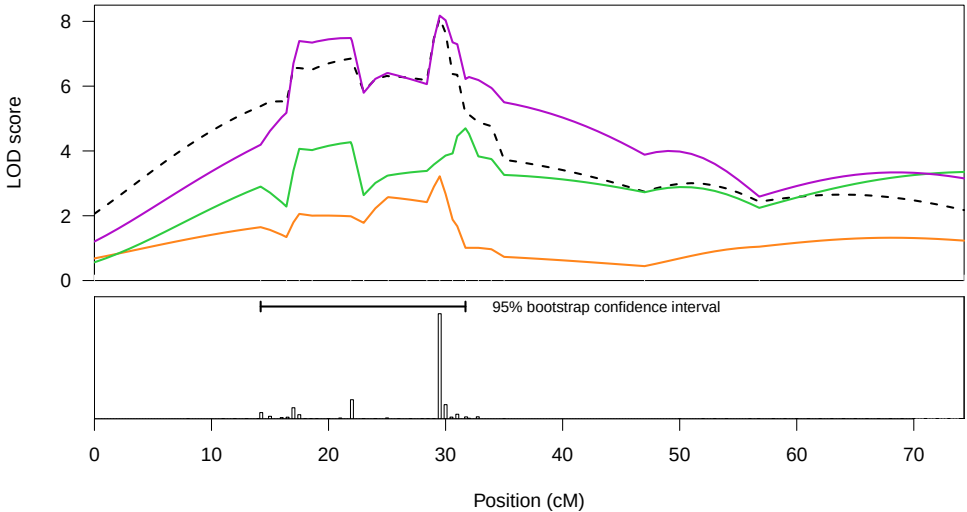
Data from Sugiyama et al. (2001) Genomics 71:70–77

Approximate Bayes interval



Data from Sugiyama et al. (2001) Genomics 71:70–77

Bootstrap CI



Data from Sugiyama et al. (2001) Genomics 71:70–77

DOI: 10.1534/genetics.106.061549

Poor Performance of Bootstrap Confidence Intervals for the Location of a Quantitative Trait Locus

Ani Manichaikul,* Josée Dupuis,† Śaunak Sen‡ and Karl W. Broman*,[1]

**Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205,* †*Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts 02118 and* ‡*Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94107*

Manuscript received March 24, 2006
Accepted for publication June 15, 2006

ABSTRACT

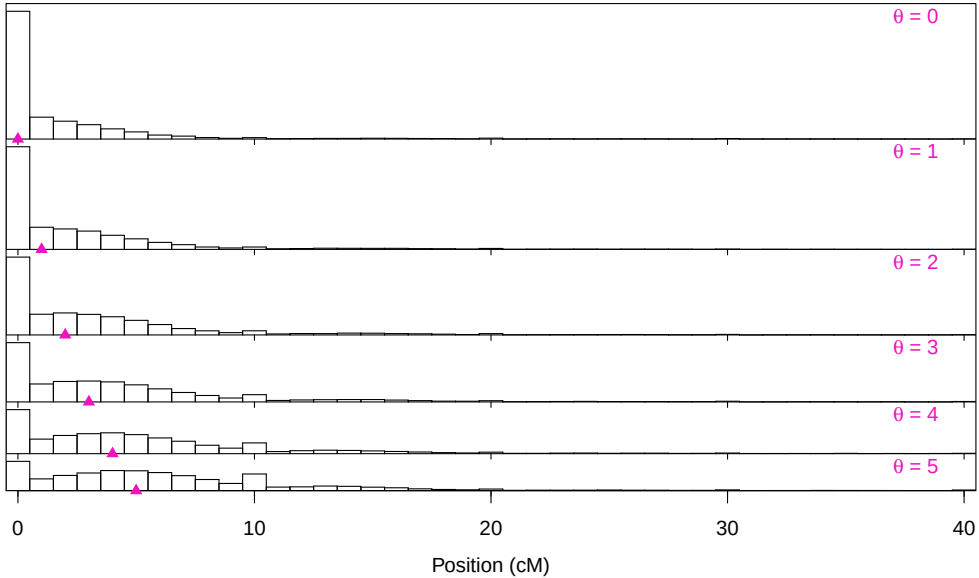
The aim of many genetic studies is to locate the genomic regions (called quantitative trait loci, QTL) that contribute to variation in a quantitative trait (such as body weight). Confidence intervals for the locations of QTL are particularly important for the design of further experiments to identify the gene or genes responsible for the effect. Likelihood support intervals are the most widely used method to obtain confidence intervals for QTL location, but the nonparametric bootstrap has also been recommended. Through extensive computer simulation, we show that bootstrap confidence intervals behave poorly and so should not be used in this context. The profile likelihood (or LOD curve) for QTL location has a tendency to peak at genetic markers, and so the distribution of the maximum-likelihood estimate (MLE) of QTL location has the unusual feature of point masses at genetic markers; this contributes to the poor behavior of the bootstrap. Likelihood support intervals and approximate Bayes credible intervals, on the other hand, are shown to behave appropriately.

