Imperial College London

Advanced Computational Methods in Statistics
Lecture 4
Bootstrap

Axel Gandy

Department of Mathematics
Imperial College London
http://www2.imperial.ac.uk/~agandy

London Taught Course Centre
for PhD Students in the Mathematical Sciences
Autumn 2014

Outline

Introduction

- Main idea:
  Estimate properties of estimators (such as the variance, distribution, confidence intervals) by resampling the original data.
- Key paper: Efron (1979)

Slightly expanded version of the key idea

- Classical Setup in Statistics:

$$X \sim F, \quad F \in \Theta$$

where X is the random object containing the entire observation. (often, $\Theta = \{F_a; a \in A\}$ with $A \subset \mathbb{R}^d$).
- Tests, CIs, ... are often built on a real-valued test statistics $T = T(X)$.
- Need distributional properties of T for the "true" F (or for F under H_0) to do tests, construct CIs,... (e.g. quantiles, sd, ...).
- Classical approach: construct T to be an (asymptotic) pivotal quantity, with distribution not depending on the unknown parameter. This is often not possible or requires lengthy asymptotic analysis.
- Key idea of bootstrap: Replace F by (some) estimate $\hat{F}$, get distributional properties of T based on $\hat{F}$.

Mouse Data

(Efron & Tibshirani, 1993, Ch. 2)

- 16 mice randomly assigned to treatment or control
- Survival time in days following a test surgery

Group	Data	Mean (SD)	Median (SD)
Treatment	94 197 16 38 99 141 23	86.86 (25.24)	94 (?)
Control	52 104 146 10 51 30 40 27 46	56.22 (14.14)	46 (?)
	Difference:	30.63 (28.93)	48 (?)

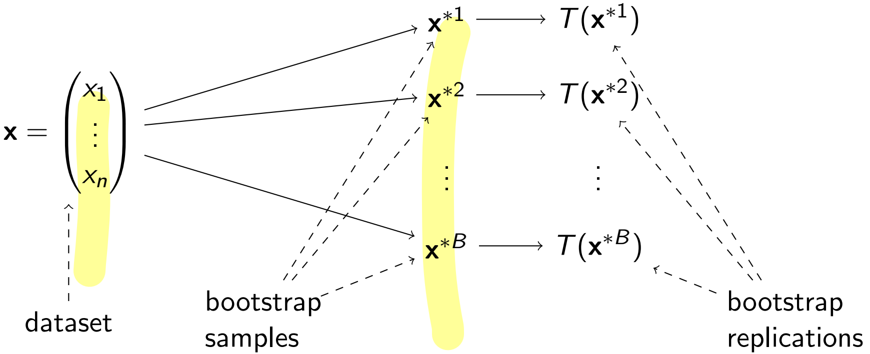
- Did treatment increase survival time?
- A good estimator of the the standard deviation of the mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample error

$$\hat{s} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- What estimator to use for the SD of the median?
- What estimator to use for the SD of other statistics?

Bootstrap Principle

- test statistic $T(\mathbf{x})$, interested in $SD(T(\mathbf{X}))$
- Resampling with replacement from $x_1, \ldots, x_n$ gives a bootstrap sample $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ and a bootstrap replicate $T(\mathbf{x}^*)$.
- get B independent bootstrap replicates $T(\mathbf{x}^{*1}), \ldots, T(\mathbf{x}^{*B})$
- estimate $SD(T(\mathbf{X}))$ by the empirical standard deviation of $T(\mathbf{x}^{*1}), \ldots, T(\mathbf{x}^{*B})$



Back to the Mouse Example

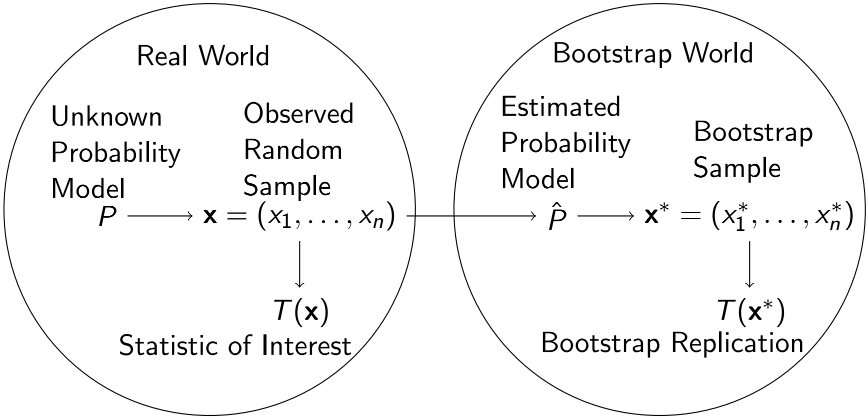
- B=10000
- Mean:

	Mean	bootstrap SD
Treatment	86.86	23.23
Control	56.22	13.27
Difference	30.63	26.75

- Median:

