MT 4113: Computing in Statistics

Computer intensive statistics
Lecture 7: The Bootstrap

Len Thomas

15 Oct 2018

Outline

Repeated sampling: the basis for inference

In classical ("frequentist") statistics, we rely on the concept of repeated sampling to make inferences about the quantity of interest – e.g., the population mean.

Repeated sampling: the basis for inference

In classical ("frequentist") statistics, we rely on the concept of repeated sampling to make inferences about the quantity of interest – e.g., the population mean.

- Hypothesis testing: if I could repeat the experiment many times and H0 was true, what proportion of sample means would be as extreme or more extreme than my value?

Repeated sampling: the basis for inference

In classical ("frequentist") statistics, we rely on the concept of repeated sampling to make inferences about the quantity of interest – e.g., the population mean.

- Hypothesis testing: if I could repeat the experiment many times and H0 was true, what proportion of sample means would be as extreme or more extreme than my value?
- Confidence intervals: if I could repeat the experiment many times, what interval would contain the population mean a specified proportion of times?

Repeated sampling: the basis for inference

In both cases, the “ideal” would be to have multiple replicate datasets.

But, we only have the one dataset to use...

Repeated sampling: the basis for inference

In both cases, the "ideal" would be to have multiple replicate datasets.

- Hypothesis testing: same data generating process as our data, but where H0 is true. Can then see how extreme our sample mean is.

But, we only have the one dataset to use...

Repeated sampling: the basis for inference

In both cases, the “ideal” would be to have multiple replicate datasets.

- Hypothesis testing: same data generating process as our data, but where H0 is true. Can then see how extreme our sample mean is.
- Confidence intervals:

But, we only have the one dataset to use...

Repeated sampling: the basis for inference

In both cases, the "ideal" would be to have multiple replicate datasets.

- Hypothesis testing: same data generating process as our data, but where H0 is true. Can then see how extreme our sample mean is.
- Confidence intervals:
  - Invert the test statistic: same data generating process as our data, but over a range of hypothesized means. Can see which hypothesized means are plausible – those where our sample mean is not extreme.

But, we only have the one dataset to use...

Repeated sampling: the basis for inference

In both cases, the "ideal" would be to have multiple replicate datasets.

- Hypothesis testing: same data generating process as our data, but where H0 is true. Can then see how extreme our sample mean is.
- Confidence intervals:
  - Invert the test statistic: same data generating process as our data, but over a range of hypothesized means. Can see which hypothesized means are plausible – those where our sample mean is not extreme.
  - Alternative: exact same data generating process as our data – i.e. with the same true population mean. Can generate a distribution of plausible sample means and assume they represent the distribution of plausible population means.

But, we only have the one dataset to use...

University of St Andrews

Computer-intensive inference

Use properties of the dataset we have to help us *simulate* repeated datasets.

Note, simulations can be:

Computer-intensive inference

Use properties of the dataset we have to help us *simulate* repeated datasets.

- Hypothesis testing – Monte Carlo tests – simulate data with H0 true.

Note, simulations can be:

Computer-intensive inference

Use properties of the dataset we have to help us *simulate* repeated datasets.

- Hypothesis testing – Monte Carlo tests – simulate data with H0 true.
- Confidence intevals:

Note, simulations can be:

Computer-intensive inference

Use properties of the dataset we have to help us *simulate* repeated datasets.

- Hypothesis testing – Monte Carlo tests – simulate data with H0 true.
- Confidence intevals:
    - Invert test statistic – simulate data with a range of values of H0.

Note, simulations can be:

University of St Andrews

Computer-intensive inference

Use properties of the dataset we have to help us *simulate* repeated datasets.

- Hypothesis testing – Monte Carlo tests – simulate data with H0 true.
- Confidence intevals:
  - Invert test statistic – simulate data with a range of values of H0.
  - Simulate data as close as possible to the data generating process – "resamples". This is the bootstrap.

Note, simulations can be:

Computer-intensive inference

Use properties of the dataset we have to help us *simulate* repeated datasets.

- Hypothesis testing – Monte Carlo tests – simulate data with H0 true.
- Confidence intevals:
  - Invert test statistic – simulate data with a range of values of H0.
  - Simulate data as close as possible to the data generating process – "resamples". This is the bootstrap.