THERE is much interest in mapping the genetic loci (called quantitative trait loci, QTL) that contribute to variation in a quantitative trait (such as body weight). Confidence intervals for the locations of QTL are particularly important for the design of further experiments to identify the gene or genes responsible for the effect. Likelihood support intervals are the most widely used method to obtain confidence intervals for QTL location, but the nonparametric bootstrap has also been recommended. Through extensive computer simulation, we show that bootstrap confidence intervals behave poorly and so should not be used in this context. The profile likelihood (or LOD curve) for QTL location has a tendency to peak at genetic markers, and so the distribution of the maximum-likelihood estimate (MLE) of QTL location has the unusual feature of point masses at genetic markers; this contributes to the poor behavior of the bootstrap. Likelihood support intervals and approximate Bayes credible intervals, on the other hand, are shown to behave appropriately.

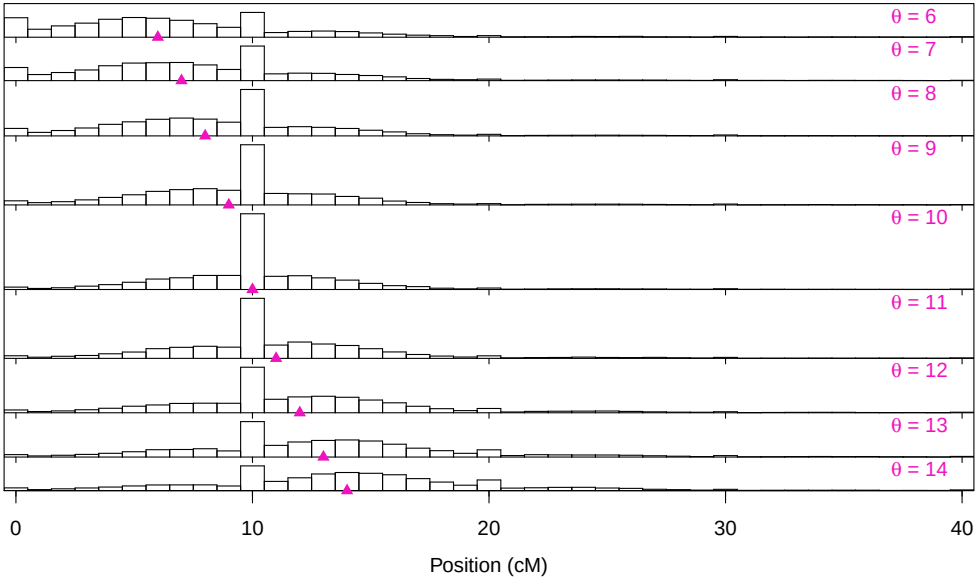
Simulation study

- Backcross, 200 individuals
- One chromosome of length 100 cM
- Markers at 10 cM spacing
- Single QTL responsible for 10% of phenotypic variance
- Normally distributed residual variation
- Varied location of QTL, at positions 0, 1, …, 100 cM
- Analysis by standard interval mapping; calculations every 1 cM
- 10,000 simulations for each QTL position
- Bootstrap used 1000 replicates