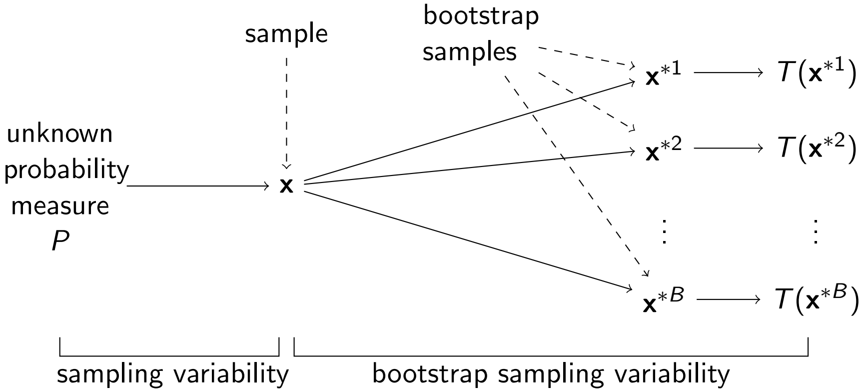
	Median	bootstrap SD
Treatment	94	37.88
Control	46	13.02
Difference	48	40.06

Illustration



Sources of Variability

- sampling variability (we only have a sample of size n)
- bootstrap resampling variability (only B bootstrap samples)

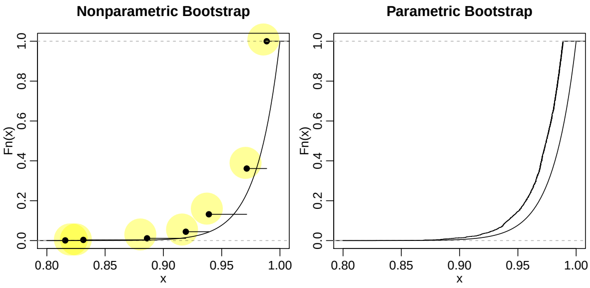


Parametric Bootstrap

- Suppose we have a parametric model $\mathrm{P}_\theta, \theta \in \Theta \subset \mathbb{R}^d$.
- $\hat{\theta}$ estimator of θ
- Resample from the estimated model $\mathrm{P}_{\hat{\theta}}$.

Example:Problems with (the Nonparametric) Bootstrap

- $X_1, \dots, X_{50} \sim U(0, \theta)$ iid, $\theta > 0$
- MLE $\hat{\theta} = \max(X_1, \dots, X_{50}) = 0.989$
- Non-parametric Bootstrap:
  $X_1^*, \dots, X_{50}^*$ sampled indep. from $X_1, \dots, X_{50}$ with replacement.
- Parametric Bootstrap: $X_1^*, \dots, X_{50}^* \sim U(0, \hat{\theta})$
- Resulting CDF of $\hat{\theta}^* = \max(X_1, \dots, X_{50})$:



- In the nonparametric bootstrap: Large probability mass at $\hat{\theta}$.
  In fact $P(\hat{\theta}^* = \hat{\theta}) = 1 - (1 - 1/n)^n \xrightarrow{n \to \infty} 1 - e^{-1} \approx .632$

Outline

Plug-in Principle I

- Many quantities of interest can be written as a functional T of the underlying probability measure P, e.g. the mean can be written as
$$T(\mathrm{P}) = \int x d\,\mathrm{P}(x).$$
- Suppose we have iid observation $X_1, \dots, X_n$ from P. Based on this we get an estimated distribution $\hat{\mathrm{P}}$ (empirical distribution or parametric distribution with estimated parameter).
- We can use $T(\hat{\mathrm{P}})$ as an estimator of $T(\mathrm{P})$.
For the mean and the empirical distribution $\hat{\mathrm{P}}$ of the observations X_i this is just the sample mean:
$$T(\hat{\mathrm{P}}) = \int x d\hat{\mathrm{P}}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Plug-in Principle II

- To determine the variance of the estimator $T(\hat{\mathrm{P}})$, compute confidence intervals for $T(\mathrm{P})$, or conduct tests we need the distribution of $T(\hat{\mathrm{P}}) - T(\mathrm{P})$.
- Bootstrap sample: sample $X_1^*, \ldots, X_n^*$ from $\hat{P}$; gives new estimated distribution P^*.
- Main idea: approximate the distribution of

$$T(\hat{\mathrm{P}}) - T(\mathrm{P})$$

by the distribution of

$$T(\mathrm{P}^*) - T(\hat{\mathrm{P}})$$

(which is conditional on the observed $\hat{\mathrm{P}}$).

Bootstrap Interval

- Quantity of interest is $T(\mathrm{P})$
- To construct a one-sided $1-\alpha$ CI we would need c s.t. $\mathrm{P}(T(\hat{\mathrm{P}}) - T(\mathrm{P}) \geq c) = 1-\alpha$.
  Then a $1-\alpha$ CI would be $(-\infty, T(\hat{P}) - c)$.
  Of course, P and thus c are unknown.
- Instead of c use c^* given by

$$\hat{\mathrm{P}}(T(\mathrm{P}^*) - T(\hat{\mathrm{P}}) \geq c^*) = 1-\alpha$$

  This gives the (approximate) confidence interval

$$(-\infty, T(\hat{P}) - c^*)$$

- Similarly for two-sided confidence intervals.

Studentized Bootstrap Interval

- Improve coverage probability by studentising the estimate.
- quantity of interest $T(\mathrm{P})$, measure of standard deviation $\sigma(\mathrm{P})$
- Base confidence interval on $\frac{T(\hat{\mathrm{P}})-T(\mathrm{P})}{\sigma(\hat{\mathrm{P}})}$
- Use quantiles from $\frac{T(\mathrm{P}^*)-T(\hat{\mathrm{P}})}{\sigma(\mathrm{P}^*)}$.

