Introduction to computational statistics

Brian Kissmer

USU Department of Biology

Oct. 10th, 2024

Learning objectives

1. Be able to describe what bootstrapping and permutation tests do, and how they work
2. Be able to perform bootstrapping and permutation in R

Today's outline

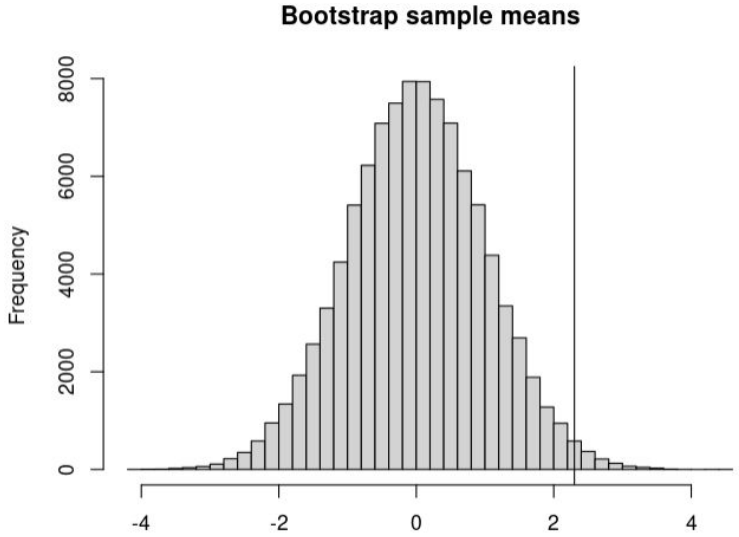
1. Computational statistics
2. Bootstrapping and permutation
3. Practice in R

Where we've been, where we're going

1. We have so far focused on simulations in biology, exploring their structure and functionality, with an emphasis on stochastic simulations
2. For the rest of the course, we will consider the use of computers in biological analysis, including:

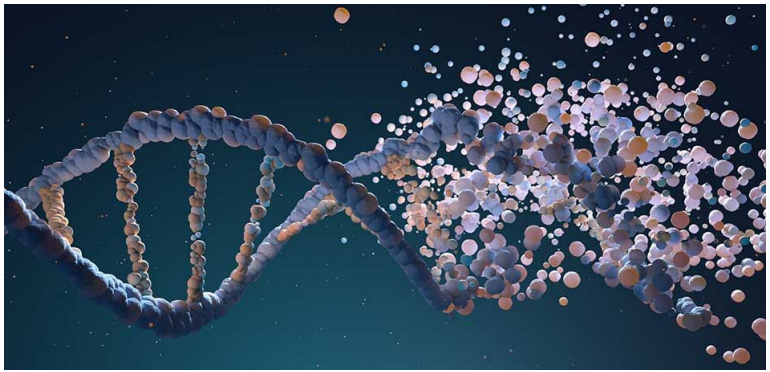
Where we've been, where we're going

1. We have so far focused on simulations in biology, exploring their structure and functionality, with an emphasis on stochastic simulations
2. For the rest of the course, we will consider the use of computers in biological analysis, including:
   a. Computational statistics



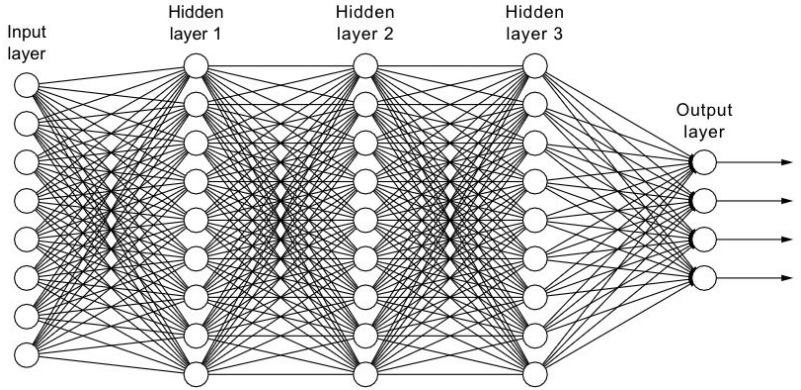
Where we've been, where we're going

1. We have so far focused on simulations in biology, exploring their structure and functionality, with an emphasis on stochastic simulations
2. For the rest of the course, we will consider the use of computers in biological analysis, including:
   a. Computational statistics
   b. Bioinformatics (genomic data analysis)



Where we've been, where we're going

1. We have so far focused on simulations in biology, exploring their structure and functionality, with an emphasis on stochastic simulations
2. For the rest of the course, we will consider the use of computers in biological analysis, including:
   a. Computational statistics
   b. Bioinformatics (genomic data analysis)
   c. Machine learning



What is computational statistics?