Note, simulations can be:

- Nonparametric: sample from the data to create new datasets.

Computer-intensive inference

Use properties of the dataset we have to help us *simulate* repeated datasets.

- Hypothesis testing – Monte Carlo tests – simulate data with H0 true.
- Confidence intevals:
  - Invert test statistic – simulate data with a range of values of H0.
  - Simulate data as close as possible to the data generating process – "resamples". This is the bootstrap.

Note, simulations can be:

- Nonparametric: sample from the data to create new datasets.
- Parametric: simulate from a distribution based on the data to create new datasets.

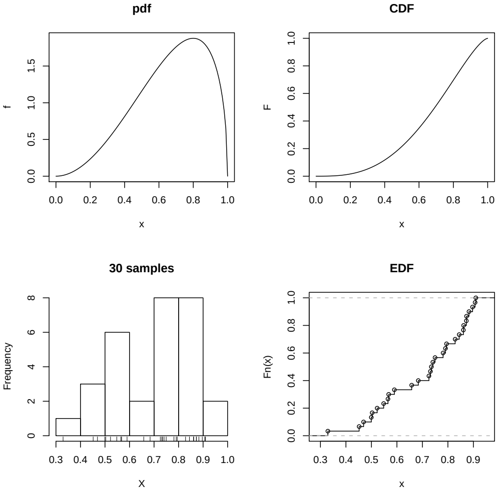
Outline

University of St Andrews

Outline

University of St Andrews

Empirical distribution function (EDF)



Nonparametric bootstrap and the EDF

- Nonparametric bootstrap:
  - generate resamples by simulating from the EDF
  - produces resamples that have the same properties as the data (c.f. previous methods where they had the same properties if *H0* was true)
  - best approximation we have to sampling again from the original data generating process

Nonparametric bootstrap and the EDF

- Nonparametric bootstrap:
  - generate resamples by simulating from the EDF
  - produces resamples that have the same properties as the data (c.f. previous methods where they had the same properties if *H0* was true)
  - best approximation we have to sampling again from the original data generating process
- How to sample from the EDF?
  - Resample with replacement from the data

University of St Andrews

Pine martin example

Original	0.13	-0.01	-0.01	0.42	-0.02	0.01	0.09	0.03	0.04	0.06	0.12	0.03
Resamples												
1	0.01	-0.01	-0.02	0.04	-0.01	0.13	0.03	-0.01	0.03	0.06	0.13	0.09
2	0.09	0.09	-0.02	-0.01	0.09	0.42	0.03	-0.01	0.04	-0.01	0.03	0.03
3	0.42	0.03	-0.01	0.09	0.06	0.06	0.06	-0.02	-0.01	0.09	0.4	0.09
.												
.												
.												
b	0.03	0.04	0.42	0.09	-0.01	-0.01	0.03	0.04	0.13	0.03	0.09	-0.01

- Notes:
  - number of times each data point occurs in each resample is a random variable
  - in b bootstrap resamples, we expect each data point to appear on average b times
  - can ensure they occur exactly b times with using a *balanced bootstrap* (but usually not worth the extra effort)

University of St Andrews

Outline

Bootstrap confidence intervals

- There are many ways to generate a confidence interval via bootstrap resampling[1]
- Today, we will cover the simplest – the "percentile method"

[1] see, e.g., Carpinter, J. and J. Bithell. (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine 19: 1141-1164.

Percentile method

- A useful approximate method for producing CIs
- Besides being relatively simple to understand and implement, it performs well in practice under most commonly-encountered situations
- Very widely used in applied statistics

Percentile method – Rationale

- Our resamples come from the same distribution as the original data

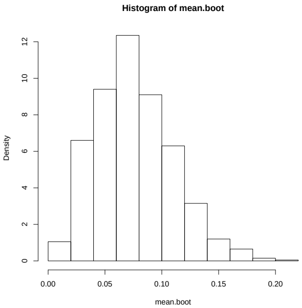
Percentile method – Rationale

- Our resamples come from the same distribution as the original data
- So can use the resamples to obtain the distribution of the quantity of interest – e.g., μ

University of St Andrews

Percentile method – Rationale

- Our resamples come from the same distribution as the original data
- So can use the resamples to obtain the distribution of the quantity of interest – e.g., μ
- Example: mean of pine martin data:



Percentile method (contd.)