Efron's Percentile Method

- Use quantiles from $T(\mathrm{P}^*)$
- (less theoretical backing)
- Agrees with simple bootstrap interval for symmetric resampling distributions, but does not work well with skewed distributions.

Example - CI for Mean of Exponential Distribution I

- $X_1, \ldots, X_n \sim \text{Exp}(\theta)$ iid
- Confidence interval for $\operatorname{E} X_1 = \frac{1}{\theta}$.
- Nominal level 0.95
- One-sided confidence intervals:
  Coverage probabilities:

	10	20	40	80	160	320
Normal Approximation	0.845	0.883	0.904	0.919	0.928	0.934
Bootstrap	0.817	0.858	0.892	0.922	0.917	0.94
Bootstrap - Percentile Method	0.848	0.876	0.906	0.92	0.932	0.94
Bootstrap - Studentized	0.902	0.922	0.942	0.949	0.946	0.944

- 100000 replications for the normal CI, bootstrap CIs based on 2000 replications with 500 bootstrap samples each
- Substantial coverage error for small n
- Coverage error $\searrow$ as $n \nearrow$
- Studentized Bootstrap seems to be doing best.

Example - CI for Mean of Exponential Distribution II

- Two-sided confidence intervals
  Coverage probabilities:

	10	20	40	80	160	320
Normal Approximation	0.876	0.914	0.93	0.947	0.949	0.95
Bootstrap	0.828	0.89	0.906	0.928	0.936	0.942
Bootstrap - Percentile Method	0.854	0.896	0.921	0.926	0.923	0.93
Bootstrap - Studentized	0.944	0.943	0.936	0.936	0.954	0.946

  - Number of replications as before
  - Smaller coverage error than for one-sided test.
  - Again the studentized bootstrap seems to be doing best.

Outline

Hypothesis Testing through Bootstrapping

- Setup: $H_0 : \theta \in \Theta_0$ v.s. $H_1 : \theta \notin \Theta_0$
- Observed sample: $\mathbf{x}$
- Suppose we have a test with a test statistic $T = T(\mathbf{X})$ that rejects for large values
- p-value, in general: $p = \sup_{\theta \in \Theta_0} \mathrm{P}_\theta(T(\mathbf{X}) \geq T(\mathbf{x}))$
  If we know that only θ_0 might be true: $p = \mathrm{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x}))$
- Using the sample, find estimator $\hat{\mathrm{P}}_0$ of the distr. of $\mathbf{X}$ under H_0
- Generate iid $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*B}$ from $\hat{\mathrm{P}}_0$
- Approximate the p-value via

$$\hat{p} = \frac{1}{B} \sum_{i=1}^{B} \mathrm{I}(T(\mathbf{X}^{*i}) \geq T(\mathbf{x}))$$

- To improve finite sample performance, it has been suggested to use

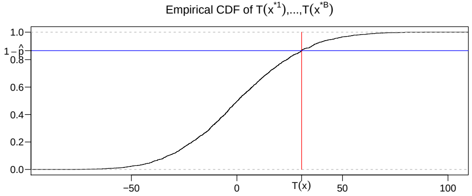
$$\hat{p} = \frac{1 + \sum_{i=1}^{B} \mathrm{I}(T(\mathbf{X}^{*i}) \geq T(\mathbf{x}))}{B+1}$$

Example - Two Sample Problem - Mouse Data

- Two Samples: treatment $\mathbf{y}$ and control $\mathbf{z}$ with cdfs F and G
- $H_0 : F = G$, $H_1 : G \leq_{st} F$
- $T(\mathbf{x}) = T(\mathbf{y}, \mathbf{z}) = \bar{\mathbf{y}} - \bar{\mathbf{z}}$, reject for large values
- Pooled sample: $\mathbf{x} = (\mathbf{y}', \mathbf{z}')$.
- Bootstrap sample $\mathbf{x}^* = (\mathbf{y}^{*\prime}, \mathbf{z}^{*\prime})$: sample from $\mathbf{x}$ with replacement
- p-value: generate independent bootstrap samples $\mathbf{x}^{*1}, \dots, \mathbf{x}^{*B}$

$$\hat{p} = \frac{1}{B} \sum_{i=1}^{B} \mathrm{I}\{T(\mathbf{x}^{*i}) \geq T(\mathbf{x})\}$$

- Mouse Data: $t_{obs} = 30.63$ B$= 2000$ $\hat{p} = 0.134$



How to Choose the Number of Resamples (i.e. B)? I

(Davison & Hinkley, 1997, Section 4.25)

- Not using the ideal bootstrap based on infinite number of resamples leads to a loss of power!
- Indeed, if $\pi_\infty(u)$ is the power of a fixed alternative for a test of level u then it turns out that the power $\pi_B(u)$ of a test based on B bootstrap resamples is

$$\pi_B(u) = \int_0^1 \pi_\infty(u) f_{(B+1)\alpha,(B+1)(1-\alpha)}(u) du$$

where $f_{(B+1)\alpha,(B+1)(1-\alpha)}(u)$ is the Beta-density with parameters $(B+1)\alpha$ and $(B+1)(1-\alpha)$.