Computational statistics is the field that combines statistics and computer science to transform data into knowledge through computationally intensive algorithms, or when big data are involved.

- Resampling methods
- Permutation/randomization methods
- Numerical optimization
- Markov chain Monte Carlo
- Neural networks and machine learning

Opening discussion

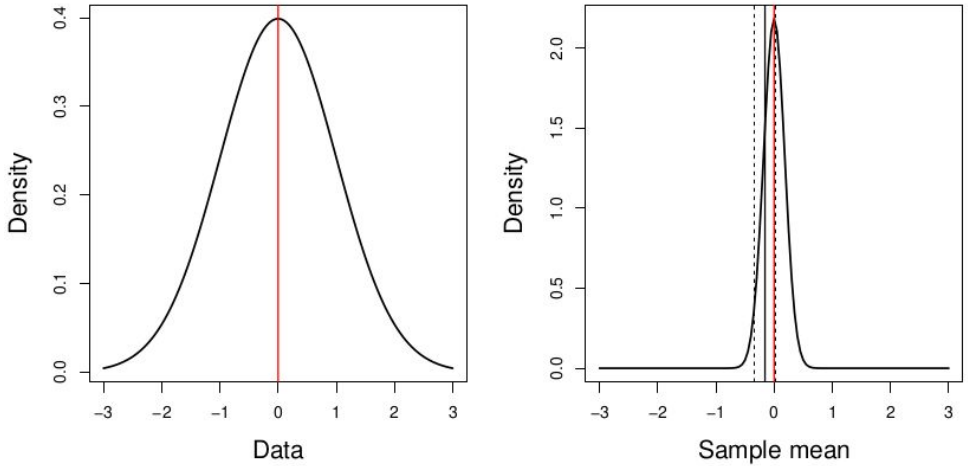
How do computers help with data analysis? Do computers simply make hard things easier (i.e., perform calculations that could be done by hand), or do they let one do things that would otherwise be impossible, or at least nearly impossible? You have 4–5 minutes to discuss these questions. Summarize your groups thoughts on the provided index card.

Bootstrapping

Bootstrapping is a statistical procedure that re-samples a single dataset to create many simulated samples. This process allows for the calculation of standard errors, confidence intervals, and hypothesis testing.

Standard error (SE): traditional approach

Sampling distribution = theoretical set of all possible estimates, approximately normal



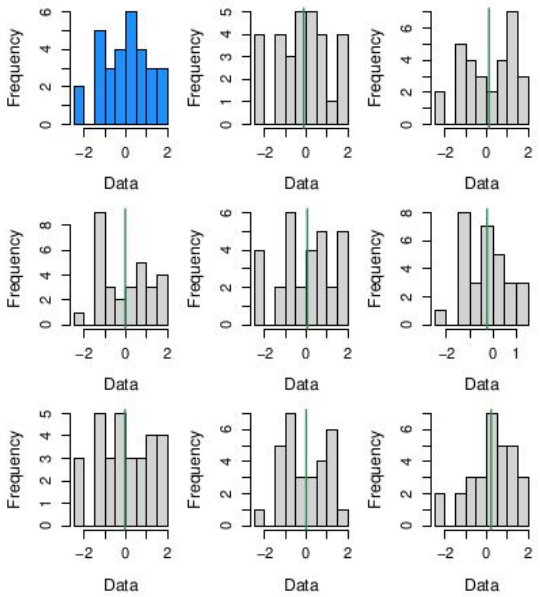
$\sigma = 1,\ n = 30$

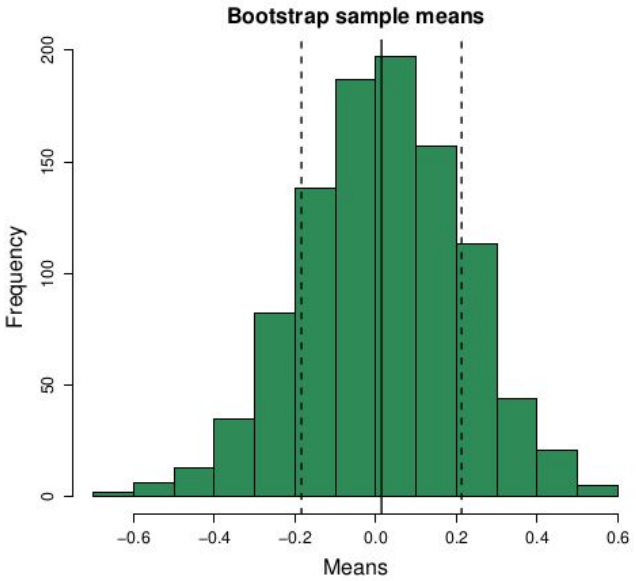
SE = average deviation between true mean and sample mean = $\sigma / \sqrt{n}$