- CI is an interval that contains the true value of μ 100$(1-\alpha)$% of the time

Percentile method (contd.)

- CI is an interval that contains the true value of μ $100(1-\alpha)\%$ of the time
- Turns out (under some conditions), the following procedure gives an interval with this property:
  - Order the μ
  - Lower limit is the $\alpha/2(b+1)\%$th value
  - Upper limit is the $(1-\alpha/2)(b+1)\%$th value

University of St Andrews

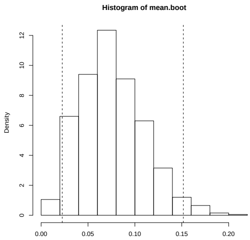
Percentile method (contd.)

- CI is an interval that contains the true value of μ $100(1-\alpha)\%$ of the time
- Turns out (under some conditions), the following procedure gives an interval with this property:
  - Order the μ
  - Lower limit is the $\alpha/2(b+1)$%th value
  - Upper limit is the $(1-\alpha/2)(b+1)$%th value
- E.g. For 999 resamples 95% limits given by the 25th smallest and 25th largest

University of St Andrews

Percentile method (contd.)

- CI is an interval that contains the true value of μ $100(1-\alpha)\%$ of the time
- Turns out (under some conditions), the following procedure gives an interval with this property:
  - Order the μ
  - Lower limit is the $\alpha/2(b+1)\%$th value
  - Upper limit is the $(1-\alpha/2)(b+1)\%$th value
- E.g. For 999 resamples 95% limits given by the 25th smallest and 25th largest

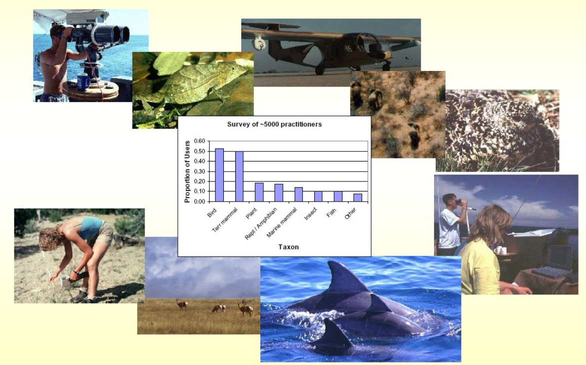


Example

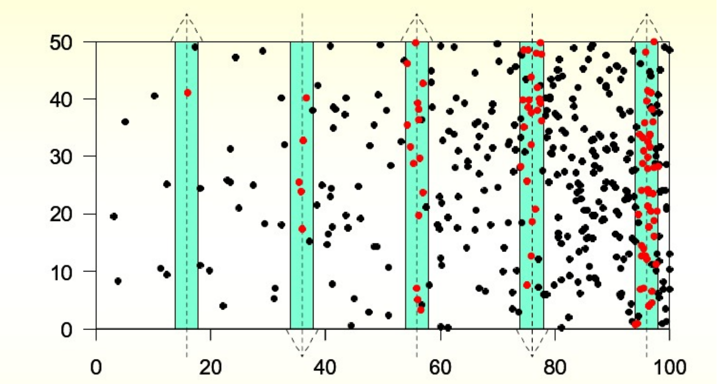
- This method is easy to apply and very general, so widely used

Example

- This method is easy to apply and very general, so widely used
- Example: distance sampling surveys of wildlife populations (other examples in the reading)



Example - distance sampling



- Dots represent groups of animals
- Red dots = detected groups;
  Black dots = undetected groups

Example - distance sampling

- Density

$$\hat{D} = \frac{\text{n seen} \times \text{mean group size}}{\text{area searched} \times \text{p detect}} = \frac{n \times \bar{s}}{2wL \times \hat{p}}$$

- Analytic variance $\hat{\text{var}}(\hat{D}) \approx \hat{D}^2 \left(\frac{var(n)}{n^2} \times \frac{var(\bar{s})}{\bar{s}^2} \times \frac{var(\hat{p})}{\hat{p}^2} \right)$
- Analytic CIs – assume $\hat{D}$ lognormally distributed
- Alternative – nonparametric bootstrap
  - Resample transects with replacement