How to Choose the Number of Resamples (i.e. B)? II

- If one assumes that $\pi_B(u)$ is concave, then one can obtain the approximate bound

$$\frac{\pi_B(\alpha)}{\pi_\infty(\alpha)} \geq 1 - \sqrt{\frac{1-\alpha}{2\pi(B+1)\alpha}}$$

A table of those bounds:

B=	19	39	99	199	499	999	9999
$\alpha = 0.01$	0.11	0.37	0.6	0.72	0.82	0.87	0.96
$\alpha = 0.05$	0.61	0.73	0.83	0.88	0.92	0.95	0.98

(these bounds may be conservative)

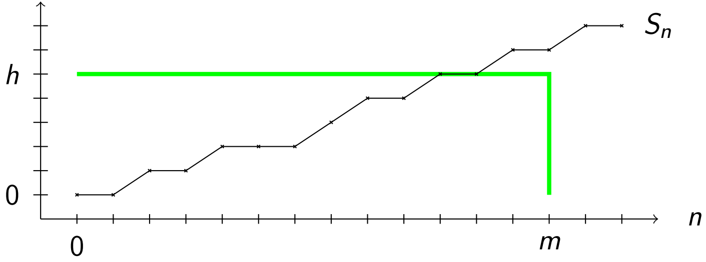
- To be safe: use at least $B = 999$ for $\alpha = 0.05$ and even a higher B for smaller α.

Sequential Approaches

- General Idea: Instead of a fixed number of resamples B, allow the number of resamples to be random.
- Can e.g. stop sampling once test decision is (almost) clear.
- Potential advantages:
  - Save computer time.
  - Get a decision with a bounded resampling error.
  - May avoid loss of power.

Saving Computational Time

- It is not necessary to estimate high values of the p-value p precisely.
- Stop if $S_n = \sum_{i=1}^{n} \mathtt{I}(T(\mathbf{X}^{*i}) \geq T(\mathbf{x}))$ "large".
- Besag & Clifford (1991):
  Stop after $\tau = \min\{n : S_n \geq h\} \wedge m$ steps

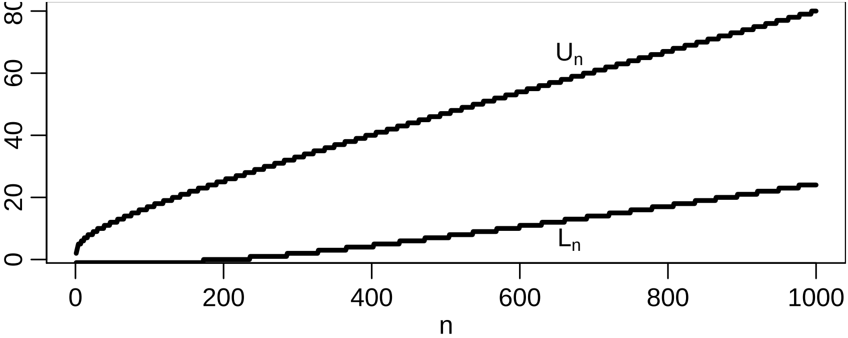


- Estimator: $\hat{p} = \begin{cases} h/\tau & S_\tau = h \\ (S_\tau + 1)/m & else \end{cases}$

Uniform Bound on the Resampling Risk

The boundaries below are constructed to give a uniform bound on the resampling risk: ie for some (small) $\epsilon > 0$,

$$\sup_p \mathrm{P}_p(\text{wrong decision}) \leq \epsilon$$



Details, see Gandy (2009).

Other issues

- How to compute the power/level (rejection probability) of Bootstrap tests?
  See (Gandy & Rubin-Delanchy, 2013) and references therein.
- How to use bootstrap tests in multiple testing corrections (eg FDR)?
  See (Gandy & Hahn, 2012) and references therein.

Outline

Main Idea

- Asymptotic theory does not take the resampling error into account - it assumes the 'ideal' bootstrap with an infinite number of replications.
- Observations $X_1, X_2, \dots$
- Often:

$$\sqrt{n}(T(\hat{\mathrm{P}}) - T(\mathrm{P})) \xrightarrow{d} F$$

  for some distribution F.
- Main asymptotic justification of the bootstrap: Conditional on the observed $X_1, X_2, \dots$:

$$\sqrt{n}(T(\mathrm{P}^*) - T(\hat{\mathrm{P}})) \xrightarrow{d} F$$

Conditional central limit theorem for the mean

$\sqrt{n}(\bar{X}_n - \mu)$
$\to N(0, \Sigma)$

- Let $X_1, X_2, \dots$ be iid random vectors with mean μ and covariance matrix Σ.
- For every n, suppose that $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$, where X_i^* are samples from $X_1, \dots, X_n$ with replacement.
- Then conditionally on $X_1, X_2, \dots$ for almost every sequence $X_1, X_2, \dots$,

$$\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \xrightarrow{d} N(0, \Sigma) \quad (n \to \infty).$$

- Proof:
  Mean and Covariance of $\bar{X}_n^*$ are easy to compute in terms of $X_1, \dots, X_n$.
  Use central limit theorem for triangular arrays (Lindeberg central limit theorem).