Standard error (SE): traditional approach

Re-sample the sample to approximate the theoretical sampling distribution

SE = standard deviation of bootstrap sample means





Standard error (SE): traditional approach

Re-sample the sample to approximate the theoretical sampling Distribution

1. Standard errors on an estimate = standard deviation of the bootstrap estimates
2.

Standard error (SE): traditional approach

Re-sample the sample to approximate the theoretical sampling Distribution

1. Standard errors on an estimate = standard deviation of the bootstrap estimates
2. Confidence intervals on an estimate = empirical quantiles of the distribution of bootstrap sample estimates
3.

Standard error (SE): traditional approach

Re-sample the sample to approximate the theoretical sampling Distribution

1. Standard errors on an estimate = standard deviation of the bootstrap estimates
2. Confidence intervals on an estimate = empirical quantiles of the distribution of bootstrap sample estimates
3. Null hypothesis tests = you can reject a null value for a parameter with some level of confidence if it is not included in the relevant confidence interval

Bootstrapping null hypothesis example

This R function takes a vector of data, a confidence level, and null value and returns whether the null value falls within the bootstrap confidence interval.

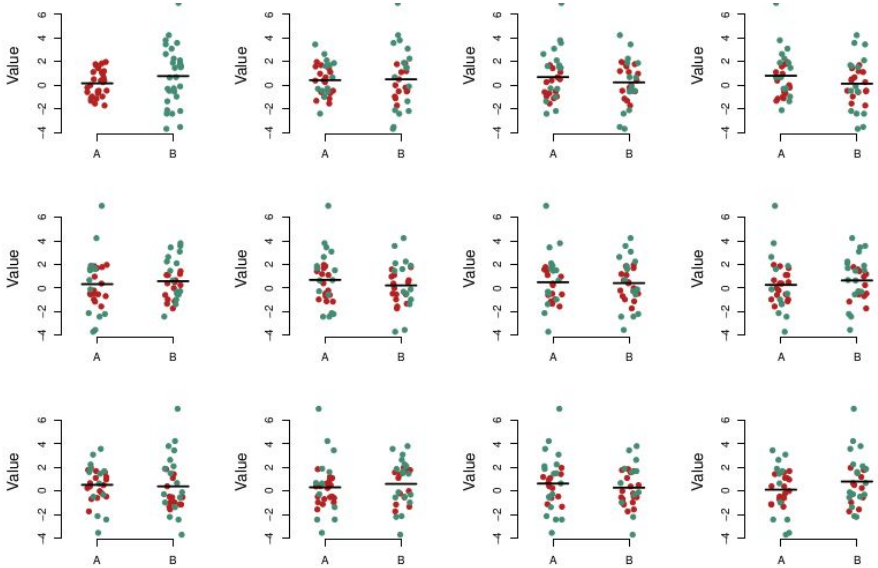
```
bootTest<-function(X=NA,conf=.95,nullVal=0){
    est<-rep(NA,1000)
    for(i in 1:1000){
        est[i]<-mean(sample(X,length(X),replace=TRUE))
    }
    lb<-quantile(est,probs=(1-conf)/2)
    ub<-quantile(est,1-(1-conf)/2)
    if(nullVal >= lb & nullVal <= ub){
        return("Fail to reject")
    } else{ return("Reject")}
}
```

Permutation tests

A permutation test or randomization test determines whether apparent patterns in data could arise by chance. The general algorithm is:

1. Compute a test statistic on the data, e.g., difference in means, correlation, etc.
2. Repeatedly randomize (permute) the labels (treatments) or covariates and recalculate the statistic
3. Use this null distribution to determine (with some level of confidence) whether the observed data can be explained by chance under the null hypothesis

Permutation tests



Permuting treatment labels (A vs. B) generates a null distribution for differences in means

R code for bootstrapping and permutations

See the handout for this week