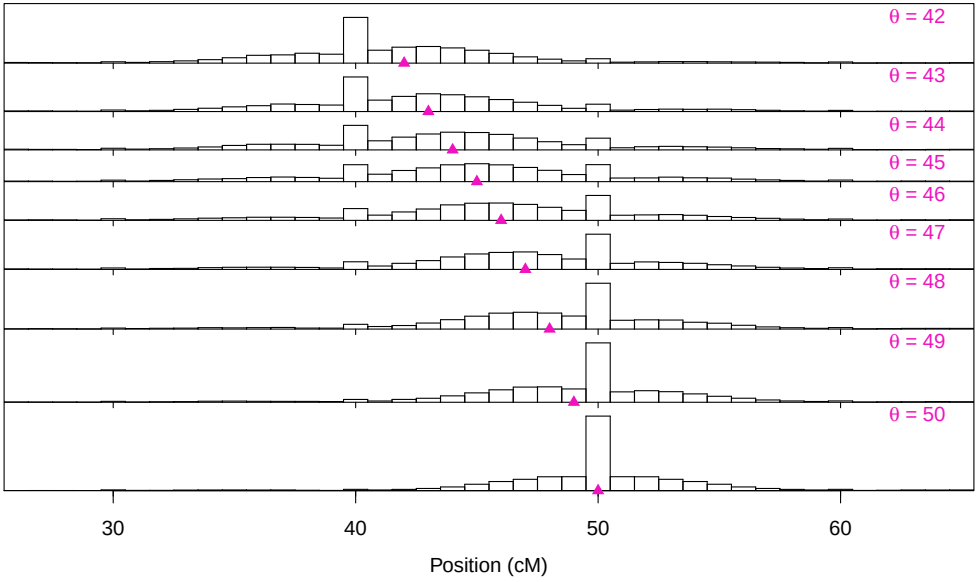
Distribution of estimated location



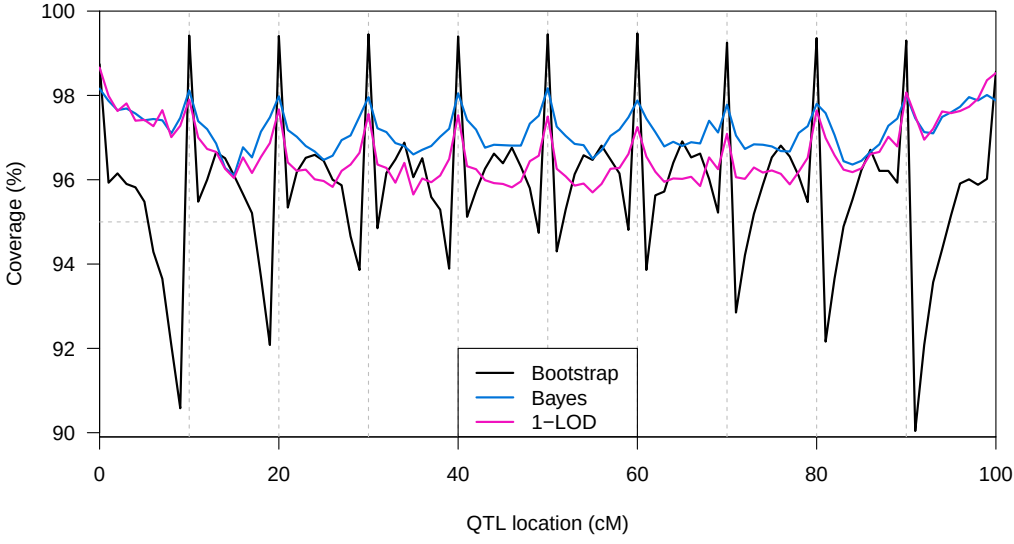
Distribution of estimated location



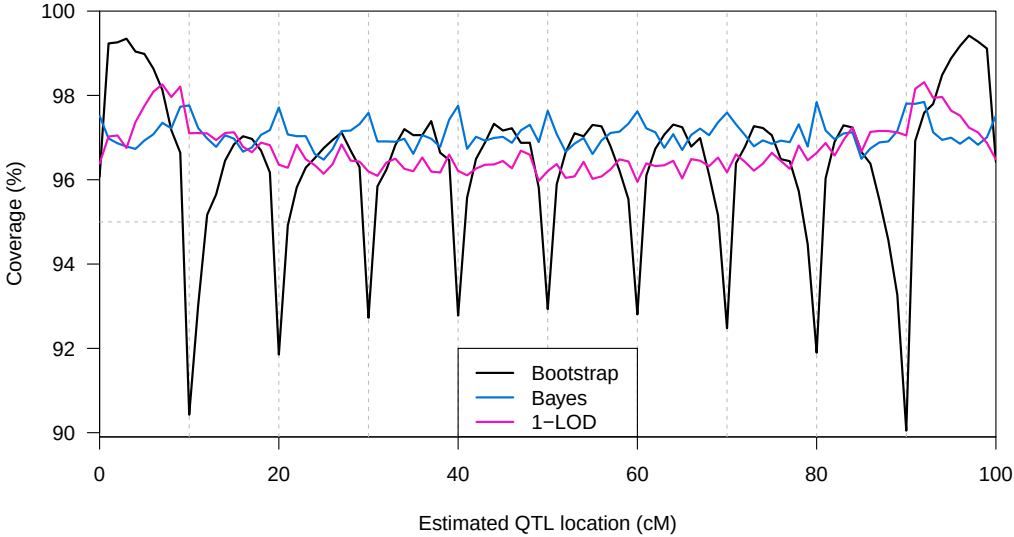
Distribution of estimated location



Coverage vs. true location



Coverage vs. estimated location



Summary

- The bootstrap can be super useful
- But it can also behave badly

  You need the distribution of $\hat{\theta} - \theta$ to not depend on θ
- If results look wonky, maybe you shouldn't trust them
- How to tell if the bootstrap works? Simulate!

- The odd tendency for the estimated QTL location to be at a marker messes up the bootstrap.