Delta Method

- Can be used to derive convergence results for derived statistics, in our case functions of the sample mean.
- Delta method: If ϕ is continuously differentiable, $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} T$ and $\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}) \xrightarrow{d} T$ conditionally then $\sqrt{n}(\phi(\hat{\theta}_n) - \phi(\theta)) \xrightarrow{d} \phi'(T)$ and $\sqrt{n}(\phi(\hat{\theta}^*_n) - \phi(\hat{\theta})) \xrightarrow{d} \phi'(T)$ conditionally.

Example

Suppose $\theta = \begin{pmatrix} \mathrm{E}(X) \\ \mathrm{E}(X^2) \end{pmatrix}$ and $\hat{\theta}_n = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \frac{1}{n}\sum_{i=1}^n X_i^2 \end{pmatrix}$. Then convergence of $\sqrt{n}(\hat{\theta} - \theta)$ can be established via CLT.
Using $\phi(\mu, \eta) = \eta - \mu^2$ gives a limiting result for estimates of variance.

Bootstrap and Empirical Process theory

- Flexible and elegant theory based on expectations wrt the empirical distribution

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

  (many test statistics can be constructed from this)
- Gives uniform CLTs/LLN: Donkser theorems/Glivenko-Cantelli theorems
- Can be used to derive asymptotic results for the bootstrap (e.g. for bootstrapping the sample median);
  use the bootstrap empirical distribution

$$\mathbb{P}_n^* = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i^*}.$$

- For details see van der Vaart (1998, Section 23.1) and van der Vaart & Wellner (1996, Section 3.6).

Outline

Introduction

- It can be shown that that the bootstrap has a faster convergence rate than simple normal approximations.
- Main tool: Edgeworth Expansion - refinement of the central limit theorem
- Main aim of this section: to explain the Edgeworth expansion and then mention briefly how it gives the convergence rates for the bootstrap.
- (reminder: this is still not taking the resampling risk into account, i.e. we still assume $B = \infty$)
- For details see Hall (1992).

Edgeworth Expansion

- θ_0 unknown parameter
- $\hat{\theta}_n$ estimator based on sample of size n
- Often,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2) \quad (n \to \infty),$$

  i.e. for all x,

$$\mathrm{P}(\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sigma} \leq x) \to \Phi(x) \quad n \to \infty,$$

  where $\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$, $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.
- Often one can write this as power series in $n^{-\frac{1}{2}}$:

$$\mathrm{P}(\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sigma} \leq x) = \Phi(x) + n^{-\frac{1}{2}} p_1(x)\phi(x) + \cdots + n^{-\frac{j}{2}} p_j(x)\phi(x) + \ldots$$

  This expansion is called Edgeworth Expansion.
- Note: p_j is usually an even/odd function for odd/even j.
- Edgeworth Expansion exist in the sense that for a fixed number of approximating terms, the remainder term is of lower order than the last included term.

Edgeworth Expansion - Arithmetic Mean I

- Suppose we have a sample $X_1, \dots, X_n$, and

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- Then
  - $p_1(x) = -\frac{1}{6}\kappa_3(x^2 - 1)$
  - $p_2(x) = -x\left(\frac{1}{24}\kappa_4(x^2 - 3) + \frac{1}{72}\kappa_3^2(x^4 - 10x^2 + 15)\right)$

  where κ_j are the cumulants of X, in particular
  - $\kappa_3 = \mathrm{E}(X - \mathrm{E}\, X)^3$ is the skewness
  - $\kappa_4 = \mathrm{E}(X - \mathrm{E}\, X)^4 - 3(\mathrm{Var}\, X)^2$ is the kurtosis.

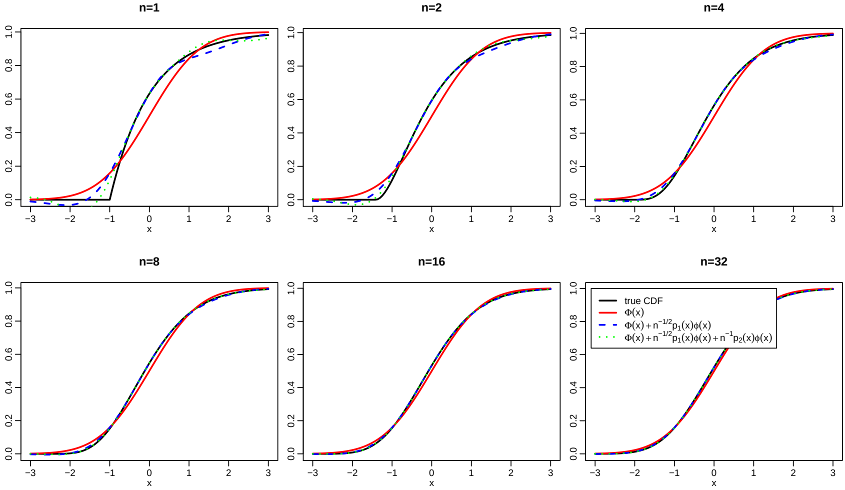
  (In general, the jth cumulant κ_j of X is the coefficient of $\frac{1}{j!}(it)^j$ in a power series expansion of the logarithm of the characteristic function of X.)

Edgeworth Expansion - Arithmetic Mean II

- The Edgeworth expansion exists if the following is satisfied:
  - Cramér's condition: $\lim_{|t|\to\infty} |\operatorname{E}\exp(itX)| < 1$ (satisfied if the observations are not discrete, i.e. possess a density wrt Lebesgue measure).
  - A sufficient number of moments of the observations must exist.