Outline

1. Introduction and recap
2. Nonparametric bootstrap
   - How to generate resamples
   - Bootstrap confidence interval
   - Summary
3. Parametric bootstrap
   - How to generate resamples
   - Summary

Advantages

- Simple to apply
- General and robust (compared with other general analytic approaches) method of setting CIs
- No need for parametric assumption about $f()$ or quantity of interest

Disadvantages

- Assumes observations/samples are iid
  - *probability integral transform* method when not identically distributed)
  - *copulas* when samples are not independent

Disadvantages

- Assumes observations/samples are iid
  - *probability integral transform* method when not identically distributed)
  - *copulas* when samples are not independent
- Performs poorly when underlying distribution is very skewed, or estimator is very biased.
  - consider *BCa (Bias corrected and accelerated) bootstrap* (see Carpinter and Bithell 2000, cited earlier).

Disadvantages

- Assumes observations/samples are iid
  - *probability integral transform* method when not identically distributed)
  - *copulas* when samples are not independent
- Performs poorly when underlying distribution is very skewed, or estimator is very biased.
  - consider *BCa (Bias corrected and accelerated) bootstrap* (see Carpinter and Bithell 2000, cited earlier).
- Requires a reasonably large number of samples – 20-30 or more (but see *smoothed bootstrap* in background reading)

Disadvantages

- Assumes observations/samples are iid
  - *probability integral transform* method when not identically distributed)
  - *copulas* when samples are not independent
- Performs poorly when underlying distribution is very skewed, or estimator is very biased.
  - consider *BCa (Bias corrected and accelerated) bootstrap* (see Carpinter and Bithell 2000, cited earlier).
- Requires a reasonably large number of samples – 20-30 or more (but see *smoothed bootstrap* in background reading)
- In complex examples, it can be difficult to see what the unit for resampling should be

University of St Andrews

Disadvantages

- Assumes observations/samples are iid
  - *probability integral transform* method when not identically distributed)
  - *copulas* when samples are not independent
- Performs poorly when underlying distribution is very skewed, or estimator is very biased.
  - consider *BCa (Bias corrected and accelerated) bootstrap* (see Carpinter and Bithell 2000, cited earlier).
- Requires a reasonably large number of samples – 20-30 or more (but see *smoothed bootstrap* in background reading)
- In complex examples, it can be difficult to see what the unit for resampling should be
- Generally only asymptotically exact as b and $n \Rightarrow \infty$

University of St Andrews

Outline

1. Introduction and recap
2. Nonparametric bootstrap
   - How to generate resamples
   - Bootstrap confidence interval
   - Summary
3. Parametric bootstrap
   - How to generate resamples
   - Summary

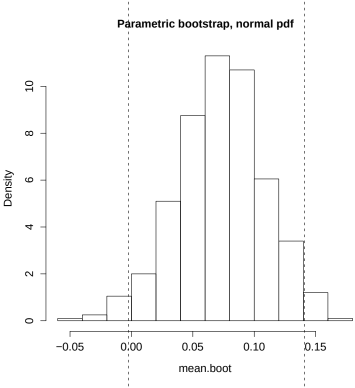
Outline

Introduction

- Parametric MC method for obtaining CIs
- Algorithm:
  - Fit a parametric model $f(\theta)$ to the data
  - Generate resamples by simulating from the fitted model
  - Use the resamples to obtain CIs (e.g., by percentile method)

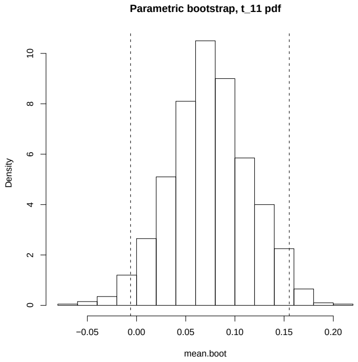
Pine martin example

- 999 resamples generated from $N(\bar{x}, s^2)$



Pine martin example

- 999 resamples generated from $t_{11}(\bar{x}, s^2)$



Outline

Advantages

- General and robust (compared with general analytic approaches) method of setting CIs
- Observations don't need to be iid

Disadvantages

- Generally only asymptotically exact as b and $n \Rightarrow \infty$
- Must assume a parametric model for $f(\theta)$