Edgeworth Expansion - Arithmetic Mean - Example

$X_i \sim \mathrm{Exp}(1)$ iid, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$



Coverage Prob. of CIs based on Asymptotic Normality I

- Suppose we construct a confidence interval based on the standard normal approximation to

$$S_n = \sqrt{n}(\hat{\theta}_n - \theta_0)/\sigma$$

where σ is the asymptotic variance of $\sqrt{n}\hat{\theta}_n$.

- One-sided nominal α-level confidence intervals:

$$I_1 = (-\infty, \hat{\theta} + n^{-1/2}\sigma z_\alpha)$$

where z_α is defined by $\Phi(z_\alpha) = \alpha$.

$$\begin{aligned}
P(\theta_0 \in I_1) &= P(\theta_0 < \hat{\theta} + n^{-1/2}\sigma z_\alpha) = P(S_n > -z_\alpha) \\
&= 1 - \left(\Phi(-z_\alpha) + n^{-1/2} p_1(-z_\alpha)\phi(-z_\alpha) + O(n^{-1})\right) \\
&= \alpha - n^{-1/2} p_1(z_\alpha)\phi(z_\alpha) + O(n^{-1}) \\
&= \alpha + O(n^{-1/2})
\end{aligned}$$

Coverage Prob. of CIs based on Asymptotic Normality II

- Two-sided nominal α-level confidence intervals:

$$I_2 = (\hat{\theta} - n^{-1/2}\sigma x_\alpha, \hat{\theta} + n^{-1/2}\sigma x_\alpha)$$

where $x_\alpha = z_{(1+\alpha)/2}$,

$$\begin{aligned}
P(\theta_0 \in I_2) =& P(S_n \leq x_\alpha) - P(S_n \leq -x_\alpha) \\
=& \Phi(x_\alpha) - \Phi(-x_\alpha) \\
& + n^{-1/2}[p_1(x_\alpha)\phi(x_\alpha) - p_1(-x_\alpha)\phi(-x_\alpha)] \\
& + n^{-1}[p_2(x_\alpha)\phi(x_\alpha) - p_2(-x_\alpha)\phi(-x_\alpha)] \\
& + n^{-3/2}[p_3(x_\alpha)\phi(x_\alpha) - p_3(-x_\alpha)\phi(-x_\alpha)] + O(n^{-2}) \\
=& \alpha + 2n^{-1}p_2(x_\alpha)\phi(z_\alpha) + O(n^{-2}) = \alpha + O(n^{-1})
\end{aligned}$$

- To summarise: Coverage error for one-sided CI: $O(n^{-1/2})$, for two-sided CI: $O(n^{-1})$.

Higher Order Convergence of the Bootstrap I

- Will consider the studentized bootstrap first.
- Consider the following Edgeworth expansion of $\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n}$:

$$\mathrm{P}\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq x\right) = \Phi(x) + n^{-\frac{1}{2}} p_1(x)\phi(x) + O\left(\frac{1}{n}\right)$$

- The Edgeworth expansion usually remains valid in a conditional sense, i.e.

$$\hat{\mathrm{P}}\left(\frac{\hat{\theta}_n^* - \hat{\theta}_n}{\sigma_n^*} \leq x\right) = \Phi(x) + n^{-\frac{1}{2}} \hat{p}_1(x)\phi(x) + \dots + n^{-\frac{j}{2}} \hat{p}_j(x)\phi(x) + \dots$$

  Use the first expansion term only , i.e.

Higher Order Convergence of the Bootstrap II

$$\hat{\mathrm{P}}\left(\frac{\hat{\theta}_n^* - \hat{\theta}_n}{\sigma_n^*} \leq x\right) = \Phi(x) + n^{-\frac{1}{2}}\hat{p}_1(x)\phi(x) + O\left(\frac{1}{n}\right)$$

Usually $\hat{p}_1(x) - p_1(x) = O(\frac{1}{\sqrt{n}})$.

- Then

$$\mathrm{P}\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq x\right) - \hat{\mathrm{P}}\left(\frac{\hat{\theta}_n^* - \hat{\theta}_n}{\sigma^*} \leq x\right) = O\left(\frac{1}{n}\right)$$

- Thus the studentized bootstrap results in a better rate of convergence than the normal approximation (which is $O(1/\sqrt{n})$ only).
- For a non-studentized bootstrap the rate of convergence is only $O(1/\sqrt{n})$.

Higher Order Convergence of the Bootstrap III

- This translates to improvements in the coverage probability of (one-sided) confidence intervals.
  The precise derivations of these also involve the so-called Cornish-Fisher expansions, an expansion of quantile functions similar to the Edgeworth expansion (which concerns distribution functions).

Outline

Introduction

- Iterate the Bootstrap to improve the statistical performance of bootstrap tests, confidence intervals,...
- If chosen correctly, the iterated bootstrap can have a higher rate of convergence than the non-iterated bootstrap.
- Can be computationally intensive.
- Some references: Davison & Hinkley (1997, Section 3.9), Hall (1992, Section 1.4,3.11)

Double Bootstrap Test

(based on Davison & Hinkley, 1997, Section 4.5)

- Ideally the p-value under the null distribution should be a realisation of $U(0,1)$.
- However, computing p-values via the bootstrap does not guarantee this
  (measures such as studentising the test statistics may help - but there is no guarantee)
- Idea: use an iterated version of the bootstrap to correct the p-value.
- let p be the p-valued based on $\hat{\mathrm{P}}$.
- observed - data $\rightarrow$ fitted model $\hat{\mathrm{P}}$;
- Let p^* be the random variable obtained by resampling from $\hat{\mathrm{P}}$.
- $p_{adj} = \mathrm{P}^*(p^* \leq p|\hat{\mathrm{P}})$

Implementation of a Double Bootstrap Test

Suppose we have a test that rejects for large values of a test statistic.
Algorithm: For $r = 1, \ldots, R$:

- Generate $X_1^*, \ldots X_n^*$ from the fitted null distribution $\hat{\mathrm{P}}$, calculate the test statistic t_r^* from it
- Fit the null distribution to $X_1^*, \ldots, X_n^*$ obtaining $\hat{\mathrm{P}}_r$
- For $m = 1, \ldots, M$:
  - generate $X_1^{**}, \ldots X_n^{**}$ from $\hat{\mathrm{P}}_r$
  - calculate the test statistic t_{rm}^{**} from them
- Let $p_r^* = \frac{1+\#\{t_{rm}^{**} \geq t_r^*\}}{1+M}$.

Let $p_{\text{adj}} = \frac{1+\#\{p_r^* \leq p\}}{1+M}$
Effort: MR simulations.
M can be chosen smaller than R, e.g. $M = 99$ or $M = 249$.

Outline

Dependent Data

- Often observations are not independent
- Example: time series
- $\rightarrow$ Bootstrap needs to be adjusted
- Main source for this chapter: Lahiri (2003).

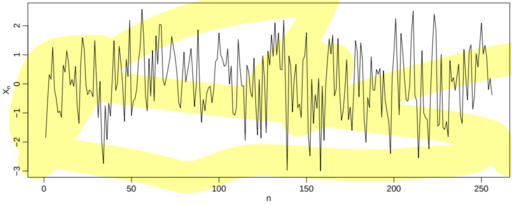
Dependent Data - Example I

(Lahiri, 2003, Example 1.1, p. 7)

- $X_1, \ldots, X_n$ generated by a stationary ARMA(1,1) process:

$$X_i = \beta X_{i-1} + \epsilon_i + \alpha \epsilon_{i-1}$$

  where $|\alpha| < 1$, $|\beta| < 1$, (ϵ_i) is white noise, i.e. $\mathrm{E}\,\epsilon_i = 0$, $\mathrm{Var}\,\epsilon_i = 1$.
- Realisation of length $n = 256$ with $\alpha = 0.2$, $\beta = 0.3$, $\epsilon_i \sim N(0, 1)$:



Dependent Data - Example II

- Interested in variance of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.
- Use the Nonoverlapping Block Bootstrap (NBB); Blocks of length l:
  - $B_1 = (X_1, \dots, X_l)$
  - $B_2 = (X_{l+1}, \dots, X_{2l})$
  - $\dots$
  - $B_{n/l} = (X_{n-l+1}, \dots, X_n)$

- resample blocks $B_1^*, \dots, B_{n/l}^*$ with replacement; concatenate to get bootstrap sample

$$(X_1^*, \dots, X_n^*)$$

- Bootstrap estimator of variance: $\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n X_i^*)$
  (can be computed explicitly in this case - no resampling necessary)

Dependent Data - Example III

- Results for the above sample:
  True Variance $\mathrm{Var}(\bar{X}_n) = 0.0114$ (based on 20000 simulations)

l	1	2	4	8	16	32	64
$\widehat{\mathrm{Var}}(\bar{X}_n)$	0.0049	0.0063	0.0075	0.0088	0.0092	0.0013	0.0016

- bias, standard deviation, $\sqrt{\mathrm{MSE}}$ based on 1000 simulations:

l	1	2	4	8	16	32	64
bias	-0.0065	-0.0043	-0.0025	-0.0016	-0.0013	-0.0017	-0.0031
sd	5e-04	0.001	0.0016	0.0024	0.0035	0.0052	0.0069
$\sqrt{\mathrm{MSE}}$	0.0066	0.0044	0.003	0.0029	0.0038	0.0055	0.0076

Note:

- block size =1 is the classical IID bootstrap
- Variance increases with block size
- Bias decreases with block size
- Bias-Variance trade-off

Moving Block Bootstrap (MBB)

- $X_1, \dots, X_n$ observations (realisations of a stationary process)
- l block length.
- $B_i = (X_i, \dots, X_{i+l-1})$ block starting at X_i.
- To get a bootstrap sample:
  - Draw with replacement $B_1^*, \dots, B_k^*$ from $B_1, \dots, B_{n-l+1}$.
  - Concatenate the blocks $B_1^*, \dots, B_k^*$ to give the bootstrap sample $X_1^*, \dots, X_{kl}^*$
- $l = 1$ corresponds to the classical iid bootstrap.

Nonoverlapping Block Bootstrap (NBB)

- Blocks in the MBB may overlap
- $X_1, \dots, X_n$ observations (realisations of a stationary process)
- l block length.
- $b = \lfloor n/l \rfloor$ blocks:

$$B_i = (X_{il+1}, \dots, X_{il+l-1}), \quad i = 0, \dots, b-1$$

- To get a bootstrap sample: draw with replacement from these blocks and concatenate the resulting blocks.
- Note: Fewer blocks than in the MBB

Other Types of Block Bootstraps

- Generalised Block Bootstrap
  - Periodic extension of the data to avoid boundary effects
  - Reuse the sample to form an infinite sequence (Y_k):

$$X_1, \dots, X_n, X_1, \dots, X_n, X_1, \dots, X_n, X_1, \dots$$

  - A block $B(S, J)$ is described by its start S and its length J.
  - The bootstrap sample is chosen according to some probability measure on the sequences $(S_1, J_1), (S_2, J_2), \dots$
- Circular block bootstrap (CBB):

  sample with replacement from $\{B(1, l), \dots, B(n, l)\}$

  $\rightarrow$ every observation receives equal weight
- Stationary block bootstrap (SB):

$$S \sim \text{Uniform}(1, \dots, n), \quad J \sim \text{Geometric}(p)$$

  for some p.

  $\rightarrow$ blocks are no longer of equal size

Dependent Data - Remarks

- MBB and CBB outperform NBB and SB (Lahiri, 2003, see Chapter 5)
- Dependence in Time Series is a relatively simple example of dependent data
- Further examples are Spatial data or Spatio-Temporal data - here boundary effects can be far more difficult to handle.

Outline

Bagging I

- Acronym for *bootstrap aggregation*
- data $d = \{(\mathbf{x}^{(j)}, y^{(j)}), j = 1, \dots, n\}$
  response y, predictor variables $\mathbf{x} \in \mathbb{R}^p$
- Suppose we have a basic predictor $m_0(\mathbf{x}|d)$
- Form R resampled data sets $d_1^*, \dots, d_R^*$.
- empirical bagged predictor:

$$\hat{m}_B(\mathbf{x}|d) = \frac{1}{R}\sum_{r=1}^{R} m_0(\mathbf{x}|d_r^*)$$

  This is an approximation to

$$m_B(\mathbf{x}|d) = \mathrm{E}^*\{m_0(\mathbf{x}|D^*)\}$$

  D^* resample from d.

Bagging II

- Example: linear regression with screening of predictors (hard thresholding)

$$m_0(\mathbf{x}|d) = \sum_{i=1}^{p} \hat{\beta}_i \mathtt{I}(|\hat{\beta}_i| > c_i) x_i$$

corresponding bagged estimator:

$$m_B(\mathbf{x}|d) = \sum_{i=1}^{p} \mathsf{E}^*(\hat{\beta}_i \mathtt{I}(|\hat{\beta}_i| > c_i)|D^*) x_i$$

corresponds to soft thresholding

- Bagging can improve in particular unstable classifiers (e.g. tree algorithms)

Bagging III

- For classification problems concerning class membership (i.e. a 0-1 decision is needed), bagging can work via voting (the class that the basic classifier chooses most often during resampling is reported as class)
- Key Articles: Breiman (1996a,b), Bühlmann & Yu (2002)

Boosting

- Related to Bagging
- attach weights to each observation
- iterative improvements of the base classifier by increasing the weights for those observations that are hardest to classify
- Can yield dramatic reduction in classification error.
- Key articles: Freund & Schapire (1997), Schapire et al. (1998)

Pointers to the Literature

Random Forest

- Efron & Tibshirani (1993) - easy to read introduction.
- Hall (1992) - Higher order asymptotics
- Lahiri (2003) - Dependent Data
- Davison & Hinkley (1997) - More applied book about the bootstrap in several situations with implementations in R.
- van der Vaart (1998, Chapter 23): Introduction to the Asymptotic Theory of Bootstraps.
- van der Vaart & Wellner (1996, Section 3.6): Asymptotic Theory based on empirical process theory.
- Special Issue of *Statistical Science*: 2003, Vol 18, No. 2, in particular Davison et al. (2003)

Part I

Appendix

Next lecture

- Particle Filtering

References I

Besag, J. & Clifford, P. (1991). Sequential Monte Carlo p-values. *Biometrika* **78**, 301–304.

Breiman, L. (1996a). Bagging predictors. *Machine Learning* **24**, 123–140.

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 2350–2383.

Bühlmann, P. & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics* **30**, 927–961.

Davison, A. & Hinkley, D. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Davison, A. C., Hinkley, D. V. & Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science* **18**, 141–157.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**, 1–26.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

References II

Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting,. *Journal of Computer and System Sciences* **55**, 119 – 139.

Gandy, A. (2009). Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association* **104**, 1504–1511.

Gandy, A. & Hahn, G. (2012). MMCTest - a safe algorithm for implementing multiple Monte Carlo tests. *arXiv:1209.3963 [stat.ME]* .

Gandy, A. & Rubin-Delanchy, P. (2013). An algorithm to compute the power of Monte Carlo tests with guaranteed precision. *Annals of Statistics* **41**, 125–142.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer Verlag.

Lahiri, S. (2003). *Resampling Methodes for Dependant Data*. Springer.

Schapire, R. E., Freund, Y., Bartlett, P. & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics* **26**, 1651–1686.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. & Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer.